Melanie Ellis

The impact of a national examination in foreign language on teaching, learning and the work of schools

WYDAWNICTWO NAUKOWE
UNIWERSYTETU PEDAGOGICZNEGO KRAKÓW

The impact of a national examination in foreign language on teaching, learning and the work of schools

Uniwersytet Pedagogiczny
im. Komisji Edukacji Narodowej
w Krakowie

Prace Monograficzne 820

Melanie Ellis

The impact of a national examination in foreign language on teaching, learning and the work of schools

WYDAWNICTWO NAUKOWE
UNIWERSYTETU PEDAGOGICZNEGO KRAKÓW 2018

Recenzenci
dr hab. Ewa Piechurska-Kuciel, prof. UO
dr hab. Krystyna Droździał-Szelest, prof. UAM

projekt okładki Janusz Schneider
łamanie Bożena Ślusarczyk-Kołaczewska

ISSN 0239-6025
ISBN 978-83-8084-105-5
DOI 10.24917/9788380841055

Wydawnictwo Naukowe UP
30-084 Kraków, ul. Podchorążych 2
tel./faks 12-662-63-83, tel. 12-662-67-56
e-mail: wydawnictwo@up.krakow.pl
http://www.wydawnictwoup.pl

druk i oprawa Zespół Poligraficzny WN UP

Acknowledgements

This book could not have happened were it not for the support of a large number of people. Thanks goes first to all the schools, teachers and principals who took the time and made the effort to take part in the different phases of the project. Particular thanks goes to Marianna Niesobska and her staff, who were tremendously helpful and offered useful constructive criticism at key moments.

Phase Three of the project would not have taken the shape it did if it were not for the late Magdalena Szpotowicz, who engaged me to work at the Educational Research Institute in Warsaw, gave me the possibility to design BUNJO, and supported me fully in Year 1 of that study, the part of that project from which come the data for the Impact Study reported here as Phase III. Anna Michałowska worked shoulder to shoulder with me on the project management for that phase. Małgorzata Proskura did magnificent work overseeing the data collection. Katarzyna Pacuszka gave useful comments on the instruments, provided the translation, and with Katarzyna Kutyłowska worked on materials for training data collectors and ran a training session. Dorota Campfield and Małgorzata Szulc-Kurpaska piloted the interview materials and offered helpful comments and insights. Agata Gajewska-Dyszkiewicz worked with me to prepare the tender documents and provided invaluable input on practical aspects. Dominika Chandler provided endless moral support. From Kamil Sijko I learnt a great deal about coding.

Special thanks for work on the coding and positive support go to Jolanta Pucek.

Ewa Piechurska-Kuciel made some very helpful suggestions on how to improve the organization of the text.

I am also grateful to the late Alan Davies, who got me interested in language testing long ago in Edinburgh, and from whom I first heard about washback.

Longitudinal research is hard work and it is easy to lose heart on the way. I have had enormous and constant positive support from my mother, and patience, tolerance and good humour from Marek.

With apologies to all who have been neglected over the last two years, but went on sending warm wishes. It really helped, thank you.

Introduction

In 1999 Poland began reform of the education system. The structure of the system of schools changed, which created a new type of school for years 7 to 9, for young people between the ages of 13 and 16, the lower secondary school, or *gimnazjum*. Alongside the structural reform came a new national core curriculum and new legislation for schools. For the first time a national system of examinations was introduced and a network of dedicated institutions created, led by the Central Examination Board, which were responsible for them. The first cohort of learners who completed the new lower secondary school sat the first national examinations in 2002 in arts subjects, mathematics and natural sciences. In 2009 papers in foreign languages were added to the suite. Prior to this event foreign languages had been tested only in the school-leaving examination, for learners in year 12, so this new examination presented a new challenge for schools and learners.

The research described in the study was initially conceived to investigate whether there was any evidence of *washback* from the new examination on the teaching and learning process in *gimnazjum*. In order to be able to state that any change had taken place it was first necessary to conduct a *baseline study* to obtain a picture of how English was being taught in mainstream lower secondary schools. This took place in 2008, the school year prior to the first administration of the new examination. Using questionnaires and classroom observation enabled an impression of the learning of English to be reported. The next phase of the study was conducted a year later, shortly after the first examination, and looked to see if there were signs of washback. Results of the first examination showed that there were in effect two different populations of learners, those who obtained high results for whom the test was easy, and a second group, similar in size, who had been challenged by the test and scored much lower. Contextual factors, such as shortage of teachers, teacher qualifications and differing lengths of school language learning experience were thought to contribute to this. In the first three years the examination was effectively a low-stakes test, with no consequences arising from the test results. This was treated as an interim introductory period. It was not until 2012 that the exam results would count, together with points from the other subject papers, to selection for upper secondary schools. However, as is often the way with pedagogical research in schools, further changes came about which were to result in the foreign language exam being revised. Changes were made to the point at which the first foreign language was introduced, lowering it to the start of primary school. In addition, the second foreign language was also to start earlier, at the beginning of Year 7, the first class of *gimnazjum.*

These changes necessitated revision of the core curriculum, which for foreign languages meant a move more closely towards the Common European Framework of Reference for Languages both in description of content and levels. Curricular standards became operationalized in terms of what the learner can do. Changes in the curriculum forced changes in the national examinations, which meant there was an opportunity to address the polarized learner population. In 2012 a revised foreign language examination was launched, with two levels, basic and extended. For those continuing the same foreign language from primary school the extended level was obligatory, in addition to the basic exam. The third phase of the research, the Impact Study, took place at the time of the introduction of the revised exam.

Messick (1996) challenges researchers to provide evidence that any changes found in the teaching or in the education system result from the examination. The research question then moved beyond whether washback was taking place, to trying to discover what was causing the exam to have effects on some teachers and some learners but not on others (Wall & Alderson, 1993) and in some schools more than in others. Using parts of interviews with teachers, school principals and learners drawn from another large scale study, together with questionnaire extracts, the main part of this project looks to fill gaps in what is known about washback to date. Moving beyond the classroom, it looks also at the school in its context and factors relating to the school and its examination results. Effects at this level are referred to as *impact*, seen as the effects of the examination felt beyond the immediate classroom situation. It is more appropriate to see the school as a complex dynamic system, as there are many factors which all interact with each other, across all areas of the school and out into the district surrounding it. The aim of this work has been to try to unravel this complex mix, to search for patterns and attempt to find factors which contribute to the coming about of washback and impact from the national examination in English. Many factors were identified, some of which are being described for the first time.

The first part of the book provides a theoretical and contextual background for the research project. Chapter One gives an introduction to the concepts of washback and impact, presents and discusses models and examines, through a review of the literature in both mainstream education and foreign language assessment, how understanding of the concepts has developed and changed as the field has grown.

In Chapter Two the focus moves to discover how tests are used in society by studying how test scores are used. The opening section gives an overview of the concept of *validity* and shows how the impact and consequences of a test are seen as an important part of this. Examinations are used for many purposes in society, as a means of control, for selection, but also to better social equity. To what extent the test or the test provider can be held responsible for the consequences resulting from use of test scores forms

part of the discussion of ethical issues connected with testing. Another aspect of this is whether a test can be considered *fair*. This deals with the notion of equity and underlines the importance that a test should not advantage any group of test-takers above any others. The final sections of the chapter discuss the use of examination results in accountability systems for evaluating the quality of education. This is investigated at depth with examples from the USA, and England. The chapter closes with an introduction to the role of examinations in the educational system in Poland. It explains how examination results are used as one element in the process of evaluation of quality in schools and as a sole measure of the effectiveness of educational outcomes.

Chapter Three traces the recent history of national examinations in Poland, placing them within the context of reform of organization of the school system and implementation of a new national core curriculum. The situation that prevailed at that time in foreign language education in schools is presented, with the aim of clarifying certain policy decisions which were made. Next, the origins of the foreign language component of the national examination at the end of lower secondary school are traced. This is followed by a description of the core curriculum for foreign languages and an analysis of the thinking behind it. We then turn to the test in English which forms part of the national examination at the end of lower secondary school and analyze its construct in an attempt to discern and describe the model of language inherent in its design. The content of the test is scrutinized and compared with the core curriculum which reveals apparent difficulties with content under-representation in terms of the testing of speaking. Analysis of the description of the test in the examination syllabus also shows that reading is weighted more heavily than other skills. The chapter then moves chronologically through the history of the examination in English, explaining the results of the first three years, and then detailing the revisions made to the test in 2012 in response to further educational reforms. This chapter provides the background for the research study of the impact of this examination which is described in the second part of the book.

Part Two of the book presents a longitudinal study of the impact of the new national examination in English at the end of lower secondary school.

Chapter Four explains the rationale behind the study, and traces the methodology used to investigate washback and impact. Cheng et al. (2015: 463) sum up what could be described as the approach taken in this research:

> Future research needs to investigate washback as a dimension of impact and explore the relationship between washback and other forces operative in the educational context.

As a large part of the data is qualitative from interviews, the specific challenges this poses are considered. The design of the study is then explained and a description given of its three phases and the relationships

between them. The research was conducted over a period of five years on three different samples and groups of participants, all in the context of lower secondary schools. The aims and research questions for the project as a whole and for each phase are explained. Next, detailed information is given about sampling and the participants in the three phases. Instruments and procedures used for data collection are introduced together with a description of the process of analysis. The study is descriptive in nature, taking a pragmatic-realist approach. Principles have been adopted from ethnography in how the data collection is designed and conducted, and from grounded theory in how the qualitative data is analyzed.

Chapter Five presents the data from each of the three stages of the project, with a focus on Phase Three, the Impact Study. Where possible, qualitative data is corroborated with data from other instruments which obtain quantitative results. In some places the qualitative data is initially treated quantitatively and then illustrated with excerpts of text. The chapter ends with a critical analysis of the limitations of the study.

Chapter Six takes the research questions from each phase of the study and discusses them, where appropriate comparing findings between respondents and across the phases. As evidence of both washback and impact were noted, the next sections use the findings to attempt to explain how the mechanisms of washback and impact work. For some aspects there is empirical evidence, while others allow for the formulation of hypotheses. Implications are made for the different stakeholders affected by the examination. At the micro-level of school these are teachers, school principals, learners and their families, while in the macro-context these include institutions providing teacher education, local education authorities, the inspectorate and educational policy makers.

Xie & Andrews (2013: 6) wrote that

> In a given educational setting contextual factors and individual factors interact and largely determine both the kind and amount of washback that occurs. However, it is still not clear in what ways these factors combine and interact with each other.

It would appear that the research presented here has helped go some way towards filling this gap.

CHAPTER 1

Washback and Impact

This book concerns the role the external public examination in foreign language plays in the teaching and learning process in lower secondary school in Poland and within the education system in general. The external examination is mandatory for all young people in mainstream education at the end of the third stage of education, currently Year 9, when they are 15–16 years old.

In this chapter we begin to consider the relationship between examinations and teaching and learning, first by discussing the effects which a test may have on a small-scale, within a classroom, on learners and teachers and on the learning and teaching process. We study the concept of washback, trace the history of how understanding of its workings have developed, through examining different models, offered by various authors, and review work done on the washback of tests in general education and also on tests of foreign language. As we follow these studies we see how research findings have contributed to current understanding of the phenomenon of washback.

In the second part of the chapter we move to a larger scale, considering the role examinations play beyond the individual classroom, in education systems, which is referred to as test impact. As impact is a complex process, it is conceptualized in different ways by different authors and we will see connections with the fields of educational evaluation, innovation and change, and social studies. We will look at different models which attempt to explain the mechanisms at work and show the large number of factors involved in the process. These models become clearer as we review their application in studies conducted on the impact of national examinations in different countries and of international examinations, used worldwide. Here we will also see how studying the effect of an international exam in different contexts is used within the process of test design and validation.

1.1. Defining terms

Testing and assessment are an integral part of education in school. *Assessment* is used here to refer to the process by which teachers gather informa-

tion about their learners: how their students are coping with the contents of the course they are following, what progress they are making, what they have achieved over a period of time, and what strengths or weaknesses they may have. Assessment may also refer to the teacher's evaluation of the contents of the course, its syllabus, or the materials used to implement it, or to the teacher's self-evaluation of the effectiveness of their work. *Testing* is one way in which assessment is conducted. *Tests* may be small in scope and informal, a quick check devised by the teacher to find out, for example, if learners have memorised a set of new words, can form negative statements correctly, or can match appropriate responses to a series of questions. This term can also refer to tests, or *examinations* (which in this work will be used interchangeably withthe word test), prepared by external bodies which administer the same test on a wide scale.

Examinations which have important significance for the test taker or for society are known as *high-stakes* tests. Madaus (1988: 29) defines these as examinations where "the test results are directly related to important rewards or sanctions for students, teachers or institutions," while Popham (1987, cited in Smith 1991: 9) refers to high stakes tests as those "whose results are used to trigger actions or decisions…" with reference to learners, teachers, principals and school districts. High stakes tests are often found at threshold points in an educational system where their results may act as selective filters, deciding on the educational future of the young people taking them. More information about the consequences of such use of test results can be found in Chapter Two.

External examinations we understand as national tests, which are developed, administered and marked by a body outside school (such as an examination board), so as to ensure that every child receives the same test, conducted and marked in the same circumstances. Such examinations are also known as *standardised tests* which include both examinations produced by public bodies at state or national level and those produced by commercial international testing agencies. These are to be contrasted with *school-based* examinations which are usually devised by a teacher, or group of teachers, administered in a particular group, class or school and marked by the teacher.

It has long been accepted in educational circles that high stakes testing has an influence on the working of school systems (Pearson, 1988: 98). In some contexts high stakes testing programmes have been designed with the deliberate intention of implementing change. In general education this approach is sometimes referred to as *measure-driven instruction* (MDI) (Popham, 1987). The belief is that if a test is high stakes, is criterion referenced (ie. tests a set of competences or skills), involves direct testing of real skills and useful knowledge, and includes a package of support for teachers in how to gear their instruction to the skills to be tested, then the examination will bring about an improvement both in teaching and in

learning outcomes (Popham, 1987: 680). In foreign language testing Pearson (1988) referred to this practice as the use of tests as *levers for change*. Andrews et al. (2002: 208) use the term *engineered washback* to refer to the attempt to implement change through the introduction or adaptation of examinations. As we will see in later sections these views are controversial.

In applied linguistics research the effects of examinations on teaching (Wall and Alderson, 1993: 41) and learning (Bailey, 1996: 259; Wall, 1997: 291) are known as *washback* or *backwash* (Hughes, 1993; Weir, 1990), terms which are used interchangeably. This section aims only to offer working definitions of terms and all concepts raised will be fully explored in later sections and chapters. Alderson and Wall (1993: 117) define washback as things that teachers and learners do "that they would not necessarily otherwise do because of the test" (emphasis removed). Green (2007: 1), taking into consideration what has been learnt about the washback phenomena in the decade following those first definitions (see review below), sees washback as "grounded in the relationship between preparation for success on a test and preparation for success beyond the test, in the domain to which the test is intended to generalise and to which it may control access."

Some researchers take a wider view, looking beyond teaching and learning. The term used to refer to studies of this type is *impact,* defined by Wall (1997: 291) as "any of the effects that a test may have on individuals, policies or practices, within the classroom, the school, the educational system or society as a whole." Saville (2010: 2), following Hamp-Lyons (1997: 299), sees test impact as a "superordinate concept" which includes washback.

A different approach to the relationship between tests, teaching and learning is to consider the impact, or influence, the test has as one of the qualities of the test. Bachman and Palmer (1996: 17) propose that "the most important quality of a test is its usefulness" and offer a model of test usefulness which includes the qualities of "reliability, construct validity, authenticity, interactiveness, *impact* and practicality" (ibid. emphasis added). This model is governed by three principles, which view test usefulness as a balance of these six qualities, all or any of which should be considered only in terms of the test in question and the context in which it is implemented. Thus when designing a test, Bachman and Palmer urge careful consideration of the consequences which may come about (1996: 35) as a result of the uses of the test results.

Messick, by contrast, places washback within the concept of validity. Rather than listing different types of validity, he unifies them by considering the "content, criteria and consequences" (1995: 742) of a test. The notion of washback is then placed within the consequential aspect of validity. Weir (2005: 44–49) extends the notion of *consequential validity,* placing it within a socio-cognitive framework, where he examines how the test is responded to in a given context. Part of this response is how scores are interpreted, one aspect of which for him is washback. Morrow (1986: 6) uses the term

"washback validity" which he sees as the quality of the effect a test has on classroom teaching (in Alderson & Wall 1993: 116).

In general education Fredericksen and Collins (1989) saw tests as part of an adaptive dynamic system of education and were concerned with issues of validity within this. As a dynamic system is always shifting and adapting to different forces, it is to be expected that the system will respond to external examinations. If there is a good match between what the test is aiming to measure and targets set out in the curriculum, then this shift in the system can be positive, as it will encourage development of everything described in the curriculum and the test can be considered *systemically valid* (1989: 29.) However, if, for example, indirect testing techniques are used, rather than testing curriculum target skills directly, then the changes effected by the examination may be quite different from those intended. Fredericksen and Collins (ibid.) offer the example of proofs in geometry to illustrate this problem. If the examination requires learners to reproduce geometric proofs, then this could lead learners to simply memorize all the proofs which might occur on the test. This would be systemically valid if the curriculum target was "to be able to reproduce proofs". If, however, the curriculum aim was for learners "to be able to develop and use geometric proofs", then the test could be seen as adversely affecting the system. The term *systemic validity* therefore refers to the role of the examination in the educational system as a whole from the perspective of interrelationships between the examination, the curriculum and teaching.

1.2. Washback

In this section we will begin by looking at how researchers have attempted to define the washback phenomenon through the use of models and how different studies have expanded and refined understanding of the processes it entails. We then move to a review of research conducted in washback looking at the effects of tests on learners, teachers and on learning and teaching, and examining some of the factors which affect whether, or to what extent washback takes place.

1.2.1. Models of washback

Many models of washback or impact have been offered and as these tend to reflect the level of knowledge about the phenomena at the time, we will view these in chronological order.

In language testing one of the earliest models of washback is that of Hughes (1993), in which he explains that a test may have influence on either the *participants*, the *process* or the *product* of teaching and learning. He views this as hierarchical, with the test first affecting the "perceptions

and attitudes of the participants towards their teaching and learning tasks" (Hughes, 1993: 3), which in turn may affect the process of how they teach or learn, and which may then have bearing upon the product of that process.

In order to effect "beneficial washback" Hughes (1989: 44–47) suggests that what is tested should be what it is hoped will be improved (e.g. including oral skills in a test to promote spoken English) and that skills should be tested directly. Tests should be criterion-referenced and linked to teaching and learning objectives. Teachers and learners need to understand the test thoroughly and, if necessary, support should be made available for teachers.

Shohamy (1992: 513) offers a "diagnostic feedback testing model" for school based assessment of foreign language. Her concern is for tests to be used for positive effect, to improve the teaching and learning process, and for this the information obtained from a test is seen as of key importance. To this end she makes the following recommendations. The test should be *communicative*, test authentic language use and follow current theories in second language acquisition. A single score does not inform the test users, rather a *detailed profiling* is needed to support *diagnosis* of strengths and weaknesses. The test needs to be clearly *connected with teaching and learning*, so that information from the test results can help change instruction. The *teachers and school directors* need to be *actively involved* in the assessment process, as it is they who will implement any changes (Nitko, 1989; Nero, 1989). The test information should be *both criterion and norm referenced*, thus give information about attainment of curriculum targets (criterion referencing), but also allow for comparison with other schools (norm referencing). *Support* should be provided to help teachers respond to the information given by the test. Shohamy, while addressing participants, process and product as in Hughes's model, sees the test as part of a cycle, with clear correspondence between curriculum goals and content, test content and results and the teaching and learning process. Commonalities can be seen in what is thought to have positive effects. Both Hughes and Shohamy consider the *positive* potential of washback.

Alderson and Wall (1993), in a seminal paper, explore the concept of washback from a theoretical and philosophical perspective, debating first whether in fact it exists or is simply a metaphor, and then discussing how to determine the bounds of what could be considered the influence of a test. They reject the notion that a test by itself is enough to affect how teachers teach or how learners learn (1993: 118) on the grounds that the test takes place in a setting where a large number of other factors come into play. They also suggest that the relationships between different factors are not easily predicted and are most likely highly complex (1993: 119), rejecting a simple cause-effect model. On the basis of this discussion, and drawing on research done to date, they propose fifteen possible hypotheses, as follows:

1. A test will influence teaching
2. A test will influence learning
3. A test will influence **what** teachers teach and 4. **how** teachers teach
5. A test will influence **what** learners learn and 6. **how** learners learn
7. A test will influence the **rate** and **sequence** of teaching and 8. **rate** and **sequence** of learning
9. A test will influence the **degree** and **depth** of teaching and 10. the **degree** and **depth** of learning.
11. A test will influence attitudes to the content, method, etc. of teaching and learning
12. Tests that have important consequences will have washback
13. Tests that do not have important consequences will have no washback
14. Tests will have washback on **all** teachers and learners
15. Tests will have washback effects for **some** learners and **some** teachers**,** but **not** for others.
(Alderson & Wall, 1993: 120–121)

Alderson & Wall emphasize, however, that these hypotheses are not meant to be exhaustive and may need additional elaboration as more evidence is collected. Their article began a quest to delineate a model of washback, which continues to the present day. Their call: "we need… to identify and examine cases where washback is thought to have occurred, and to see how and why it did or did not occur" (1993: 121) and their stress of the need for empirical data has shaped much of the subsequent research in the field of washback in applied linguistics.

Bailey (1996: 264), in a paper written for the test provider Educational Testing Service (ETS), draws on the work of Hughes and Shohamy to build a "basic model of washback." She lists Hughes' participants (students, teachers, materials writers, curriculum designers, and researchers) and links these to associated products. The relationships between these elements is viewed as more complex than envisaged by Hughes, with new materials, or curricula, feeding into teaching and learning, research results informing both teaching and the test itself and all of the other products also potentially having bearing on the test.

With reference to Alderson and Wall's fifteen hypotheses (1993: 120–121) Bailey predicts processes that learners may engage in (1996: 264) as the result of being faced with a test. These include undertaking specific learning tasks, making use of test-taking strategies, seeking additional help or instruction, or redirecting time and effort away from other activities to allow for test preparation. Bailey, like Alderson & Wall, predicts that washback may have *negative* effects, in certain circumstances.

One of the areas which may be negatively affected, again extrapolating from Alderson and Wall (1993), is that of the teaching programme. Shohamy (1992: 514) warns of potential "narrowing of the curriculum" if there is over-emphasis on student test scores on external examinations, rather than on promoting meaningful learning. The teacher will focus time and attention

on those aspects which will appear on the test and ignore, or skimp over, other areas. We are reminded of the importance of ensuring a good fit between the test requirements and curriculum targets.

Watanabe (1996), investigating the relationship between university entrance examinations and teaching approaches used in test preparation courses in Japan, found that *teacher factors* mitigated the nature and extent of washback. These factors were related to the teacher's "educational background and/or experiences" (1996: 330) and "beliefs about effective teaching methods" (ibid.).

An additional factor hypothesized by Watanabe was that of *the proximity of the test*. Where the test was imminent, a closer match was found between teaching style and the test questions than where the test was more distant. This provided initial evidence in support of Alderson & Wall fifteenth hypothesis. Further support for this was provided by Alderson & Hamp-Lyons (1996), who found that whether a class in the US was a Test of English as a Foreign Language (TOEFL) preparation class or not affected how the teacher taught and what they taught, but that the effect of the test was outweighed by individual differences in the teacher's teaching style. They hypothesized that other factors, such as size of class and materials used were also contributory factors. As a result they proposed a re-working of the fifteenth hypothesis to read: "Tests will have different amounts and types of washback on some teachers and learners than on other teachers and learners" (1996: 296). They also added refinements to the original hypothesis regarding amount and type of washback: that it will be influenced by *the status of the test* (high or low stakes), "the extent to which the test is counter to current practice", and proposed additional hypotheses about consideration of the choice and design of test preparation materials and methods teachers use, and how open teachers and materials writers are to innovation (ibid.) Shohamy, Donitsa-Schmidt & Ferman (1996), confirmed that test status was a factor, as was status of the language being tested and the purpose of the test. Like Watanabe, they also found test proximity to be a factor affecting teaching behaviour. They hypothesized that test characteristics, such as the way skills are tested, also influence the type and extent of washback (1996: 316)

Cheng (1997), investigating the influence of the introduction of a revised syllabus and new examination system in Hong Kong secondary schools, introduced the notion of *washback intensity*, which she defines as "the degree of washback effect in an area or a number of areas of teaching and learning affected by an examination" (1997: 43). This was in response to the fact that during her study she found that some areas, such as the content of the teaching and choice of course books, were much more strongly affected by the changes than others (1997: 49–50).

Burrows (2004: 125), like Andrews (1994a) and Heyneman (1987: 260), found *choice* an element of washback theory. Based on observations of

four teachers who had been identified from survey and interview data as having differing responses to the implementation of a new classroom based assessment programme, she found that a teacher could make a *conscious decision* whether a new system would have a washback effect on their teaching or not. A teacher could decide to adopt a new system as proposed, or resist it altogether. Alternatively, they could take on only parts of it, or change it in some respects. On the basis of this, drawing on references in the literature on educational change (Markee, 1997; Fullan & Park, 1981; Fullan with Stiegelbauer, 1991; Hargreaves, 1994) and work done by Wall (1996), Burrows proposed a new model for washback within curriculum innovation, focused on the teacher, which takes into consideration the teacher's beliefs, assumptions and knowledge (BAK), (Woods, 1996: 195). It goes beyond earlier models of washback, which proposed that teachers respond in different ways to a new test or assessment scheme, by suggesting that patterns may be predicted in the teacher responses, relating to their BAK and their feelings about the new test or assessment. She draws parallels to models of educational change (Markee, 1997; McCallum et al., 1995) and particularly to work done in curriculum innovation.

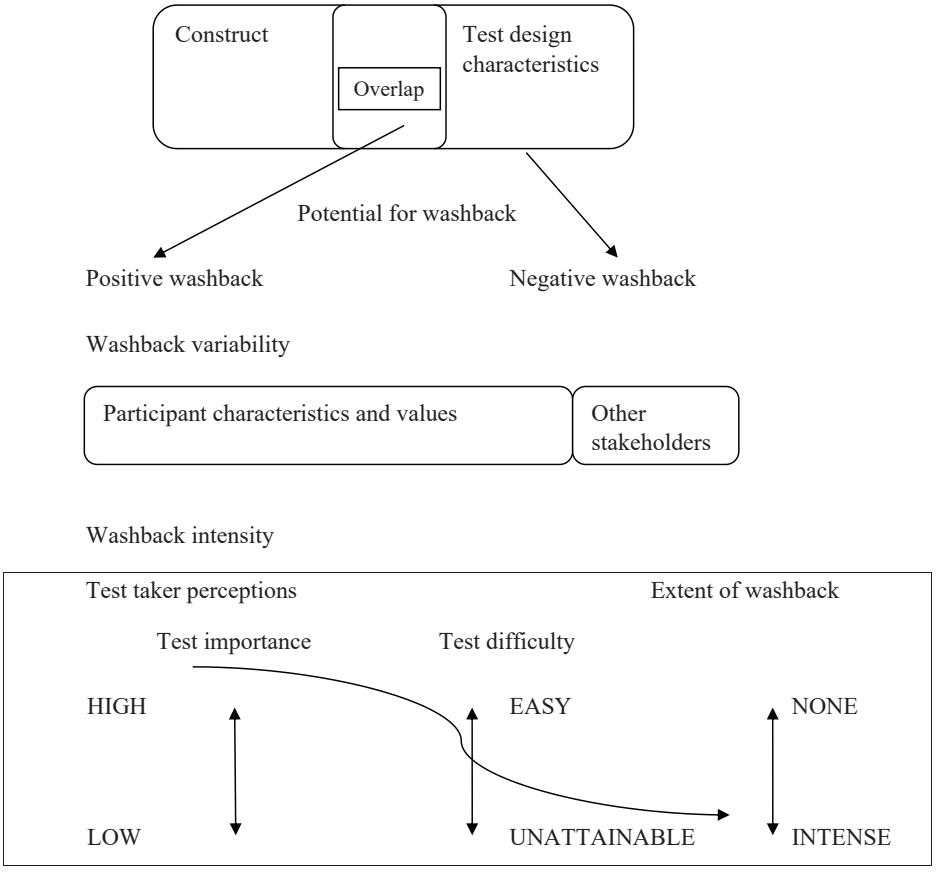
In 2004, in the foreword to a collection of papers on washback, Alderson, looking back on a decade of research since the 1993 Alderson and Wall "Does washback exist?" article, stated that there is now sufficient data to show that washback definitely exists, but that now the challenges are to answer the questions "what does washback look like?", "what brings washback about?" and "why does washback exist" (2004: ix). Subsequent research has consequently started trying to find ways to explain how washback works and relationships between the many factors which have been identified.

Green (2007) builds a model of washback which takes into consideration all of the aspects which have been described above, that is *washback direction* (i.e. negative or positive), the *stakes* of the test, *washback intensity* and *variability* (which includes characteristics and values of the participants) (2007: 12–25).

Green's model has three layers (see Figure 1.1), the first of which considers that whether a test will have negative or positive effects can be foreseen by examining the relationship between the *test construct* and the curriculum with which the test is associated. By test construct we understand how the test designers view language and how they apply their beliefs about language in practice (Alderson, Clapham & Wall, 1995: 17). For example, if the examination is to assess a learner's reading ability, this means the test designer must decide what the notion "reading" means (e.g. when, what, why, how, who), what different skills it entails, and in what ways these can be tested. The construct will consequently have direct influence on the content of a test. Where there is a close match between the construct of a test and the construct of the curriculum and the content of a test and the content of the curriculum, the potential is that the test will have a positive effect on the

teaching and learning process (Resnick & Resnick, 1992; Green, 2007: 13). However, a good construct/content/ curriculum match alone is not enough, a key factor is also the way in which the content and skills included in the examination are tested. Where the way they are tested closely reflects how they are taught, then there is greater potential for positive washback (Weir, 1993: 28). Where there are differences between the test and the curriculum, either in terms of construct, content, or in the application of theory in practice, then there is potential for negative washback. Likewise, if the way the content or skills tested differs from the way these are usually taught, then it is likely that teaching will shift to mirror the way these areas are tested (ibid.). This is a more complex relationship than that of construct/ content/curriculum, as it can be deliberately exploited with the intention of implementing change in classroom practice (Popham, 1987; Pearson, 1988). This use of tests for conscious manipulation will be explored fully in Chapter Two. Here it is enough to say that where there is mismatch, there is potential for negative washback, although this may not necessarily be the case.

Figure 1.1. Green's 2007 Model of washback (based on Green 2007: 24)



In the second layer of the model, washback variability, Green includes aspects of the participants which have been found to influence washback. These include the teacher's knowledge and understanding of the requirements of the test (Wall & Alderson, 1993: 67; Chapman & Snyder, 2000; Qi, 2005); and "acceptance of test demands" (Green, 2007: 24) with reference to findings in Alderson & Hamp-Lyons, 1996; Watanabe, 1996; Burrows, 2004; Cheng, 2005; and Wall, 2005 which refer to factors in how teachers perceive the test. The work of these researchers has shown that there is considerable variability in how different teachers respond to a new test. Also included as a factor at this level of the model are "resources to meet test demands" (Green, 2007: 24) which refers to work done mainly in developing countries (Chapman & Snyder, 2000) where the washback from tests was found to be influenced by a lack of resources to meet the new requirements, either in terms of material resources (e.g. recording equipment to teach listening skills), or in terms of resources in the teacher, such as skills and language ability.

The next layer of the model, concerns washback intensity (Cheng 1997, discussed earlier). Within this layer Green includes aspects of how stakeholders perceive the importance of the test (see e.g. Shohamy et al., 1996 discussed above), or the importance of the information the test gives to the test-taker (Crooks, 1988; Black & Wiliam, 1998). Attitudes of teachers towards a new test, and what they consider its implementation may entail, have been seen to affect the intensity of washback from it (e.g. Smith, 1991; Burrows, 2004), as have their own training or skills (Wall & Alderson, 1993; Watanabe, 1996). A second aspect included in this layer of the model is perception of the difficulty of the test on the part of learners and their teachers. Crooks (1988) suggests that a test should be perceived as presenting an attainable challenge, as if it is thought to be too easy it will not have a motivating effect (Mehrens, 1998). Green (2007: 25) therefore claims that washback will have the greatest intensity when the test is perceived as at an appropriately challenging level. He also predicts that the intensity will be at a maximum when test outcomes are seen to be of greater importance than overall development of language and when these attitudes prevail in the teaching-learning context. The model also incorporates the notion of *seasonality*, expecting that washback intensity will increase as a test nears (Bailey, 1999; Watanabe, 1997) and may change over time as the test becomes accepted as a norm (Shohamy et al., 1996, Wall, 1996).

Shih (2009: 199) offers a model of washback on teaching to extend the one proposed by Burrows. When investigating the General English Proficiency Test (GEPT) in Taiwan, she found that extant models of washback were inadequate to explain her findings. She compared classes in two university departments, in one of which taking the GEPT was a requirement, and found that "the objectives of the course and the relation of the course to the school's policy" (2009: 198) were key in deciding whether the test would

have influence. Like Burrows, she found evidence that teacher beliefs were a factor, but added to this the more general notion of how teachers felt about the role of tests in teaching and learning and about the GEPT in the particular context. An additional factor suggested are the consequences of the test results for the teacher. If evaluation of the teacher, or decisions about their employment, is made on the basis of test outcomes, then teacher accountability will contribute to washback. By contrast, however, Shih suggests that if the test results have no impact on the teacher, then even if the test is seen as important to the learners, the effects will be lessened. The aspect of accountability will be explored more fully in Chapter Two.

Shih aims to offer a fuller dynamic and interactive model, synthesizing previous research findings, including contextual factors and demonstrating their interplay. Contextual factors are divided into three areas: National, social or broader educational factors; School-level factors; and Course-level factors. The next areas in the model are Test factors and Teacher factors, both of which are affected by the contextual factors. It is also suggested, based on Shohamy et al. (1996), that all aspects of "washback phenomena may evolve over time" (Shih, 2009: 200). All of these areas impact on teaching in terms of its content, methods used, school assessment practices, how the teacher talks, time allocated for test preparation, homework assigned, atmosphere in the lesson and levels of stress and anxiety (2009: 199).

To complement this model, Shih (2007: 151) offered a model of washback on student learning. Like the model of washback on teaching, it covers the context, but this is described in terms of factors extrinsic and intrinsic to the learner. The extrinsic factors include socioeconomic, school and educational factors; family, friends and colleagues; and personal factors. Intrinsic factors include individual differences, personal characteristics and personal perceptions of the test. Test factors are also listed. All of these areas impact on the "students' learning and psychology" (2007: 151), defined as the content of learning, time spent learning, test anxiety, motivation for and strategies used in learning (ibid.). Additional factors included in the model are the "results of the test" and 'subsequent learning" (ibid.) and, as with the model of washback on teaching, the factor of time is added to allow for changes taking place progressively. The complexity of these two models reflects how knowledge and understanding of washback has developed since the 1993 hypotheses.

In an attempt to understand more about the mechanisms of washback Xie (2011, 2013) and Xie and Andrews (2012) investigated the relationships between learner factors and how they prepared to take a test. Xie & Andrews (2012) applied expectancy-value motivation theory to a study among college students preparing for a high-stakes English proficiency test (College English Test Level 4, CET4) in China. Structural equation modelling enabled them to find a direct relationship between the skills which learners perceived were needed in the test and the strategies they used to prepare

for it. If learners had good understanding of what the test entailed this had an influence on their preparation strategies. Learners' perceived value of the test and its purpose explained 24% of the variance on test preparation.

This study can be seen to expand understanding of the effect of washback on student learning. It helps to indicate some of the nature of how the washback mechanism influences learners through a series of "filters". First, learners have to understand what skills are needed to take the test and perceive that they could succeed in these skills. A second filter applies, which is learners' perception that these skills are worth the investment. In this way we begin to understand why washback applies in some contexts and not in others, and varies from learner to learner. Research by Qi (2005), which found a mismatch between the skills the test designer intended to test and the skills teachers perceived as necessary to take the test, could be similarly interpreted, with the filtering effect of teacher perceptions explaining the diminished effects of the intended positive washback.

In another attempt to amplify understanding of the mechanisms of washback Zhan & Andrews (2014), in a case study of three non-English major students, investigating out of class preparation for CET4 (see above), applied Dörnyei's (2005) theory of possible selves. They found a relationship between the type of washback affecting an individual and how they envisioned themselves as a CET4 test-taker (p. 71). This suggests that another of the 'filters' may be related to the relationship between test-taker self-perceptions and the test.

The complex nature of these recent findings goes some way towards providing possible evidence for the model of washback proposed by Shih (2007).

Models have progressed from a simple linear cause-effect design, towards much more complex inter-active, multidimensional models and most recently to proposals of models of dynamic systems. As the basic tenet of a dynamic system is that every component part in it can relate to any other part of it and cause reactions in any direction and of any strength which have the effect of producing new configurations that may be quite unexpected (Thelen & Smith, 1996: 54), it is difficult to draw a model of a dynamic system of washback. What can be done, however, is to list, on the basis of washback models proposed to date, components which are most likely to be parts of the system. In the next sections of this chapter we attempt to show these components in a series of tables, which sum up findings from research.

Defining washback for the purposes of this study

This study takes washback as the influence that external standardised tests have on teachers and learners and on teaching in the classroom and on the learning process both in the classroom and outside it (Hamp-Lyons, 2000: 586; Wall, 1997). We restrict the notion of washback to the influence

of tests themselves and not the workings of the results of those tests, which we will place in the macro context and interpret as part of a wider educational system (Bachman & Palmer 1996: 30). Washback is consequently seen as a finely tuned part of this wider system of influence, which we will refer to as test impact (Wall, 1997; Shohamy, 2001; Green, 2003: 6 in Hawkey 2006: 7). The macro context will be considered below in the section referring to impact and the uses and effects of test results in Chapter Two.

1.2.2. Review of washback research

Several reviews of research into washback have been conducted. One of the earliest of these, an extensive insight into the field of general education in the USA, was Kirkland (1971), entitled "The effects of tests on students and schools." Despite this being a very early study, the issues raised in it continue to be key concerns and for this reason will be outlined here. One important point raised is that although a great deal is written about the effects of tests there is relatively little empirical evidence available to support such claims (Kirkland, 1971: 306). We shall see that this continues to be the case throughout the literature (e.g. Cheng, 2004: 148).

Kirkland (1971)

Kirkland separated the research under review into three areas: the effects of tests on learners, on teachers, and on the working of schools. First considered are the effects of tests on learners. How a test is perceived by the learner, by stakeholders and by society in general is thought to influence the test taker's self-concept. Where a test is trusted as providing accurate results, the learner may make judgements about themselves on the basis of the scores they obtain on the test. However, test scores may be only part of the picture, with other factors influencing how learners decide on their level of achievement (Goslin, 1967).

There is some evidence that when the test score was lower than the learner expected, this had a negative impact on their self-image (Hills and Williams, 1965). Brim et al. (1969) found that attribution of cause played a role in the interpretation of test scores, with those who believe in their own effort, those who have high self-esteem and self-awareness getting more useful information from the test results than those who attributed scores to fate (op. cit.: 309).

How learners performed on tests in the past has an effect on how they expect to perform on subsequent tests and low past performance is associated with lower scores later (Feather, 1966; Sears, 1940; Kagan, 1968). Learner expectations were also found to influence how long learners are prepared to work on a problem task. Where learners expected to succeed on a test they worked longer with more demanding problems than those who had low expectations (Battle, 1965; Feather, 1963, 1966).

Learner expectations of results can also be negatively influenced by teachers (Smith, 1952) and how teachers "label" learners. Osler (1954) found that what teachers told learners about their level of ability influenced test results. With randomly assigned learners, where learners were told they were in the lowest ability group, their test scores were found to be the lowest, and where learners were told they were in the most successful group their scores corresponded. The researcher concludes that it is how the learner perceives their ability that is the decisive factor. A study by Brookover, Shailer and Paterson (1964) had similar findings.

Self-concept has been found to be associated with a sense of achievement. Where learners experience success over time they have fewer emotional problems, while learners who repeatedly fail are more likely to be disturbed (Glidewell & Stringer, 1967; Modu, 1969).

While the relationship between motivation and testing has been the subject of much research, findings, according to Kirkland (1971: 312), are inconclusive. However, there is evidence that the learner's attitude towards the test and how they will perform on it influences test outcomes. Where learners feel they are sure to fail their results are negatively affected (Anastasi, 1968; Anastasi & Cordova, 1953; Wrightstone, 1963).

The information learners receive from doing a test motivates learners, on the condition that they are positively oriented towards the test (Pollaczek, 1952; Wrightstone, 1963). Where learners were encouraged to use a key to check their own responses, discuss answers and score themselves, they scored higher on later tests (Curtis, 1944; Flock & Saggar, 1968; Sassenrath & Gatherick, 1965). The amount and type of feedback received also affects test results, with specific comments given by the teacher found to be associated with higher scores on subsequent tests (Page, 1958).

A number of studies indicate that there is a relation between self-concept, motivation and achievement on tests, which is described as level of aspiration. Where learners experience success they receive positive responses from those around them, which contributes to a positive self-image, which, in turn, increases the learner's expectations of further success and confidence in their ability to achieve (Atkinson, 1964; Atkinson & Feather, 1966; Crandall, 1963; Taylor, 1964). Sewell & Shah (1968), found parental encouragement to be associated with higher learner aspirations.

The type of test task affects how learners learn and prepare for the test, with tests of factual information leading to rote memorization, as opposed to tasks where learners are required to write an essay which led to higher order skills such as organisation and finding relationships between pieces of information (Cook, 1955; Eurich, 1931; Merkhofer, 1954; Meyer, 1935; Terry, 1933).

A large number of studies reported in Kirkland (1971: 318–19) concern the question of anxiety and testing. Although findings are conflicting, generalizations can be made (Ebel, 1965; Goldman, 1961; Hill & Sarason,

1966; Ruebuch, 1963; Sarason et al., 1960). It should be remembered that the studies referred to were conducted in the USA. High anxiety is generally found to have a negative effect on test performance, although not for all learners or in all situations. However, moderate levels of anxiety may enhance test performance. When a learner considers a test important and has high levels of test anxiety, this is more likely to be associated with a negative effect on performance than when the learner is less anxious. Anxiety about tests is related to ability, with less able learners more likely to be more anxious. Test anxiety was also found to be associated with socio-economic status, ethnicity, level of peer acceptance, emotional issues and living in a rural community. While anxiety and gender were not found to be associated among primary students, girls in junior high were found to be more anxious about tests than boys. How students rate their levels of anxiety and how these are rated by their teachers or psychologists is not consistent. As learners progress up the school system levels of anxiety increase. Test anxiety is related to some personality factors, in particular to negative self-concept.

When considering the effect of tests on parents, Kirkland (1971: 330) notes a lack of empirical research, which does not allow many generalisations to be made. However, Maller & Smallenberg (1963) suggest that parents are concerned that testing in school should not have negative consequences for their children.

A summary of the effects of tests on learners will be given in Table 1.3, where effects found in studies of language tests on learners are joined to these effects noted by Kirkland.

Reviewing the effects of tests on teachers Kirkland (1971: 330–335) reports research from the 1930s (in Tyler, 1966) where teachers began to teach according to the objectives of an examination, rather than what was specified in the curriculum. Brickman (1946) reported that teachers exchanged the course book they used previously, for test preparation material and past test papers in response to an external examination. Conversely, research also indicates that some teachers claim not to be affected by tests. Traxler (1958) and Sandler (1959) found little effect on methods of teaching. Goslin (1967) found little evidence of the impact of testing on curriculum, although the same study found that 40% of teachers reported explicit teaching of test-taking strategies and that 18–32% of teachers (depending on level of school) said they had changed their teaching methods. A national survey in the USA found that the majority of teachers felt tests had a positive effect on their teaching programmes (NEA, 1962). Teachers show a tendency to align the type of tasks they use in class to the types of tasks appearing on standardised tests (Goslin, 1967). Kirkland concludes (1971: 335) "tests have relatively little influence on what is taught in the schools, or how it is taught, except in the case of tests used for college admission."

In reviewing the effects of tests on schools (1971: 336–338) Kirkland reports some influence on teaching and programmes, which is generally felt to be positive. The use of test scores to compare schools is seen as increasingly expected by stakeholders.

To conclude this overview of Kirkland's (1971) review of the effects of tests and testing, relationships were noted between testing and its effects on learner attitudes, psychology and motivation. Some effects were also noted on teachers and schools, although these, at the time, appeared to be of less concern than the influence of testing on learners. It will be of interest to see how these findings change with time and with the increasing incidence of the use of standardised testing.

Taking Kirkland's division of the effect of testing on learners, teachers and schools we will now examine how research develops in both general education and in research on the testing of foreign languages. Teachers and learners will be dealt with first and the effects of testing on schools in the section on impact.

Effects of tests on teachers

Whether an examination will affect the work of a teacher and so bring about washback has been widely researched. Many different factors have been found to have effect. In Table 1.1 is a summary of empirical research conducted on factors in teachers found to have influence on whether washback takes place, its extent or intensity. Research written in *italics* is from the field of general education. All other research applies to the teaching of foreign languages.

Table 1.1. Empirical research on washback: teacher factors influencing washback (italics indicates research studies in general education)

Teacher beliefs
Beliefs about effective teaching Watanabe, 1996 Beliefs about assessment Burrows, 2004; East, 2014 Beliefs about test preparation Alderson & Hamp-Lyons, 1996
Teacher attitudes to their work
Teachers' openness to change Alderson & Hamp-Lyons, 1996: 296
Attitude to teaching/exam preparation materials Wall & Alderson, 1993; Andrews, 1994; Alderson & Hamp-Lyons, 1996: 286; Read & Hayes, 2003; Cheng, 1997; Wall & Horák, 2006

Teacher and the test
Understanding of test Wall & Alderson, 1993: 67; Chapman & Snyder, 2000; Qi, 2005; Cheng, 1997 Knowledge about test Wall & Alderson, 1993: 67 Perceived quality of test Shohamy et al., 1996; *Smith, 1991*
Perceived importance of test Shohamy et al., 1996
Perceived expectations of learner performance on the test Chen, 2002
Teacher experience and skills
Teacher education or training Watanabe, 2004; educational background Watanabe, 1996 Training for new test/assessment implementation Munoz & Alvarez, 2010; Wall & Anderson, 1993
Learning experience Watanabe, 2004
Length of teaching experience Lam, 1994; Shohamy, Donitsa-Schmidt & Ferman, 1996
Ability in the target language taught Qi 2005; Shih, 2009
Familiarity with range of teaching methods Wall & Alderson, 1993: 67

Factors in the teacher found to affect washback

It becomes immediately clear from the number of factors in the teacher thought to have an influence on whether and to what extent washback is likely to occur (see Table 1.1), that the phenomenon is complex. Care needs to be taken in generalising on the basis of these studies, even though they all use empirical data, as they vary considerably in terms of size of sample, research methodology and research context.

One useful way of summarising many of the factors in the teacher found to play a role in washback, is to consider the beliefs, attitudes, knowledge of the teacher (Burrows, 2004). Beliefs about what constitutes effective teaching (and effective ways of preparing for a test), beliefs about the role of the examination and its place in the teaching-learning process, which may be tempered by the teacher's perception of the status and significance of the examination (Shohamy, Donitsa-Schmidt & Ferman, 1996), or by what the teacher perceives as their learners' views on the examination (Alderson & Hamp-Lyons, 1996), and the teacher's philosophy of how languages are learnt, are all factors which filter if, what, how or how much washback will take place.

Let us consider some examples of the effects of teacher beliefs. Watanabe (1996), studying whether a university entrance exam in Japan promoted use of grammar-translation in secondary school, observed differences in the approach to the teaching of grammar between two teachers in a Japanese school teaching exam preparation classes. One teacher claimed that they felt effective teaching required explanation and translation of grammatical issues in a text to enable learners to have full understanding, while the other felt that enabling the students to achieve communicative competence was of prime importance. This showed that what the teacher believed was an effective way to teach had a greater effect than the exam itself. East (2014), in a study of how teachers in New Zealand secondary schools were responding to the introduction of new high-stakes continuous assessment of foreign language speaking, found that many teachers believed the new assessment process was more useful than a previous scheme, with suggestions that some felt it was more authentic and more reflective of communicative methodology than the former summative testing. Teachers, however, had concerns about practical implication, which East interpreted as indicating a lack of understanding of the construct of the continuous assessment. Intended to be used to collect a series of samples of spontaneous learner-learner interaction during normal activities on the course, it seemed that some teachers were viewing the assessments rather as a series of tests of speaking, set apart from the normal teaching process. In other words, the teachers' experience of how speaking had hitherto been assessed may have inhibited understanding of the new scheme, or they were interpreting the new scheme from the perspective of their perceptions of summative, rather than continuous assessment. Their beliefs about how speaking should be assessed were impacting on their views and understanding of the new assessment scheme.

Teacher attitudes are the next area to consider. One of the most extensive areas investigated is what teachers think about the examination, or assessment practice, in question. Within this can be distinguished whether the teacher perceives the test as a reliable and valid measure, and whether it reflects the curriculum they are expected to teach. Whether the teacher feels comfortable with the proposed changes and considers the new requirements feasible in their context is another area of concern. Cheng (1998) reported teachers were worried about a new examination in schools in Hong Kong and about how their students would perform. Kiss-Gulyas (2001) reporting teacher responses to a new matriculation exam in Hungarian schools, found similar fears. Shohamy et al. (1996) reported teachers in Israel expressed negative feelings towards a test of Arabic as a foreign language, as they considered the language to be of low prestige, and the test to be of dubious quality and have issues with validity. By contrast, they had more positive feelings towards a test of English which they considered a language important, useful and relevant for their learners. They were, however, anxious about their learners' performance and results.

Burrows (2004), in her study of the effects of the introduction of a new assessment programme for adult migrants in Australia, found that some teachers claimed that they had made no changes in their practice on the grounds that they found that the new programme offered nothing positive, while others with different perceptions of the programme had adopted the new ways of assessment. Alderson and Hamp-Lyons (1996) found that many, although not all, of the teachers in their study were negative to the TOEFL and preparing their learners for it and that exam preparation classes were more serious than general English lessons. Contrast between exam and non-exam classes was also noted by Read & Hayes (2003) in their study of IELTS. Both Watanabe (2000) and Read & Hayes (2003) reported that the atmosphere in the exam preparation lesson seemed determined by the attitude of the individual teacher and could be positive, while Wall & Horák (2006) found generally positive attitudes from teachers working in TOEFL preparation classes. Smith (1991) noted that teachers in general education in the USA were highly sensitive to the results of external tests and would take action to avoid negative consequences caused by low scores. She also reported that teachers found learners were stressed by the tests, which in turn affected how the teachers felt.

In order for a teacher to accommodate a new examination in their teaching process they first need to understand its construct, which involves knowing about its content, the ways in which different aspects of language and skills are tested, how these will be assessed and according to what criteria. Chapman & Snyder (2000) reported that teachers in Uganda had difficulty with a new examination as they either appeared unable to understand the changes it demanded of them, or were reluctant to introduce new teaching tasks in class. Qi (2005) found a mismatch between what teachers in China perceived was being tested in the National Matriculation English Test (NMET) and what was intended by the test designers, which meant that the anticipated change to a communicative teaching approach did not take place.

The level of understanding of the new examination may be influenced by the teacher's own experiences, both as teacher and learner, and by their education. Watanabe (1996), investigating two teachers in his case study of the washback from a university entrance exam in Japan, found that how teachers taught grammar in high school seemed to be related to their educational background. The teacher who held a postgraduate degree in linguistics had a tendency to explain grammar and use translation regardless of whether the course was to prepare for the exam or not, while the second teacher, who had been a teacher in high school, translated and explained more in the exam preparation classes, seeming to cater for the differing needs of students, as was expected in school. Watanabe also suggested that the teachers' own experience may have been a contributory factor, as the second teacher had graduated from the university for

which he was preparing his students to take the entrance exam. The first teacher by contrast, had attended a more prestigious university and would consequently have had different expectations, based on his own experience. Lam (1994) noted a difference in the washback effect from a university entrance exam in Hong Kong between more and less experienced teachers of English. Teachers with more years' experience were found to be more affected by washback than less experienced teachers and to respond to the examination with less innovation, relying more heavily on course books. They also showed more tendency to *teach to the test*, using past papers for test-specific practice.

The skills which the teacher has either for the teaching of the subject or, in the case of language teaching, in the foreign language itself have been found to be factors which influence washback. Watanabe (2004), in a study of the washback from university entrance exams, found that how teachers decided to teach in Japanese high school related to their own teaching and learning experience, with one teacher, for example, commenting that they did not know how to teach listening well and so were relieved that it was not part of the university entrance exams for which they were preparing their learners, thus justifying the fact that they spent little time on it in class, while another teacher explained that the ways they chose to teach were those that they had been taught by, implying that this was the primary factor influencing their teaching methodology and not the examination. In both these cases factors in the teacher appeared to mitigate the effects of washback from the examinations. Qi (2005), investigating the National Matriculation English Test (NMET) in China, found that the fact that teachers themselves appeared to have little experience of using English communicatively seemed to contribute to a misunderstanding of the aims of the writing task, where appropriate use of language in context was intended to be a criterion for rating the task. Teachers in the classroom were found to focus more on accurate use of language than on appropriacy, as this was what they were more familiar with.

A new examination may mean that the teacher will need to have certain skills in their teaching repertoire, such as how to organise pair and group work in class, or will mean that they need to develop these skills to meet the new challenge. This training or education aspect may also refer to the support a teacher is given to help them cope with the new examination. Wall & Alderson (1993) felt that if teachers in their study of the impact of the new O Level examination in Sri Lanka had received adequate training in how to adapt their methodology to teach the new national course books and had been given fuller information about the intentions of the examination, then they would have been better equipped to meet the aims of the new programme. Munoz & Alvarez (2010), in a study of washback following the implementation of a new programme of classroom oral assessment in Colombia, found that teachers in an experimental group who had received

support and training on assessing speaking were observed to have adopted practices more closely allied to the intentions of the test designers and appeared to have a better understanding of the aims of the test than those in a control group where no such support had been made available.

The teacher's response to the test may also be affected by their perception of how they think their learners will perform on it. Chen (2002), investigating washback from the Basic Competency Test (BCT) for competitive entry to upper secondary school in Taiwan, found that where teachers considered their learners would do well on the test, there was greater evidence of washback in their course and lesson planning.

Evidence of washback in teaching

Washback in the classroom can take many forms, the most common of which have been found to be in the content of what is taught (see Table 1.2). The teacher may make decisions to change or adapt the content of the syllabus, or scheme of work, they use to reflect what they perceive to be the focus of the test, referred to as *aligning the syllabus to the test*. Ferman, (2004) for example, found that changes in the oral component of the Israeli matriculation exam in English, intended to promote the development of communicative proficiency, led to a significant increase in the focus on oral activities in class and also to changes in the organisation of lessons, with more pair and group work noted. It was reported, on the basis of survey and interview data, that should the oral component be removed from the test teachers would stop teaching oral skills. In this context such alignment of what is taught to the test is an indication of positive washback, as the changes noted in teaching were the changes the authorities intended by introducing the revised exam. However, Ferman also found that the oral exam impacted on the selection of texts for reading. Teachers deliberately chose literature which would be easier for students to talk about, as oral reports on readings from the course formed part of the new exam. In this way the teachers *narrowed the scope of the content* of their teaching in the belief that this would benefit the learners' performance on the test, an effect which can be interpreted as negative washback.

Changes in teaching in response to an exam may be made to the detriment of other content which the teacher considers is not part of the test, or is of less prominence in the test, a phenomenon known as *narrowing the syllabus*. For example, Wall & Alderson (1993) and Alderson & Wall (1993) found that teachers in Sri Lanka altered the content of what they taught in secondary school English lessons to focus on reading and writing, which were tested in the new O Level exams. Speaking and listening, which were not tested, were given less class time to allow for the focus on the other skills. Alternatively, the teacher may prioritize skills, or content, that they perceive to be focal parts of the test and spend a greater length of time on them than in the past. Lam (1994), for example, found that teachers

from Hong Kong secondary schools in a questionnaire study claimed they focused in school not only on those skills which were new in the revised university Use of English entrance exam, but primarily on those which received heavy weighting in terms of the allocation of points. Special lessons were allocated to prepare learners for these parts of the exam.

Teaching content may also be influenced by the test in the type of tasks which are selected for classroom use. Stecher, Chun & Barron (2004), for example, researching in general education, found that teachers changed how they taught writing in response to new standards-based assessment in schools in Washington state, USA. The introduction of open writing tasks in the assessments led to overt teaching of writing and an increase in class time spent on writing activities. The teacher may adapt the type of tasks used when teaching to mirror those which appear in the test, changing, for example, the tasks commonly used for checking reading comprehension to cover a wider range of sub-skills. Content can also be changed through selection of materials, with the teacher choosing material designed specifically to prepare learners for the test in question, be this a course book, or exam-practice material. Wall & Horák (2006), for example, found that teachers used TOEFL preparation books not only to teach, but also as the syllabus for their courses, confirming earlier findings by Alderson & Hamp-Lyons (1996) looking at the same exam.

Research findings on whether teachers change *how* they teach in response to tests are mixed. One possible reason for this could be in the research methodology of some of the studies. Teachers do not make rapid changes in how they teach, rather this happens gradually over time (Cheng, 1997, 1999, 2005), and so such changes will most likely be detected only in longitudinal studies. A second factor is the research instrument used for data collection. Change is not something of which we are always aware and as a result may not be captured adequately in self-report questionnaires, which are a commonly used form in washback studies. Longitudinal studies which include observation of the same teachers, including a period of observation from before the introduction of the new test, would be the most effective way of recording changes in methodology (Cheng 2008: 359). Such research is problematic from a logistical perspective and consequently infrequent. It is also difficult to be able to say conclusively that any change observed has occurred directly and exclusively as a result of a new test (Cheng 2008: 358), as the way any teacher teaches is likely to evolve over time in response to a very large number of other factors, such as teacher education events, peer-observation and changes in the learners they teach. For this reason observation needs to be accompanied by follow-up interviews with the teacher, or questionnaires developed in response to points arising from the observation. Glover (2014) suggests that the fact that few studies show changes in methods used by teachers may be an artefact of how 'methodology' is operationalized. The problem being that

this is a broad concept and if the researcher predicts the changes which might be anticipated as a result of a new exam or assessment process these may not adequately take into consideration individual teaching styles.

Table 1.2. Empirical research on washback: Washback effects in teaching

Washback effects reported (italics indicates research studies in general education)
Content Wall & Alderson, 1993; Wall, 1999, 2005; Cheng, 1997, 2005; *Smith, 1991*; Andrews, 1995 in Andrews, Fullilove & Wong 2002: 210; Shohamy et al., 1996; *Pedulla et al., 2003*; Ferman, 2004; Lam, 1994; *Stecher et al., 2004*
Method **Effects found** Lam, 1994; Shohamy, 1993, Saif, 2006; Ferman, 2004; Glover, 2014 Effect found but varied according to teacher Alderson and Hamp-Lyons, 1996; Watanabe, 1996, 1997; Burrows, 2004; Read and Hayes, 2003; Andrews et al., 2002 **No effect found** Wesdorp 1982: 130; Wall and Alderson, 1993; Wall, 1999, 2005; Cheng,1997a, 1999, 2005 "very slowly, reluctantly and with difficulty" (in Tsagari, 2007: 34); Qi, 2004, 2005
Teacher-made assessment **effect found** Wall & Alderson, 1993: 66; Wall, 1994; Wall & Horák, 2006 **No effect found** Wesdorp, 1982
Teacher talk Glover, 2014
Time allotment made in class for test preparation *Smith, 1991; Sturman*, 2003; *Pedulla et al., 2003*
Atmosphere in class Watanabe, 2000; Alderson & Hamp-Lyons, 1996; East, 2014
Attention paid to skill area which is on the test Wall & Alderson, 1993: 52; *Stecher et al., 2004*
Teaching test-taking strategies *Smith, 1991a: Sturman, 2003*
Use of exam preparation materials Lam, 1994; Andrews, 1994b; Shohamy, 1993; Alderson & Hamp-Lyons, 1996; Cheng, 1997b; Read & Hayes, 2003; Wall & Horák, 2006; Ferman, 2004
Adapted materials to meet learner needs Saif, 2006: 28; Lam, 1994
Classroom assessment practice Alderson & Wall, 1992, Wall, 1994, Wall & Horák, 2006 **No effect found** Wesdorp, 1982

Where changes in methodology have been found, it has often been noted that they are individual to the teacher, with different responses found in teachers in the same context (Watanabe, 1996), or even in the same school (Alderson & Hamp-Lyons, 1996). While these differing responses may be attributable to teacher factors, some researchers found differences depended on the course being taught (Hayes & Read, 2004). Qi (2004, 2005) noted that while teachers changed *what* they were doing in class (e.g. they had increased time spent on listening, reading and writing in class, as had been intended by the reform of the National Matriculation Test in China), they had not changed *how* they were teaching, indicating that the examination had failed to bring about the deeper pedagogical change it had been designed to effect. The finding that examinations bring about superficial, rather than profound changes in teaching was also noted in Cheng (2005) and Wall (2005). Other studies, such as Wall & Alderson (1993) or Wesdorp (1982) found no evidence that teachers had changed their methodology.

Teachers have been found in some situations to change how they assess their learners in response to a new test, or in what scoring system they use (Alderson & Wall, 1992; Wall, 1994; Wall & Horák, 2006), adapting their practices to match the exam. Wesdorp (1982), however, found no such change. Another type of response is for the teacher to instruct learners in how to approach and prepare for the test and to introduce them to specific test-taking strategies (Smith, 1991; Sturman, 2003). This may also be reflected in the way teachers talk to their learners (Glover, 2014), where explanations of how to tackle test tasks may be given, which increase the proportion of teacher talking time. Classroom time may also be allocated to doing practice tests, or practising test tasks (Smith, 1991; Sturman, 2003; Pedulla et al., 2003), particularly in the weeks leading up to the examination.

To sum up, the influences of a test in the classroom vary according to a large number of factors in the teacher and the decisions they make about how to respond to the test. In some teachers these responses will be intense and the changes in their classroom teaching clearly evident, while others may be less overtly affected. The reasons for differences in teacher responses can partly be explained by the wider context in which they work and by pressure, or perceived pressure, from other forces. These will be considered in the section about impact below and in chapter two, where we discuss the use of test scores.

Washback on the learner and learning

While substantial research has been done into the effects of new tests on teachers and teaching, less work has focused on the effects on the learner or the learning process (Bailey, 1999: 14; Spratt, 2005; Tsagari, 2007: 54; Cheng, 2008: 360).

Table 1.3. Empirical research on washback: Learner factors found associated with washback (italics indicates research studies in general education, see review of Kirkland, 1971, earlier)

Anxiety Cheng, 1998; *Ebel, 1965;* Ferman, 2004; *Goldman, 1961; Hill & Sarason, 1966;* Huhta et al., 2006; *Pedulla et al., 2003*; *Ruebuch, 1963; Sarason et al., 1960;* Shih, 2007; Shohamy et al., 1996
Individual differences Ferman, 2004; Gosa, 2004; Hahn et al., 1989; *Paris et al., 1991;* Shih, 2007; Shohamy, 1993
Motivation Berwick and Ross, 1989; Green, 2007b; Huhta et al., 2006; Shih, 2007; Watanabe 1997, 2000, 2001 Goals & Actions Booth, 2012 Possible selves Zhan & Andrews, 2014 **No effect found** Cheng, 1998: 297
Anxiety Cheng, 1998; *Ebel, 1965;* Ferman, 2004; *Goldman, 1961; Hill & Sarason, 1966;* Huhta et al., 2006; *Pedulla et al., 2003*; *Ruebuch, 1963; Sarason et al., 1960;* Shih, 2007; Shohamy, 1993
Self-efficacy/self-concept *Atkinson, 1964; Atkinson & Feather, 1966;* Cheng, Andrews & Yu, 2011; *Crandall, 1963; Glidewell & Stringer, 1967; Goslin, 1967; Modu, 1969; Taylor, 1964;* Xie, 2010; Xie & Andrews, 2012; Zhan & Wan, 2013
Attitude to test Berwick and Ross, 1989; Cheng, 1998; Huhta et al., 2006; Read and Hayes, 2003; Shohamy, 1993, Shohamy et al., 1996 Perceived importance of test *Paris et al., 1991* Perceived difficulty of test Gosa, 2004; *Hills and Williams, 1965*; *Anastasi, 1968*; *Anastasi & Cordova, 1953; Wrightstone, 1963*
Previous test experience *Feather, 1966; Sears, 1940; Kagan, 1968; Glidewell & Stringer, 1967; Modu, 1969;* Gosa 2004; Stoneman, 2006; Murray, Riazi & Cross, 2012
Influence of others on the learner
Family, friends Hargreaves, 1997. Murray, Riazi & Cross, 2012; Shih, 2007; *Sewell & Shah, 1968*
Information, feedback from teacher *Brookover, Shailer and Paterson, 1964;* Green, 2007a; *Osler, 1954; Page, 1958; Pollaczek, 1952; Smith, 1952; Wrightstone, 1963*

Learners and washback

Alderson & Wall (1993) include learner attitudes and learner responses to the content of and process of learning in their fifteen hypotheses about

washback. Kirkland (1971), reviewed earlier, gave substantial evidence of the effect of tests in general education on learners' feelings and attitudes. Ferman (2004) found learners reported high levels of anxiety in response to the new oral component of the Israeli matriculation exam in English and noted that learners with average ability in language were more anxious than others. Gosa (2004), using diaries to investigate learners preparing for a matriculation examination in Romania, found that learners differed in their feelings about the exam but that its effects were felt at home as well as in class. Cheng (1998) found that while learners felt that a new exam in Hong Kong schools would encourage them to work hard, they had doubts about whether the exam measured everything they were learning. Murray, Riazi & Cross (2012) studied the beliefs, opinions and emotions of test takers facing the exam required for foreign-trained teachers to qualify to practise in Australia. Effects were found for the participants' personal experiences of the test and also for the influence of others in determining the candidates' attitudes towards the test. Shohamy (1993) found that different learners responded differently to the same examination of Arabic as a foreign language in Israel. Some two-thirds of learners in the study reported positive effects of the exam such as increased motivation, while the remainder felt anxious and had doubts about the validity of the exam. Almost all of the learners in the same study, by contrast, felt anxious about the test of English.

Motivation in relation to tests has been investigated by Watanabe (2001), who found results varied according to how learners perceived the difficulty of the Japanese matriculation examination, which serves as a university--entrance test. Where they thought the test was challenging but that good results were possible, it was found to be motivating. Berwick and Ross (1989) however, found that this motivation was short-lived, as first year university students were reported to have no clear goals. It could be argued therefore, that the Japanese matriculation examination is motivating for some, but only for a limited period and because of the use which is made of its scores (see Chapter Two). Green (2007b), studying examination preparation, found that it was motivation towards the exam, coupled with learner attitudes, rather than differences in the test preparation programmes, that contributed more to the differences in outcomes between groups.

Cheng, Andrews & Yu (2011) found an effect for self-efficacy in Hong Kong schools, where learners who rated their English more highly undertook more tasks related to the exam both in and out of class. The intensity of this activity was related to the timing of the examination and was found to rise as the exam drew closer. Similar findings are reported in Xie (2010), Xie and Andrews (2012), and in Zhan & Wan (2013), looking at the College English Test Band 4 (CET4) in the tertiary sector in China.

Dividing factors in the learner from the effects on learning is an artificial division, made for the purposes of classifying past research in this

paper. More recent research points to the complexity of relationships in the washback process. Shih (2007) hypothesizes that factors in the learner may be affected by factors outside the learner. For example, feelings about a test may be influenced by what others, such as teachers or parents, say about the test. El-Ebyary (2009: 2) argues for intra-washback and inter-washback, where inter-washback develops as a result of interaction with others and may be intentional or unintentional. The teacher may set out to deliberately change the learners' perceptions of a test or, by contrast, learners may develop ideas about a test on the basis of rumour, incorrect information, or alternatively, may misunderstand information which has been given to them. Placing this process in a dynamic model El-Ebyary explains that feelings and attitudes in the learner are in constant interaction with others and the environment and so in a state of frequent change. Booth (2012), taking an approach based on socio-cultural activity theory, argues that no two learners will respond in the same way to a test, despite apparently being in the same context (based on Lantalf & Pavlenko, 2001), as each will construct their own response differently, according to their own interpretation, and as a result of the interaction of different intrinsic and extrinsic factors.

Washback on learning

Bailey (1996: 264–265) suggests that a test may affect learners in two main ways: in *what* they choose to focus on in their learning and in *how* they go about this. Learners may decide that they need more work on certain aspects of language and so take action, for example, by deciding to listen to more to recordings, or by watching TV programmes, or films in the target language. They may deliberately study the examination guidelines, or syllabus, and spend time practising the grammar, or functions they find there. They may focus on specific vocabulary areas which they know will be covered in the examination and do exercises to develop their knowledge of words, or read material on related topics. How they go about this will depend on the individual and the context in which they are learning. Some learners may look for help outside school, choosing additional test-preparation courses, seeking tutoring, or working together with their peers. They may also work individually at home using test-preparation materials. Others may look for support in school outside class time, asking teachers for advice, borrowing additional materials, asking for information about sources, or asking teachers for individual guidance. Social-economic factors (Alderson & Wall 1993: 61) and logistics may come into play, with less affluent families unable to afford extra classes, or tutoring, and children from rural areas having less choice of, or less access to, extra-curricular activities because of transport limitations. Learners may also influence what happens in class, for example, requesting that the teacher explain practice test-tasks, asking her/him to do practice tests with them, or to give them feedback on test-related

performance. It should be noted, however, that there are several instances in the research literature where it is reported that teachers perceive that learners require them to do test-related work in class, but where this is contradicted by data from the learners (Alderson & Hamp-Lyons, 1996: 286).

Table 1.4. Empirical research on washback: Effects on learning

Learning
Focus of learning Cheng, Andrews & Yu, 2011; Gosa, 2004; Qi, 2005
Test preparation strategies Cheng, Andrews & Yu, 2011; Gosa, 2004; Huhta et al., 2006; Ferman, 2004; Shih, 2007; Stoneman, 2006; Xie, 2011, 2013, 2015; Xie & Andrews, 2012; Watanabe, 1992
Learner achievement Andrews, Fullilove & Wong, 2002; Saif, 2006 **No effect found** Read & Hayes, 2003; Wesdorp, 1982; Cheng, 1998

Learners have been found to be influenced by the content of tests. Gosa (2004), in a study of the influence of the school-leaving certificate examination in Romania on learning, found evidence from learner diaries that what learners focused on, both at home and in class, were test-related tasks, and that they almost exclusively ignored speaking and listening, which were not included in the examination (ibid.: 226). Qi (2005) also found evidence that learners chose to focus on the test content.

Learners find different ways to prepare for tests. Shih (2007), found that learners responded to the challenge of a new speaking component on the General English Proficiency Test (GEPT) test in Taiwan by developing their own strategies to compensate for the fact that speaking was not commonly practised in class and by finding ways to practice outside class. Similarly, Xie & Andrews (2012) found a direct correlation between the skills learners perceived were needed in the test and test preparation strategies. Where learners had a good understanding of what the test entailed and understood its purpose, this had a direct influence on how they prepared for the test. It was found that learners whose awareness of the skills needed on the test was closest to those given in the test specifications, tended to make more use of metacognitive strategies and to prepare more intensively.

Test preparation strategies have also been also found to be related to how learners perceive the importance of the test (Xie & Andrews, 2012). Xie (2015) found that learners spent more time preparing parts of the test for which there were more marks than other parts and that learners who had positive attitudes towards the validity of the test prepared more intensively than those who did not. Gosa (2004) found that there was a relationship between how learners perceived the examination and the effort they put into preparing for it. Those who thought the examination would be easy were found to make little effort. Watanabe (1992) found, however, that the

strategies different learners used to prepare for an examination varied according to language level, motivation, and learning experiences. This was confirmed by Ferman (2004), who found that lower level learners studied more intensively than more able learners and were more likely to undertake coaching in an attempt to improve their performance. Stoneman (2006) reported that how the learners prepared for an exam was affected by their previous experiences of learning and of examinations. Cheng, Andrews & Zu (2011: 4) in a study of the effects of School Based Assessment (SBA) in Hong Kong, found differences in attitude and approach between learners according to their perception of their linguistic competence. Those who felt their language was better were in general positive towards both the SBA and the external exam and did more language tasks both in class and outside school. The study also revealed that learner attitudes towards the SBA were related to the attitudes of their parents. The authors suggest that the concept of washback in the learner appears to be as complex as the concept in the teacher.

There has been little research on the relationship between washback from exams and learner achievement (Spratt, 2005). Andrews, Fullilove & Wong (2002) found some indications that the new oral test on the matriculation exam in Hong Kong had improved learners' oral skills. Further investigation, however, revealed that learners had employed rote learning and the fact that memorized phrases were used inappropriately in oral test tasks suggested they did not understand them. Saif (2006) found significant differences in oral proficiency between an experimental and control group taught by the same teacher, where the experimental group had followed a course designed to prepare the participants (teaching assistants) for a specialized oral examination and the control group a more general program. The experimental group were shown to have increased their scores on oral test in a second administration, using a parallel version, which followed the end of the 12 week 48 hour course. The findings of both these studies are limited and do not allow generalisation to a wider population. Other studies found no effect on learner achievement (Cheng, 1998; Read & Hayes, 2003; Wesdorp, 1982).

In this section we have looked at the notion of washback, how it is defined and how it has been investigated by researchers, both in general education and in applied linguistics. We have traced the development of the field of washback research from Alderson and Wall's (1993) fifteen hypotheses about the nature of the phenomenon, through a series of models which have developed in complexity as understanding of washback has grown. Shaped by an in-depth account of Kirkland's (1971) review of the effects of tests on teaching and learning, a summary of the relationships in applied linguistics research between factors in the teacher and washback and washback effects found in teaching has been offered, followed by a corresponding summary of work on the learner and learning. For

further information the reader is recommended the most recent review of washback (2015) by Cheng, Sun & Ma.

1.3. Impact

As we have discussed above, we see washback as concerned with the effects of tests on teaching and learning and we shall now move beyond this to look at the wider context which may be influenced by tests, recalling Wall's (1997: 291) definition: "'Impact' refers to any of the effects that a test may have on individuals, policies or practices, within the classroom, the school, the educational system or society as a whole." Chalhoub-Deville (2009: 119) deconstructs the notion of 'test' and considers that impact is "the influence of test content, results, and practices on learning, instruction and the curriculum." The view of impact as encompassing washback is held by Bachman and Palmer (1996), McNamara (1996, 2000), Hamp-Lyons (1998), Shohamy (2001), and Green (2007). Cheng, Sun & Ma (2015) take a slightly different view and discuss three areas affected by testing. First "the consequences of testing" (p. 438), which they define as "the direct results of (mis)use of test scores" (ibid.). Second, washback, which they see as "the influence of testing on teaching and learning" (ibid.), and finally, impact, viewed as the influence of tests in society.

In this work it has been decided to separate discussion of the use of test results from this section and deal with it in the next chapter. However, unlike Cheng et al. (2015) we consider "test consequences" as falling within impact. The decision to deal with uses of tests separately in this publication is a purely practical one, taken with the intention of making the concepts and the substantial amount of literature associated with them more digestible.

1.3.1. Impact and washback

The aim of this section is to show the differences between the notions of test washback and impact and how they are operationalised.

If we take the metaphor of the making of a film, we can consider studies of washback as close up, detailed shots, while impact studies pull back the lens to capture a bigger picture, taking in the wider surroundings and showing the more complex inter-relationships between the test and the macro-environs. We now see not only the teacher and their students in a classroom, but that class within a school, that school in a district, that district in a region, a province or state, and a country. At the same time the class is part of a series of systems: micro-systems in the school, where norms are determined internally by the teacher, or the head of department, or the school principal, and successively larger systems where norms are determined by external regulations, established by the school district,

the larger educational administration, or ultimately by state or national institutions. As these systems fan out from the epicentre, which is the classroom in a school, the number of factors which interact with them grow. Whilst in the micro-system in school it is relatively simple to define which parties, or *stakeholders*, are concerned with a test, but as these systems move out into the macro-context the number of interested parties grow and become somewhat less easy to identify. There is also the sense that as the focus shifts from the class in a school to the wider educational setting so the 'ownership' of the test and its consequences change, moving from being a matter for a particular teacher and a particular learner or group of learners, to being a matter which is increasingly less personal and more of general public interest. Impact studies therefore tend to be concerned with more abstract issues than those looking at washback, considering concepts of test quality, such as validity and reliability and the functioning of a test with relationship to the curriculum and in terms of educational policy. Madaus, (1985: 616) illustrates this more nebulous quality of impact when he writes: "The debate about the use of tests in policy matters is really a debate about what we want from our schools. It is a debate about educational values and competing philosophies of education." We shall see that the impact of examinations is an emotive issue and that examinations and assessment as part of educational policy frequently become political in many contexts. Epistemological issues such as the ethics of a test, its impact in a society in terms of concepts such as equity and fairness are also the focus of impact studies. Varghese (1998: 49), writing about project evaluation vs project impact, illustrates this in the following example:

> An evaluation of adult literacy programmes may indicate the total number of persons made literate by the programme. An impact study of the programme will focus on the social implications of the outcomes…It will also ask, for example, whether the reading habits of the community improved.

Test impact can also be seen in this way, looking at how a test functions in a society and what effects it brings about.

1.3.2. Conceptualising impact

As test impact has such wide-reaching scope, determining how to operationalise it is a primary concern. Researchers have conceptualized impact in various ways, treating the introduction of a new examination in an educational system as an innovation (Wall & Alderson, 1993; Wall, 2005), or by taking the study of test impact as one aspect of educational evaluation (Hawkey, 2006). Other approaches view impact as one aspect of consequential validity (Messick, 1996, see Chapter 2) and so place the study of impact within the process of validation of a test (Kane, 1990, 1992, 2013, see Chapter 2), or see beneficial impact as a desirable quality of a good test

(Bachman & Palmer, 1996; Saville, 2010). Following Messick's suggestion (1996: 252) that "rather than seeking backwash as a sign of test validity, seek validity by design as a likely cause for backwash" Cambridge ESOL planned a pro-active approach in their international examinations with the aim to "work for positive impact" (Saville, 2010: 2), most particularly at the design stage. (This will be explored further below). Bachman & Palmer (2010) take the idea of positive impact as being one of the qualities of a test further and propose the building of an argument to demonstrate the beneficial consequences of a test, supported by evidence, which they call an Assessment Use Argument (AUA). Kane (e.g. 2013) also proposes an argument-based approach, but for the purposes of test validation. (see Chapter 2).

Wall & Alderson (1993: 68) place the introduction of a new examination within the field of educational innovation.

> if an exam is to have the impact intended, educationalists, and education managers need to consider a range of factors that affect how innovations succeed or fail and that influence teacher (and pupil) behaviours. The exam is only one of these factors.

A new high-stakes examination is seen in this sense as an innovation, having the potential to bring about beneficial educational change. Wall (2005) is also concerned with the notions of innovation and educational change. She quotes Nicholl's definition (1983: 4):

> An innovation is an idea, object or practice perceived as new by an individual or individuals, which is intended to bring about improvement in relation to desired objectives, which is fundamental in nature and which is planned and deliberate.

It is an innovation theory approach which Wall takes in her analysis of the introduction of a new national examination in Sri Lanka, data for which was collected as part the O Level Evaluation Project (2005: 5) and later in the TOEFL Impact study (both discussed below).

Hawkey (2006), reviewing studies of impact, looks at impact within studies of educational evaluation. According to Rea-Dickins & Germaine (1992: 7), evaluation can be carried out either to explain or confirm what is already happening, "to confirm the validity of features of the classroom context" (Rea-Dickins & Germaine 1992: 8), or to gather information within a process of "planned change" or innovation (ibid.), where the evaluation serves as providing initial information which will lead to subsequent change in, for example, the teaching programme or materials. Evaluation in this sense is exploratory, assessing positive and negative features of the situation and hypothesizing possible reasons for its existence or factors contributing to it. By contrast, Weiss (1998: 331), also within the context of educational evaluation, sees impact as "the net effects of a programme

(i.e. the gain in outcomes for program participants minus the gain for an equivalent group of non-participants)" and also considers the impact of the programme in the wider context. Thus impact is seen as following, rather than preceding, innovation and to be concerned with changes which have already taken place and the benefits experienced.

Evaluation, like assessment, can be focused on *process* in formative evaluation (i.e. taking place during the implementation of an innovation), or on *products* in summative evaluation (i.e. taking place at the end of the innovation) (Trochim, 2006). Impact evaluation falls within summative evaluation (ibid.). Albee (1999), concerned with the field of development and aid programmes, defines impact evaluation as assessing "change and its significance in relation to effectiveness, efficiency, relevance, impact and sustainability" (Albee, 1999: 4).

The term evaluation may also be also used in a slightly different sense, to refer to monitoring the quality, or effectiveness, of the work of a school. Mizerek (2010: 21), in this context, sees educational evaluation as an applied form of social research, the aim of which is to gather information which is used for making value judgements about decisions made and actions undertaken. One of the tools used to make decisions about a school's performance are standardised tests, functioning either at district, state or national level and the results obtained by learners. We will explore this aspect further in the context of *accountability* in the next chapter.

1.3.3. The nature of impact

The word 'impact' has two main meanings: "the action of one object coming forcibly into contact with another", or "a marked effect or influence" (Oxford dictionaries online). There is also the verb "impact on" meaning "to have a strong effect on someone or something" (ibid.). Thus when discussing the impact of examinations it is important to identity who or what is being affected by them. Taylor (2000) identifies a wide-ranging group of stakeholders who are interested in the outcomes of a test. These include the same parties as those affected by washback: the learners, parents or carers and teachers, but also others: "school owners, receiving institutions, government agencies, professional bodies, employers, academic researchers" (Taylor, 2000: 2). Working with the testing agency Cambridge ESOL, Taylor is concerned with the impact of a commercial suite of tests, rather than a national test within an educational system. She also considers those who are concerned with aspects of the test itself, rather than simply its outcome, such as teachers, learners, and school owners, materials writers and publishers, who are interested in the test construct, format, and assessment criteria as well as in how the test is conducted. Finally there are those who are directly involved in either test construction (such as item writers and consultants) or test administration

(such as raters, examiners and invigilators). Taylor (2013) changes her perception of stakeholders to one of those involved at the 'assessment core' and those on the 'periphery' (Taylor, 2013: 408), where those at the core possess expert knowledge of testing and are directly involved in test production or research on testing, and those in the next layer, surrounding the core, are teachers whose understanding of testing is less profound, but who are still informed, while those on the periphery (such as policy makers and the general public) are lay people in terms of their understanding of assessment issues. Taylor argues that as the levels of assessment literacy (i.e. knowledge and understanding of testing and assessment) are different in each of these stakeholder groups, so the information about the test which is communicated to them by the agency responsible for the test should be differently presented and worded (Taylor, 2013: 409–410). Pill and Harding (2013: 382) echo this notion of informed and uninformed stakeholders when they write:

> Those groups who use language test scores as the bases for decisions, but who are not actively involved in the construction of test materials, may make assumptions about tests, testing processes and outcomes that are at odds with what is intended or can be endorsed by the language testing community. Such misconceptions may have serious consequences for decision-making based on test scores…

Thus they place the 'language testing community' at the core and 'decision--makers' at the periphery. They underline a key issue in the impact debate: that it is most often those at the periphery, who do not have a specialist knowledge of testing, who decide what use should be made of test scores. In the next chapter we will see the mechanisms which this can bring into play. To conclude, when considering test impact we need to identify stakeholders, and to consider their level of understanding of the test in question and of assessment in general.

The next aspect of impact that needs to be identified is *what* may be affected by an examination. Here we find an overlap with what has already been discussed in washback, if we consider the micro-context. However, our main concern now is the macro-context and the role of tests within larger systems. If we consider national education systems we can find instances of tests impacting on programme content and design (e.g. Qi, 2005), on assessment practice (e.g. Munoz & Alvarez, 2010), on educational materials (e.g. Cheng, 1997) and on teacher education programmes. The aspects of the test which have this kind of impact are its construct, content, assessment criteria, and the procedures associated with its administration. Moves from internal, school-based testing to external, standardised testing impact on stakeholder perceptions of the fairness and objectivity of the examination and may also change attitudes towards ethical issues, such as cheating (Bartmańska, 2007).

High-stakes tests are often found to be introduced into educational systems with the aim of raising standards of teaching and learning, which are measured in terms of increase in learner scores on the tests. The theory behind this is that introducing a test based on standards where the results have some consequences will encourage teachers to do their best to get their students through the test with the best possible scores (Popham, 1987). This may be done for personal or professional satisfaction, or with a sense of moral responsibility that their students get the best possible chance in life as a result of the test scores. This thinking continues with the belief that such tests are motivating, both for the teacher, who wants the best for their learners and for the learner, who wants the personal satisfaction of having done their best in the test. Thus the idea is that working towards the test will improve the quality of the teaching and learning experience.

Popham (1987) argues that a change in the role of tests can improve instruction. Advocating that a well-conceived and designed test can be used to lead the planning of instruction, rather than be added after a period of instruction as an afterthought, he proposes *measurement-driven instruction* (MDI), theorizing that the test will act as a "powerful curricular magnet" (Popham, 1987: 683). He suggests that responsibility lies with the test designers and constructors to ensure that the skills being tested are appropriate for the creation of good instructional materials. In this way tests should be based on a manageable number of clearly defined targets, be criterion-referenced and test both higher and lower order skills, and support should be offered to teachers together with instructional materials.

Additionally, when a high-stakes testing programme is initiated the situation can be deliberately manipulated by including specific content, task types or processes which are new to the context. The launch of such a test is usually accompanied by extensive explanation and demonstration, and includes a program of teacher education events together with training materials. Here the thinking is that, through having to prepare learners for such content, tasks or processes on the test, teachers will be forced to develop their own knowledge and skills and change their pedagogical approach. An additional mechanism in this process are new teaching materials which are designed to promote the changes (Wall & Horak, 2011: 134). Unfortunately, this is a naïve view, as Madaus (1985: 615) skeptically summarises,

> In short, testing becomes the amazing grace that will save the world from illiterate graduates. If it is associated through legislation with strong measures of accountability, testing will bully the instructional system into line. Inevitably, test scores will rise, and policy makers will be able to point to the righteousness of their gospel. If this be madness, there is method in it.

As we saw in the review of studies of washback, the context of teaching and learning is highly complex with very many variables dynamically interacting with each other. Tests do not work in school in a simple cause-

-effect relationship with predictable, positive outcomes. Researchers have therefore hypothesized *potential* positive consequences of high stakes testing and conjunctures which would seem to favour them coming about. Cizek (2001) proposes ten such potential consequences: first, that as the teacher is required to focus more on the whole of the curriculum, they develop their content knowledge, techniques and expertise (here he assumes that the test content covers the whole of the curriculum, i.e. is *aligned* to the test). As high--stakes tests are for all children, including those with learning difficulties, teachers are required to ensure adequate education for everyone, and so equity will be enhanced. If the test scores have consequences this means that teachers learn more about the tests, assessment processes and marking. This may positively affect classroom assessment practice (Goldberg and Roswell, 1999/2000). Data obtained from tests helps teachers make more informed decisions, analyze the outcomes of their work and adapt programme content. Data can also be used for decisions about funding, however, although given by Cizek as a positive benefit, this is potentially controversial as we will see below. As the education system strives to meet the new challenges there will be a greater variety of schools and courses on offer, giving parents more choice. High-stakes tests play a key role in developing policy on accountability of schools and promoting debate on effectiveness. Debate and decisions about the content and format of high-stakes tests involve not only groups of experts, but also consultation with the wider teaching community. This leads to reflection on a wide variety of related topics and so being involved in high-stakes testing may affect assessment practices in school (Banerjee, 2000). For example, based on personal experience, completion of examiner training for the Polish *matura* (school-leaving) examination by several members of staff of the Foreign Language Teacher Training College in Zabrze led to our systematizing development procedures for year-end tests in the college (on the basis of which decisions were made about student promotion). Templates for detailed test specifications were introduced, together with a requirement for the production of written assessment criteria, marking keys and scoring information. A review procedure for tests was also implemented, where the test, specifications and syllabus were compared and proposed rating and scoring procedures were reviewed. This was followed by introduction of anonymity on examination essay papers through the use of identification codes to ensure rater impartiality, and double marking to improve test reliability. All of these changes followed from discussions promoted by our involvement in the examiner training course, in which we reflected on reliability, test fairness and accountability. Without doubt, the changes improved our assessment practice.

Cizek (ibid.) continues with two further possible positive consequences of testing. As a high-stakes test is the object of scrutiny this pushes the testing body responsible for it to raise the reliability and validity, rating

and scoring procedures and implementation of the test, thus improving overall quality.

There is some empirical evidence to support Cizek's hypotheses that high-stakes testing may bring about positive outcomes. Popham (1987) showed that basic reading and mathematical skills had improved in six states and one district in response to high-stakes testing and that there were marked increases in scores for African-American and economically deprived young people in Texas, South Carolina and Connecticut. Bishop (1998, 2000), in a series of studies comparing Canada with other countries, found a significant positive relationship between presence of high-stakes external tests and scores on TIMSS (Trends in International Mathematics and Science Study) and IAEP (International Assessment of Educational Progress). Rosenshine (2003: 4–7) found that states of the USA which had high-stakes testing programmes with clear reporting procedures performed better on NAEP (National Assessment of Educational Progress) tests of maths and reading in the 4th and 8th grade over a four year period than those states which did not have high stakes testing. He therefore concluded that, contrary to Amrein and Berliner's 2002 claims, high-stakes testing had a positive impact on student learning in general, not only on what was being tested in the high-stakes test. However, he points out that not all of the states with high stakes testing performed equally well on all of the concurrent measures and suggests that further research is needed.

Frederickson & Collins (1989), like Popham and Ciszek, also theorize how high-stakes tests can be used to good effect, but they approach this from the perspective of how a testing system can be designed in order to have positive consequences. Their first concern is for the content of the test itself and the tasks it contains. The test should provide good coverage of the curriculum and ensure that not only the content matter, but also the processes it aims to teach (such as problem-solving) are tested. The test tasks should reflect real world use of skills and knowledge and be as authentic as possible. Tasks should test skills directly and the skills selected should be those which are key in the learning process. They recommend access to sample performances of each of the key skills which are tested, at a range of different levels of attainment, to provide clear indicators. To ensure rating criteria which will produce scores appropriate for their proposed use, the end-users should be involved in the process of drawing up the assessment criteria. For example, if this is to be a matriculation examination which will be used for selection to tertiary education, then representatives of such institutions should take an active part in the production of the assessment criteria. A training program for scoring the test should be made available, not just to those who will score the test itself, but also to teachers and learners, so they fully understand the process. This transparency of the system will raise motivation and help to direct teaching and learning. The authors also recommend that the testing system promotes the practice of self-assessment

through offering preparation that provides feedback designed to help the learner analyze the strengths and weaknesses of their performance, using the scoring criteria and available exemplars. Finally, they suggest that the high-stakes test be accompanied by extended tasks or projects done by the student which are designed to demonstrate they are able to apply the key skills in practice. These would be rated according to the key skills, against a set of exemplars. Such tasks could form part of classroom practice and be assessed by the teacher. They conclude: "The goal of assessment has to be, above all, to support the improvement of learning and teaching" (Frederickson & Collins, 1989: 32). Thus the importance of the testing agency in providing training materials, exemplars and a transparent process is fundamental.

The literature on assessment, however, is weighted on the opposing side, those who are against high-stakes testing. Madaus (1988), in a classic article, outlines six principles to demonstrate how the process theorized by Popham (1987) has gone awry. The first of these deals with perceptions of the importance of a test. Madaus argues that defining a test as high- or low-stakes is inadequate, what matters is how important its results are perceived to be by those involved and he adds that such perceptions are open to distortion by the media with, for example, the publication of league tables ranking school results. Secondly, Madaus describes the mechanism he refers to as an application of Heisenberg's principle of uncertainty, which is that when a qualitative indicator is used as a benchmark for measurement, then it is expected it will cause changes in the things it is attempting to measure. Thus, testing will change the educational process, but not necessarily for the better, rather by causing distortion. These distortions feature in the next principles described. If the results of the test have important consequences then "teachers will teach to the test" (p. 37), meaning that the instructional process will focus exclusively on the test content. This will mean that, in effect, past test papers replace the curriculum. In addition, teachers will be influenced by the form of the questions on the test and match their teaching to them. Finally, if test scores are the primary decision-makers in certification or qualification, then "society tends to treat test results as the major goal of schooling" (p. 43), a situation which Madaus warns is particularly detrimental if the test scores are also used to evaluate the quality of a school, or of the educational process. In this event the second principle operates and the test score ceases to be an effective measure of learner ability.

Popham (1987: 681–682), disclaiming criticism of MDI, outlines the claims made against the use of high-stakes testing. These are that it narrows the curriculum and promotes a focus on lower level skills (in response to multiple-choice testing of basic skills, prevalent at the time in the US). The teacher's creativity is limited by having to focus on what is to be tested and this also makes the curriculum repetitive and predictable. Student

motivation lowers as their main aim becomes the test content, rather than the development of the skills the test is designed to assess. Finally, tests as 'drivers' are not appropriate, as assessment should follow instruction and not lead it.

Summarizing the writings of Smith & Rottenburg, Mehrens (1998: 7) adds to this list of the negative consequences that standardized high-stakes tests might have

> (1) external testing reduces the time available for ordinary instruction, (2) testing affects what elementary schools teach – in high stakes environments, schools neglect material that external tests exclude, (3) external testing encourages use of instructional methods that resemble tests, and (4) "as teachers take more time for test preparation and align instruction more closely with content and format, they diminish the range of instructional goals and activities" (1991: 11).

This fifth consequence is expanded by Smith (1991) who, in an in-depth longitudinal study, traced the impact of high-stakes testing on the work of schools in Arizona. She found that teachers felt demoralized, guilty and angry in response to their efforts being reduced to numbers and published in lists of school rankings. The teachers reported feelings of powerlessness, in that they were no longer in control of the learning process, but rather felt put under pressure by principals to improve results. They also felt vulnerable in that if they did not conform to these expectations this might have consequences for their future employment. Finally, Chapman & Snyder (2000: 457) point out that standardized tests may lead to "cultural, ethnic and gender biases," as it is challenging to account for diverse groups in one test.

To sum up, we have considered the nature of impact, in particular who is affected (which stakeholders) and what may be affected. In discussion of what may be affected we looked at the thinking process behind the advocating of the use of high-stakes testing to improve teaching and learning and the potential benefits it promises and then outlined arguments which point out the potential detriments of using high-stakes tests to lever educational systems. It should be noted that in the most part what has been presented are all theoretical arguments, with little empirical support. Mehrens (1998) cautioned in a review of research conducted at that time, that there was inadequate evidence available to draw conclusions for either case. We will return to this issue in the next chapter.

1.3.4. Models of how impact works

We have seen arguments proposing relationships and interactions between high-stakes testing and teaching. In this section we will examine three models which propose *how* this takes place. The first, based on the work

of Chapman & Snyder (2000), demonstrates further the thinking which has been outlined above. The model will be presented and then discussed in order to show the mechanisms which could take place. The problems inherent will then be explained. The second model comes from the work of Wall (2005) and is based on Henrichsen's Model of Diffusions of Innovations in ELT. Wall first used this model to interpret the introduction of a new examination system in Sri Lanka. Finally, we look at Saville's (2010) model of impact, designed for use by the Cambridge ESOL examinations board with the aim of improving the quality of their suite of examinations.

Chapman & Snyder's 'Conventional wisdom' model (2000)

On the next page (Figure 1.2) we show a version of Chapman and Snyder's model of how high stakes testing is purported to lead to improved student learning outcomes. The authors, with the illustration of examples taken from national examination reforms in general education from different, mainly developing, countries explain mechanisms which may operate in various contexts which influence whether impact takes place, how and to what extent. These mechanisms do not appear in the diagram. We take in turn each of the points from the column on the left and describe, reading across the rows, mechanisms which filter or obstruct the process intended.

The first point is that test results help authorities decide where to allocate resources and to identify low-achieving schools or districts. The problem with this hypothesis is that it is based on a fallacy, which is that additional resources alone will improve instruction. Further, if resources mean additional funding, then on what is this to be spent? Who decides and on what grounds? It also assumes that it is logistically possible to increase resources in low-achieving schools, which in practice may not be the case, particularly if "increasing resources" is interpreted as getting well-qualified and experienced staff to relocate to work in remote or undesirable areas. Second, it assumes that school principals and teachers know what to do with the resources they receive but, even if they do, there is no guarantee that this will improve standards of instruction or raise test scores. The argument proposed is seen as invalid because of the unstated assumptions upon which it is based.

The second point argues that tests can drive instructional change. Chapman and Snyder use two examples to show that this *might* be the case, but that this is not necessarily so. Trinidad and Tobago changed the Common Entrance examination at the end of primary education from a multiple-choice only format, to one which included essays, on the thinking that writing an essay involves application of higher-order thinking skills. It was expected that the change would push teachers to use more essay writing in class, and so move away from a focus on memorization. After initial resistance and a period of adjustment, during which test results were lower than in previous years, teachers did indeed change their practice and

Figure 1.2. Model showing 'conventional wisdom' about high-stakes testing leads to improved student learning outcomes. Based on Chapman & Snyder (2000: 466)

How tests can be used		**Action/Result**		**Impact on schools**		
Test results decide funding of schools	>	Low-performing schools/regions receive increased funding	>	Increased funding designated to raise quality of teaching	>	
Changes in test content or tasks places new demands on teachers	>	Poor test outcomes raise alarms among teacher and parents	>	Increased funding/resources to low-performing schools	>	
Publication of results of tests serves as incentive for improvement among teachers	>	Societal milieu places pressure on teachers	>	Teachers respond and change their teaching to improve student outcomes	>	Student outcomes improve
Governments compare student outcomes with other nations and increase spending on education	>	Low results on international tests widely reported in the media	>	Increased resources, or more varied resources, designated to schools	>	
Results give teachers useful information	>	Inappropriate aim for high-stakes tests				

learners developed the use of argumentative skills in writing. By contrast, however, in Uganda, with similar changes in a national examination, also designed to promote more challenging teaching tasks, a study which included classroom observation and interviews (Snyder et al., 1997) found that many teachers were unable to change their teaching in accordance with suggestions made by the national examination board. They felt threatened and the changes seemed to involve risks which were too great. Chapman & Snyder explain that in the Ugandan culture success in examinations was seen as the achievement of an individual rather than as the result of the action of a teacher. Thus, the notion that changing the examination tasks would change instructional practice was likely to fail, because of a gap in the teachers' understanding of what was intended. Mehrens (1998), citing examples from the US, echoes that teachers must be ready and able to change for positive consequences to come about. Wall & Horak (2008: 73) also discuss this issue when considering the impact of revisions to the TOEFL examination on teachers' classroom practice.

To conclude, teachers must fully understand the reasons for changes in examinations, how they themselves need to change and what this means they need to do in their classrooms with their learners as a result. The educators then need to be able to explain what they intend and why to their students and finally to implement these changes successfully. The examples from Trinidad & Tobago and Uganda show that changes are often met with resistance from different stakeholders, particularly teachers and parents, who are likely to oppose any changes if they feel that their children will do worse than before, or that others may now do better. They do not want their children to be the subject of educational experiment and prefer what is old and familiar (Chapman & Snyder, 2000: 461). Change theory shows that resistance is a natural stage in response to what is unknown, but that it can be overcome through information, explanation and support (Head & Taylor, 1997: 168). This is important for policy makers and educational administrators to understand and to plan for.

The third point in the model is that test scores can be used as an incentive for teachers to improve their teaching. Here the assumption is that competition between schools is expected to lever change. Chapman and Snyder provide two examples to show that this *may* happen, but only in favourable circumstances. In Kenya, with the aim of improving quality in low-achieving institutions, a reform in the Certificate for Primary Education included giving schools information about their results, together with those of other schools in the vicinity. Diagnostic information and suggestions for how to improve the school's results were also supplied. Over a five year period, despite a slow start, it was found that low-achieving schools made significant progress. By contrast, in Chile, in a similarly motivated reform, results were communicated to the media, rather than to schools (Schiefelbein, 1993). The media then broadcast information about results in over-simplistic and

misleading ways, possibly due to a lack of understanding of its complexity. Over time no improvements were noted in learner achievement and differences between schools remained. Schiefelbein explains that Chilean teachers had low expectations of their learners and so were satisfied if learners met them. If learners exceeded their expectations teachers did not feel the need to make any changes, while if learners under-performed, teachers placed the blame on factors over which they had no control. As a result it was unlikely that students' test scores could have brought about instructional change or improvement.

To sum up, while information about test scores may succeed as an incentive for teacher improvement, whether it works in practice will depend on how information is communicated, and the extent to which teachers feel responsible for learning outcomes. It may also be the case that teachers do not understand what they need to do to improve student scores, or may lack the necessary skills or subject knowledge to do so. If the changes are perceived as radical this may lead to feelings that the situation is unfair, as it will cause a mismatch between how learners are taught and how they are tested, because teachers continue in their existing practice. The mismatch can result in lower scores, which in turn could cause dissatisfaction among stakeholders. Chapman & Snyder (2000: 472) conclude that while tests can be successful in improving teaching, it should not be assumed that this will happen:

> It has failed as often as it has succeeded, usually because implementing the strategy failed to understand the intermediate conditions that had to be met for changes in test content, format or use to have the desired impact on teachers' classroom practice.

Hattie (2015a) goes further, arguing that policies which focus on attempting to reduce differences on test scores *between* schools are misguided, as the greatest variance is found *within* schools (based on PISA 2009 findings) (ibid.: 2). As a result, he advocates that the way forward is by building collaboration between teachers in a school and on developing the quality of school leadership (Hattie, 2015b).

The next point in the Chapman & Snyder model is that governments can compare scores on international tests with those of other countries and that this can serve as an incentive for making changes in the funding of education. The assumption here is that these are international tests, such as PISA, where the same tests are administered to representative samples of students of similar age in OECD countries. One difficulty is that league tables of results alone do not give information about educational policy and so do not necessarily explain what action could be taken (Chapman & Snyder, 2000: 465). Although each nation has a different context, challenges and culture, comparison of results "focuses minds in education ministries around the world like nothing else" (Sir Michael Barber in Coughlan article: "PISA

tests: UK stagnates as Shanghai tops league table," BBC, 3 December 2013). However, the OECD do also produce more contextual information, such as *PISA 2012 Results in Focus* (2014), where Poland is a case study for showing significant growth in achievement in reading and mathematics between 2003 and 2012, and in which a new project on allocation of resources, entitled the OECD Review of Policies to Improve the Effectiveness of Resource Use in Schools is also mentioned. To sum up, while rankings of scores attract media attention and promote debate, additional information and research on whether apparently successful policies would be appropriate in a given context is needed before action is be taken.

The final point in the model is that national high-stakes testing gives teachers information which they can use to identify and help students with problems. Chapman & Snyder point out that this assumption is based on a misunderstanding of the nature of large-scale tests and their intention. They argue that these are usually norm-referenced tests, which very often, for reasons of expediency, employ closed tasks (such as multiple choice) as these are quick and easy to mark and which, if well-constructed, lead to high reliability. The practicality issue, i.e. that such tests must be economical to administer, not take too much time, or involve too many pages, also tends to limit the number of items covering a given point in the curriculum, or even leads to the de-selection of some curriculum points, which severely restricts the amount of information the results give about attainment of individual curricular targets. To be able to generalize reliably there needs to be good coverage of all parts of the curriculum, which means an adequate number of test items based on each curriculum point. In this event teachers would be able to understand which parts of the curriculum need more attention.

Another issue is that high-stakes tests are not usually based on one year's learning, but rather come at the end of a stage of schooling, such as the end of primary, or lower secondary, or at the end of compulsory education, for matriculation purposes. This means that even if the test does give information about attainment it is not clear at which point in the educational stage this content was learnt by the test-taker, and, consequently, which teacher was involved. Tests designed to give feedback on learning are much more appropriately criterion-referenced, focused on smaller, more targeted parts of the curriculum and are best conducted in-class by the teacher as formative, continuous assessment during the school-year. Such results can then be acted upon at once, channeling support where needed, or adjusting the teaching schedule to do remedial work. A national test at the end of an educational stage gives information about the cohort which has finished that stage and which has most likely left the school. Analysis of results may offer general ideas for areas to work on, but cannot be used to help those students who took the test. Nor does it guarantee that any problems they experienced will be problems for the next cohort. In short, national high-stakes tests are not really very helpful in providing information for identifying or helping

students who have problems, this should be done with formative assessment by teachers.

Chapman and Snyder's model is helpful in developing understanding of the complexity of the interaction between the numerous variables at play when high-stakes tests are introduced to education systems. It underlines the dynamicity of these interactions and the all-importance of context. There can be no one-size-fits-all solution, as even small nuances may lead to the test failing to deliver the aspired-to results. We also see that there is a need for high levels of assessment literacy among educational policy makers. Failure to fully understand the construct of a test, the purpose intended by its designers, or the relationships between testing and teaching, can lead to a high-stakes test being used inappropriately. We return to this issue in the next chapter.

Wall's application of Henrichsen's (1989) Model of the diffusion/implementation model in ELT

Wall took part in the Sri Lanka O Level Examination Project, a four year long evaluation of curriculum reform (1988–1991), which had introduced new course books focused on the development of speaking and listening skills, followed by the introduction a national English test, and associated teacher training. The new examination was envisaged as the motor to drive change (see Alderson & Wall, 1989, 1990, 1991, 1992; Wall & Alderson, 1993). The final year of the study was designed to assess the impact of the new examination, which at that time had been in place for two years. It was planned that seven specially trained Sri Lankan team members would each follow the work of one teacher over two years in seven schools, making 49 schools in total, in different parts of the country. Data collection was to include observation of six lessons over the two year period, gathering of materials used and post-lesson interviews. As with many longitudinal studies, changes occurred in the data collection team and this caused fluctuation in the number of schools taking part. Analysing the data obtained from the observations, Wall found a number of questions arising concerning the reasons why teachers had taken decisions or actions and about their opinions and feelings on the materials and the examination. It was decided to conduct a series of group interviews in late 1991, of which only part of the data was included in the final evaluation report. Wall (1999) returned to the complete interview data set to conduct new analysis in the light of her other subsequent work on educational innovation and her increased understanding of project evaluation in general education. For this analysis she adopted a model for the diffusion of innovation by Henrichsen. This model had been used to analyse changes which had taken place in English language education in Japan between 1956 and 1968, prior to the introduction of a new project. It compiled elements from other models of innovation, but expanded and adapted them to suit his context. Henrichsen's focus was on

innovations which had taken hold, and what the reasons were to explain why they persisted while others did not.

The model has three stages: the situation *prior to* the innovation, the situation *during* the implementation of the innovation and the follow-up stage tracing the *consequences* of the implementation. Each of the three stages will now be dealt with in turn.

Prior to the innovation

Henrichsen holds that, in order to understand why some innovations take effect and others do not, it is vital to have a deep understanding of the context into which the innovation will be introduced. To be able to grasp which effects have been caused by the innovation after its implementation, first it is necessary to have a clear picture of the situation before the innovation is introduced. In this pre-stage he includes four areas:

- Context into which the innovation will be introduced: "Characteristics of intended user system",
- "Characteristics of intended users of the innovation",
- "Traditional pedagogic practices",
- "Experiences of previous reformers" (Henrichsen 1989: 80).

Within each of these areas he considers all the factors which could potentially come into play to support or hinder the innovation. Wall (2005), applying her version of this model, within the "Characteristics of the User System" investigated the situation in schools and classrooms, but also the geographical setting, and the wider political, cultural and economic context. She examined teachers as her "intended users", but these could also include other stakeholders such as learners, school principals and parents. The characteristics of teachers Wall considered included attitudes, abilities, education, and aspects of their personal and professional situations. Henrichsen (1989: 80) defined these characteristics as the "attitudes, values, norms and abilities" of the users. "Traditional pedagogic practices" were defined by Wall (ibid.) through classroom observation, interviews and analysis of teaching materials. She did not consider the final area, as she found it did not apply in her context.

During the innovation

At this stage Henrichsen includes aspects of the *Process* of the innovation, aspects connected with the *Receiver* and a range of factors which might "facilitate/hinder change" (1989: 80). The Process includes four areas:

- "Source (Innovator)" (Henrichsen 1989: 80) – referring to the origin of the innovation, and what motivated it,
- "Message (Innovation)" (ibid.) – referring to the idea of the innovation, its clarity, detail and accessibility,

- "Plans and strategies" (ibid.) – referring to the ways in which the innovation is to be seeded, promulgated and supported and the model of change which is to be adapted, e.g. top-down; social-interaction etc.,
- "Channels of communication" (ibid.) – referring to how information about the innovation is to be transmitted and how feedback is to be gathered (e.g. use of change agents; media campaign etc.).

The *Receiver* includes three aspects: Awareness, Interest, and Evaluation

Henrichsen then includes a long list of factors which may affect the implementation, categorised in four areas: Within the innovation itself; Within the resource system; Within the User System; and Inter-elemental factors, which are those that interact between these different systems.

Factors within the innovation will be described first. Henrichsen adopted six of these from other sources (see below) and added five of his own.

- Originality (categories from Pelz, 1985) – is the idea local and new? is it adapted from elsewhere? is it borrowed from elsewhere but with little modification?
- Complexity (Rogers, 1995: 242) – how easy or difficult do those involved find the new idea to understand or put into practice?
- Explicitness (Henrichsen, 1989: 84) – how clearly is the new idea described in all respects?
- Relative advantage (Rogers, 1995: 212) – when compared to what it is replacing, how much better is the new idea considered to be?
- Trialability (Rogers, 1995: 243) – is it possible to try the new idea out on a small scale?
- Observability (Rogers, 1995: 244) – to what extent will the outcomes of implementing the new idea be observable?
- Status-ideas perceived as related to higher social status "import legitimacy and attract attention to an innovation" (Henrichsen, 1989: 85).
- Practicality (Fullan 1991: 72) – is the new idea a good fit for the situation in which it is to be tried out? does it include clear guidelines for implementation?
- Flexibility/adaptability (Henrichsen, 1989: 85) – can the new idea be modified to suit individual contexts?
- Primacy (ibid.) – is this the first introduction of a change, if not, earlier changes may block or impinge on the new idea?
- Form (ibid.) – is the proposed change 'abstract' i.e. ideas, theories, or 'concrete' i.e. accompanied by materials.

The final stage of the model are "Consequences which occur as a result of the innovation." At this stage Henrichsen evaluates whether the end users have adopted the change, or rejected it. There is also a time dimension in the model, which allows the researcher to see if the acceptance or rejection holds after a period of time, or whether, after initial rejection, there is a later take

up of the innovation, or the converse, a later rejection after an initial take up. The model also allows for different types of response, which it classifies on three further clines of 'Direct-Indirect', 'Manifest-Latent' and 'Functional-Dysfunctional'. Henrichsen (based on Rogers & Shoemaker 1971: 37–38) draws attention to an 'authority decision', where implementation of the innovation is mandated and so teachers are required to comply. He warns that this may lead to a negative response, or to teachers paying lip-service to the reform, giving the appearance of complying, but in fact retaining their 'old' beliefs and practices.

To sum up, Henrichsen's model of the diffusion of innovation in ELT shows the complexity of the process and draws attention to the fact that change takes place slowly over time, with a large number of factors potentially supporting or hindering successful implementation of change. We can see similarities at the different stages with the models of washback described earlier. The elements of "Characteristics of the Intended User" in the Prior phase, and those of the Receiver in the Implementation phase bear resemblance to factors described by Shih in her models of the learner (2007) and the teacher (2009). Burrows (2004), likewise considers teacher attitudes as a mitigating factor in the degree to which washback occurs. Green (2007) places a layer of 'filters' in his washback model, which are the attitudes and values of the stakeholders involved. In this way we see how ideas of how washback comes about are compatible with the Henrichsen model of diffusion of innovation. Where a high-stakes test is introduced into an education system with the aim of bringing about change in teaching practice we can consider this an innovation. Wall (2005) demonstrated that Henrichsen's model could be adapted for this purpose in her evaluation of the impact of the introduction of a new national examination in Sri Lanka (discussed below).

Saville's model and meta-framework: Impact by design

Saville (2010), following Bachman and Palmer (1996), views impact as one of the qualities of a good test, together with reliability, validity and practicality. Milanovic and Saville (1996), working for the test producer Cambridge ESOL, conceived guidelines for building sound tests, with a consideration for test impact included at each stage, the aim being to produce "impact by design" (Messick 1996: 252).

Saville takes the view that "educational processes take place within complex dynamic systems with interplay between many sub-systems and 'cultures'" (2010: 4). He stresses the importance of different stakeholders in the process. In his model of impact the complex dynamic system is viewed as a series of concentric circles, with the micro-context (the learner, teacher and class) in the centre, moving outwards through increasingly larger circles of school, community, region and country (ibid.).

From the experience of three impact projects (reported in Hawkey 2006), Milanovic and Saville's 1996 guidelines (Plan-Support-Communicate-Monitor and Evaluate, see Saville: 2009: 256) were expanded into a model for impact which includes four dimensions: 'reconceptualising impact', 'impact by design', 'procedural basis for knowing about effects and consequences' and 'theory of action' (Saville, 2010: 5). This model widens their original concept of impact by adding a socio-cognitive and social-constructivist perspective (Weir, 2005) and includes the view that both negative and positive responses to a test could be anticipated. Central to the idea of validity in the model is the notion of matching the test construct and purpose to the contexts in which the test is to be used (Messick, 1996, see Chapter 2). Cambridge ESOL take the view that a test is part of an 'assessment system' which interacts with factors in the context of use. Saville stresses that, over time, how and why a test is used changes, according to the local situation, which may lead to it becoming more distant from its original purpose, or intentions. Validity of the system is consequently not constant, but is emergent, dynamically responding to the context over time. Impact studies, conducted over time, are therefore a necessary part of the model for building sound tests, as they can monitor changes which occur in the test contexts, provide evidence of any negative effects and allow for remedial action to be taken by the test developers (2010: 5).

In the first dimension of Saville's model, the reconceptualization broadens the view of exam impact, placing it in a wider context than previously, within geo-political and language education systems. This therefore implies that the theoretical base needs to widen beyond applied linguistics, and consider research in the social sciences, educational change and innovation. We need to remember that the exams in question here are those of an international testing agency (Cambridge ESOL), whose tests take place worldwide and cover a wide range of age groups. This is quite different than what has been considered by Chapman & Snyder, or Wall, who were concerned with high-stakes national examinations within educational systems.

In the second dimension the test producers resolve that anticipating impact needs to be part of the process of building tests. One aspect of this is ensuring that the test construct is adequately detailed, as this is necessary to analyse whether there is a good fit between the test and the contexts where it is in use. The belief here is that: "the overall validity of an assessment system is an emergent property resulting from a test interacting with contexts over time" (Saville, 2009: 256). Part of the interaction is communicating clear, detailed information to stakeholders about the test construct, procedures, marking and scoring, appropriately presented so as to be understandable by the different end-users involved.

The third dimension focuses on "finding out and understanding what is going on" (Saville, 2009: 257) through iterative impact studies which show the working of the tests in real-use in various contexts. These find-

ings would then feed back into the test production process which would respond to any negative consequences that had been identified. The fourth dimension relates to this, in that it seeks to design a plan of action for how the third dimension could be implemented, thus setting out "*what* needs to be done, and *when and how* to do it" (ibid.: 258).

Similarities in the concept of anticipating negative consequences and working to avoid them can be found in Social Impact Assessment, (SIA) which is taken from social science and is most often applied in projects such as the building of dams, or large-scale mining which will have impact on communities and societies (although also used in other contexts such as tourism.) SIA is designed to be used at any of the stages of such a project and follows four main phases:

- Understand the issues
- Predict, analyse and assess the likely impact pathways
- Develop and implement strategies
- Design and implement monitoring programs

(Vanclay, Esteves, Aucamp & Frank, 2015: 7)

While a full description is beyond the aims of this publication, parallels can be found with both what Saville intends and with Henrichsen's model. While Henrichsen has a phase prior to the implementation of an innovation, Saville looks to "reconceptualise" and widen the definition of impact. In SIA, the first phase "Understand the issues" encompasses both of these. Within this phase are more theoretical aspects, such as "understand the proposed project; clarify role and responsibilities; and social area of influence" (ibid.) which could be considered similar to Saville's (re)conceptualization and to Henrichsen's "Characteristics of the intended user system". This is followed in SIA by more practical aspects "Community profiling" and "scope issues" which involve identifying the main issues of concern and those who are affected. This is done through direct contact with the communities in the area where the project is to take place: "Inform communities" and "inclusive participatory processes", which can be understood as collecting information from the stakeholders. Here there are correspondences with Henrichsen's Prior phase and "Characteristics of Intended Users" and the equivalent of "Traditional Pedagogic practices" – here widened to be a profile of the community which will be affected. The final action in this SIA phase is to "assemble baseline data", as was the case with the Henrichsen model.

In the second SIA phase various impacts are predicted, including indirect and cumulative impacts. Also anticipated are "affected party responses", a factor which perhaps could be seen in Chapman & Snyder's model, although they showed actual responses based on empirical findings after the implementation of new exams, rather than predicting what response might have occurred. We could perhaps, however, on the basis of their model, in the light of detailed information about context and antecedents,

predict participant responses in other contexts. The next action in this SIA phase is to consider "Significance of [the] changes." One way to do this is to prioritize them on a risk assessment matrix, where severity of consequence is plotted against the likelihood of its occurrence. The next action is to plan alternatives to the project.

In the third phase SIA plans how to deal with impact. The first stage is to "Address negative impact" by, for example, use of mitigation hierarchy. This starts from actions seen as the most desirable and continues to those considered the least desirable.

1. Avoid
2. Reduce
3. Repair
4. Compensate in kind
5. Compensate by other means.
   (Vanclay et al., 2015: 51)

To some extent we could consider that the concept 'seek validity by design' in Saville's model is an attempt to avoid negative impact, certainly in terms of offering clear information about the test construct and its intended purposes. Whether this in the end will lead to avoidance of negative impact is another question. Alderson (2014), echoing earlier sentiments expressed in Alderson & Wall (1993), states:

> I don't think washback is necessarily caused by the test. It is caused by the use or misuse of the test. Of course we knew that the use of the test relates to validity, but is a knife the cause of murder?
> (Lancaster University eprint, no page numbers)

We could argue that the attempt to draw up models of test impact in itself constitutes the desire, at least, to lead to its avoidance or reduction, simply through deepening our understanding of how it works. What the SIA model appears to offer, however, are interesting practical guidelines, some of which may assist in identifying negative impact and dealing with it. One such strategy, for example, in the third SIA stage, is to 'establish a grievance mechanism' which allows all those involved to voice their concerns. To ensure that this is not simply used cosmetically, a complete plan for who is to investigate, verify and respond to grievances, how and when, forms part of the mechanism. Throughout the SIA process ethical issues are key. Having and showing respect for all parties involved is stressed. Respect here is demonstrated by acknowledging concerns raised, treating them all seriously, and seeking to respond to and accommodate them.

This overview of SIA has been brief and selective, but some aspects of the approach seem to bear promise for adaptation in investigating the impact of tests in education. Chalhoub-Deville (2009: 125) draws attention to the proactive aspect of SIA and, in the context of *No Child Left Behind*

(NCLB) in the US, proposes that cooperation between testing experts and policy-makers prior to its implementation could have mitigated some of the negative consequences. She advocates use of SIA at the planning stage of educational reforms.

1.3.5. Review of Studies of Impact

As there is a complex and dynamic relationship between the different contexts and the different stakeholders we will see in the review of impact studies of language tests below that notions of impact become blurred with notions of washback. It is precisely the dynamic nature of the interactions between individuals *within* the school (pupils, teachers, school principals), and with those coming in from 'outside', yet having a vested interest in what takes place in a given school (parents, external evaluators, local educational authority representatives) which causes this blurring. The same 'players' are involved both at the micro level, that is within the school with an in-looking focus, and at the wider level of the school in a community, where those same 'players' (teachers, school principals, pupils) and actions relating to them become the subject of interest of a wider audience. Washback, understood as what teachers and learners do in response to a test in their teaching and learning, does not take place in isolation, but in a specific social and educational context, in a community, in a region, in a country. Neither is immune to the other and there is a constant process of interplay between the levels. Thus impact studies may include aspects more often described as washback, but differ from studies of washback in that their main focus of interest is on the wider implications of a test, over and beyond the classroom and the school. We will also see (in Chapter 2) that other factors, ostensibly unrelated to the 'players', such as legislation and educational policy, enter the picture, interact with other features and filter through to the micro-classroom level.

Saville (2009: 25) sums up:

> Language test impact concerns the effects of a test or assessment procedure on various systems in a range of contexts. Because of the social and cultural implications, the scope of impact research includes the effects of a test on people's attitudes and motivations, their views of educational standards and choice of suitable qualifications, as well as ethical issues such as fairness, discrimination and equal opportunity.

Studies of the impact of language tests

This section will provide an insight into studies of the impact of language examinations. In the first part we look at studies of reforms involving examinations in the school system in different contexts: the 'O' Level examination in Sri Lanka, the Certificate of Education in English exam in

Hong Kong and the Progetto Lingue 2000 project in Italy. The second part describes studies on the impact of international exams in various contexts around the world and focuses on the IELTS impact study, the TOEFL impact study and the recent work of Cambridge English assessment.

The Sri Lankan O Level Impact study

We begin with the Sri Lankan Impact study which has been referred to earlier, as this was the first reported study of the influence of a national examination in English. Information about the context of the reform was given above in the introduction to Henrichsen's Model. This study was undertaken by an external body, Lancaster University, at the request of the Sri Lankan government, with the purpose of evaluating the examination and washback from it to classrooms (Wall & Alderson, 1993: 44). The research began with the collection of data for a baseline study, conducted six months before the first exam administration, and this was followed by drawing up a list of hypotheses of the forms washback from the exam might take. From the beginning the exam was designed to be based on the new textbooks and to promote a communicative approach to teaching, which were explained in a series of guides for teachers. Thus, it was hypothesized that positive washback would be seen in balanced use of all aspects of the course books, in teaching which followed the approach outlined in the teacher guides, and in the use of classroom tests designed to assess the course book content and marked using the criteria explained in them, which paralleled the criteria to be used in the exam. The drive of teachers affected positively by the exam would be to teach the content of the course book, as they understand this is related to the exam. Negative washback was predicted to be revealed by an increased focus on reading and writing at the cost of other skills, as these skills were the primary focus of the exam. In addition it was expected that teachers would select only those task types that reflected exam content, omitting others and that they would use past papers and overt examination practice. Their main drive would be to prepare learners for the exam, rather than teaching the course book.

The main study included a series of rounds of observations held over a two year period, somewhat disrupted by the unstable political situation in the country at the time. The observations showed no clear evidence of negative washback in terms of use of the course books, (thus indicating that the course books were having the desired effect), but revealed that the *way* teachers taught was predominantly similar to how it had been during the baseline study, suggesting that the desired impact on teaching methodology, as described in the teacher guides, was not taking place. It was found later in interviews that the reason for this was that teachers did not fully understand the idea of the intentions of either the exam, or of the course book. The interviews also revealed that there had been a lack of training about the new exam (some 70% of teachers had not been trained)

and that there was limited access to printed information about the exam (50%) and to the Teacher Guides (40%). The final set of observations, one month before the exam, showed that teachers were ignoring speaking and listening (not tested in the exam) and focusing on other skills, thus indicating a narrowing of the syllabus to reflect the content of the exam. The observers also noted that exercises being done were exam-related. As observations alone raised many questions, additional data was collected via teacher interviews, questionnaires, and analysis of materials and tests prepared by teachers. The researchers stress the importance of obtaining mixed data to facilitate a deeper understanding of the impact process. They also point to discrepancies between what teachers claimed in questionnaire data and what was observed in class, underlining the need to validate what teachers claim. In conclusion, it could be said that teachers had difficulty analysing the content of the new exam, understanding its intentions or understanding what changes this necessitated in their classroom practice. This was the case even among those who had access to the Teacher Guides. This points to the crucial importance of teacher training and support during the curriculum and exam reform process. The researchers sum up:

> If an exam is to have the impact intended, educationalists and education managers need to consider a range of factors that affect how innovations succeed or fail and that influence teacher (and pupil) behaviours. The exam is only one of these factors.
> (Wall & Alderson 1993: 68)

They warn that the washback process is far more complex than envisaged in their 15 hypotheses (Alderson & Wall, 1993) and that the belief that a good test is enough to drive positive consequences is naïve.

The Hong Kong Certificate of Education in English Study

Cheng (1997, 1999, 2004, 2005) carried out a longitudinal study which aimed to investigate responses to the introduction of a new examination in Hong Kong state education, the Hong Kong Certificate of Education examination in English, which is held at the end of Year 5 secondary and has the purpose of deciding which learners can continue to the final two years of schooling (as in the UK system) and prepare for university entrance and which should leave school, making it an important high-stakes test. Changes in the examination followed changes in the curriculum, which became Target Oriented and focused on the ability to use language in the real-world through an integrated task-based approach. This was intended to change teaching practice, shifting towards greater active learner participation and away from teacher-fronted classes. One difference, however, was that the examination is norm-referenced, as its purpose is for use as selection, while the curriculum has criteria as targets. The examination mirrors the curriculum in that it includes tasks integrating input from reading and listening with writing

activities and includes an oral part with discussion and role play activities. The Hong Kong educational authorities believed that the changes would help to "narrow the gap between what happens in the exam room and the real world" (HKEA, 1993: Appendix C, in Cheng, 1997: 42). While Cheng, in her book-length account, describes her research as "a washback study", it can be seen from her aims that the sphere of interest goes beyond the classroom.

> The aim of the study was to observe how the whole education system would react in the context of change in its assessment practice and to attempt to discover the implications of the washback effect on the teaching of English in Hong Kong secondary schools.
> (Cheng, 1997: 41)

Wall (1997: 295) suggests that this is in fact an impact study, because of the wide focus of interest. Cheng examined the processes which took place at both the macro- and micro levels, with particular focus on attitudes and behaviours towards teaching, learning, methodology and materials.

Cheng used mixed methods over a two and a half year period to collect data, including quantitative information from surveys and qualitative information from interviews and classroom observation. Informants included representatives of the Hong Kong Education Authority (HKEA), school principals, teachers, panel heads (subject team leaders in school), and students. In the first phase of her research, which she describes as the macro-context, she studied available documentation on the examination and conducted interviews with key decision-makers to form an understanding of the intentions of the reform. She also observed training sessions run by publishers and tertiary institutions where information about the proposed changes were explained, and approaches and activities discussed and demonstrated (Cheng, 2005: 117–122). Publishers had been fully briefed with regard to the changes by the HKEA and so to some extent played the role of change agents in the process, disseminating information about the new exam, listening to teachers and discussing their concerns with them and providing training in strategies for dealing with the change.

In the micro-context Cheng was able to visit schools and compare two cohorts of students, both taught by the same teachers, where one were preparing for the new examination, while the other, a year ahead, were still working to the former exam (Cheng, 1999). The study included a baseline phase before the introduction of the new curriculum and then the main phase a year later in the first year of its implementation leading up to the first administration of the new exam. Three of the teachers who were subjects in the baseline phase were the subjects of the main study and both phases included observations when lessons were also filmed. Data included notes taken during the lesson and analysis of lesson transcripts and these were analysed using a scheme based on COLT (Frohlich, Spada & Allen, 1985). As each of the three teachers taught both the old and the new curriculum it

was possible to compare their teaching in both types of class. The teaching materials in use were also analysed.

It was found that the content of what was being taught changed between the baseline and the main study, with the types of task that were to be in the examination appearing in the main phase. While some changes were observed in how the teachers taught, with regard to the balance of teacher and student talk, this varied considerably between teachers and classes on the whole remained teacher-dominated (Cheng 2005: 226). Two of the teachers whose teacher-talking time was found to increase in Year 2 rationalized this by the fact that the exam was new and so more explanation was needed for their learners to understand what was intended. In the lessons of all of the teachers an increase in the amount of group work was observed in the second year, although the intensity of its use varied between teachers. However, the total amount of class time spent on learner-led activities was not found to vary between the two years. Teachers commented in post-lesson interviews that they were unsure how to approach teaching with the new curriculum and so interpreted it in their own ways, trying out ideas for tasks and integrated activities. No past papers were available, so the main source of exam-linked materials were those in the new textbooks in Year 2, while past papers led in Year 1. No increase in the use of authentic materials was found.

In the teacher survey study a questionnaire was issued twice, first during the year when the old examination was still in use, and then in the second year, when the new exam was to be administered, allowing for comparison. Cheng (2004) found that attitudes to the new exam were largely positive and that these became stronger in Year 2. As the exam came nearer, teacher perceptions of the changes made in it became closer to the changes intended by the HKEA, suggesting a positive washback effect (Cheng 2004: 15). In response to what they intended to change in their teaching, the teachers' strength of response to "use more oral and listening tasks" was significantly greater than in the first year, also in line with the HKEA's intentions. While more teachers understood that the exam was intended to be "more practical and closer to real life" in Year 2, a decreasing number intended to use "more real-life tasks" in class. Cheng (2004: 158) comments: "What the teachers think they would like to embrace in terms of the new exam and they actually do in teaching might not necessarily match." She hypothesizes that teachers seem willing to change classroom activities, but not their methodology or teaching behaviours. Teachers voiced concern about the level of their students' English in the face of the new exam and one possible response to this may have been a noted increase in the survey in the use of English as the classroom medium of communication. Alternatively, this could be seen as a washback effect from the oral exam, but the data do not make this clear. While teachers initially expressed concern about textbooks in the first round of the survey, this concern was significantly less in the second

round, suggesting that the new textbooks, which had become available in the interim, were perceived as helpful for preparing learners for the new exam, thus indicating a washback effect from published materials.

Student questionnaires, like those of the teachers, were administered twice, but in this case to different cohorts, one preparing for the old exam and the second preparing for the new exam. According to students, English was more commonly used as the classroom language in Year 2 (preparing for the new exam) than before, a perception which differed from the parallel question in the teacher survey, where teachers felt there was no difference. Learners also perceived changes in classroom activities, with increased amounts of pair and group work, discussion and tasks which integrated skills, indicating a washback effect. Their impression about the predominant teaching style also changed, with more learners declaring that lessons were communicative in Year 2. More students were found to be attending additional English lessons outside school in Year 2 which may have indicated a concern to improve for the new exam.

In conclusion, Cheng (2005) confirmed that washback is a change process to which all involved respond individually (ibid.: 245). She found that the HKEA's intentions in bringing change to classrooms had been partly successful, in that it changed what was being taught but not how, indicating a superficial, rather than profound change. One of the limitations of this study, however, was that it looked at a very early stage of the implementation of a curriculum and new exam. As change takes place only gradually over time it is possible that greater changes could have occurred later. This piece of research was one of the earliest to use both qualitative and quantitative data and to describe the full procedure in detail, thus it serves as valuable information for further investigations.

The Progetto Lingue 2000 Studies

With the intention of raising standards in foreign language education, to meet EU targets and improve the competiveness of Italian students, the Italian Ministry of Education launched a wide-reaching reform, throughout the school system. Taking the Common European Framework of Reference (2001) as the guiding principle, it aimed to encourage the development of learners' pragmatic communicative competence through the encouragement of a more interactive, communicative approach to language teaching, with the intention of displacing the traditional grammar-translation approach predominantly in use in schools. CEFR based targets of A1 at the end of primary, A2 for the of middle school and B1 for matriculation were set, aiming to increase continuity throughout the system. An additional aim of Progetto Lingue (PL) was to increase the use of a variety of teaching materials and promote use of ICT in the classroom. Learners were given the opportunity to join extra-curricular foreign language courses held, usually in the afternoon, in their schools. These courses, funded by the Ministry, each

lasted approximately 30 hours, were modular in form, for 15 students of similar language ability, regardless of age, and were to have clear objectives which could also serve for assessment purposes (Lopriore, 2002: 208–211). Teachers were contracted to teach on these courses and paid for them over and above their usual salary. The reform envisaged that learners on these courses would certify their levels of competence by taking external examinations offered by recognized international testing agencies. It was also Ministry policy that monitoring for the purposes of evaluation and validation of the project would take place at local, regional and national levels.

Cambridge ESOL, one of the bodies approved by the Ministry to offer examinations for certification of the CEFR levels of competence, applied to conduct a small-scale impact study. Their aims were two-fold: to assess the impact of the PL2000 on teaching, pedagogic materials and assessment practices, but also to evaluate the role and appropriateness of Cambridge ESOL examinations in the process and identify ways to support teachers in their use (Hawkey, 2006: 44–45). Although not overtly stated, we assume that a further aim was to promote the Cambridge exam suite and to increase the number of test takers in Italy. This is based on information that baseline data was gleaned from information on Cambridge ESOL test takers in Italy prior to PL2000 and that one of the measures of project outcomes was to compare results on Cambridge exams for learners who had participated in PL2000 courses with those who had not.

The Progetto Lingue 2000 (PL2000) impact study made use of mixed methods, including student and teacher questionnaires, classroom observation, and both individual and focus group interviews. Document analysis was also used (Hawkey, 2006: 77). Informants included school principals and managers, teachers, parents and learners. Additional data came from entries to an essay competition, partly sponsored by Cambridge ESOL, where teachers were invited to write on the topic 'What PL2000 means for me as a teacher' (Hawkey, 2006: 83). Seven schools, in different parts of the country and representing different stages of education, were selected as case studies. The schools chosen as case studies had chosen to take Cambridge ESOL exams (Saville, 2009: 147). Following a brief pilot, which took place in three of the schools, each school was visited twice over the course of one school year, once early in the year and a second time, towards the end (Saville, 2009: 133). On each visit 1–2 lessons were observed and filmed. Interviews were arranged, in some cases focus group meetings took place, and students completed questionnaires. There was variation between the amount of research activity in the different schools, with student questionnaires the only constant measure, with the exception of the sole elementary school in the sample which was excluded from the survey study as it was felt learners were too young (ibid.).

The study aimed to evaluate washback from PL2000 on teaching methodology, behaviours, materials and assessment practice and on the

attitudes of teachers, learners, school principals and parents. The number of Cambridge ESOL exam entries and learner results was given as a measure of 'performance' on the program (Hawkey 2006: 135). The effect of PL2000 on school principals was another area of interest, as was impact on in-service teacher education and resource centres for teachers.

It was found that teachers involved had a clearer understanding of a communicative approach to language teaching (CLT) and felt more confident about teaching in this way. They interpreted this as using more of the target language and giving more opportunity for use of the language, particularly in oral skills. Hawkey reports that learner views differed somewhat, with lower perceptions of the amount of learner talk (ibid.: 138). Hawkey (2006b: 245) compares rank ordering of student and teacher perceptions of frequency of classroom activities and found that while teachers placed pair discussions second, learners placed it in eighth. One area of difficulty appeared to be lesson planning and related class management. Analysis of observations showed that teachers needed to plan more carefully so as to bridge smoothly from one activity to the next, to set up activities efficiently and avoid learners being bored or inactive. Excessive amounts of teacher talk were reported in 50% of the 20 lessons observed. There also seemed to be difficulty with error correction and with teacher understanding of the differences in focus between accuracy and fluency-based activities. Teachers were observed to be interrupting learners during speaking to correct them at inopportune times. However, in focus discussions and interviews teachers themselves identified planning and classroom management as areas for development and this was pinpointed as a focus for in-service seminars. In sum, CLT appeared to be being attempted on the courses, with teachers actively seeking to learn more and improve their practice, which indicates a positive washback effect.

The impact study found that some of the PL2000 courses adopted external exam-related course books, while others chose books with a CLT approach. A wide range of different additional materials were observed in use, including ICT, suggesting that the Ministry aim was being met in this respect.

With regard to attitudes, teachers reported that taking part in the courses was having a positive impact on learner motivation, seen as greater engagement and interest. The teachers themselves responded to the PL2000 initiative on the whole with enthusiasm. One drawback was seen as the increased workload the project entailed, particularly in terms of administrative duties.

School principals, taking a different perspective, were positive that focus on communication would lead to better opportunities for learners in the future, and felt that PL2000 was beneficial, as the school programs were enhanced. They saw that it had benefits in terms of training for the teachers involved and reported developments in teacher language proficiency. Students were also perceived as making progress, being more motivated and that increased numbers were interested in joining courses for the following year. PL2000 was perceived as satisfying parental demands for increased

foreign language provision and as being well-received by parents. However, there did seem to be some drawbacks, with questions about equity arising. One head reported that those learners who most needed support did not attend the courses, rather those from 'the most motivated families' (Hawkey 2006: 149). As courses were offered outside normal school hours this was a potential issue. There was a suggestion that holding courses in afternoons was problematic in terms of organization for the school and staff as well, with teachers of other subjects concerned that time was being taken away from doing homework. This was echoed in some of the parent interviews.

Little direct information about students was obtained as this was limited to questionnaires, which were quite brief (see Hawkey, 2006: 212). Comments about students were also gathered from interview data with other participants. Students reported that they felt they were making progress, especially in speaking skills, and that external exams made them work harder. In terms of how they used English outside class, it was seen that there was an increase in use of the internet, reading books and letter writing during the school year. This question is not directly related to the course or to the exam, and so it is hard to describe this as a washback effect. Increase in internet use is just as likely to be the result of improved access (this was 2001–2002) as involvement in PL2000.

Parents felt positive about the program, seeing the ability to communicate in English as a key skill for their children's futures. Many reported their offspring were satisfied with and enthusiastic about the courses.

Cambridge exams were seen as enjoying prestige and having wide recognition. They were considered to provide a series of goals for learners, which was perceived as motivating. The fact that they covered all four skills and were linked to CEFR showed that they were in keeping with PL2000 aims. Some doubts were expressed however about the way results are reported and about the lack of detailed information given to individuals about their performance.

Cambridge ESOL responded to information obtained from the impact study by preparing and planning a program of seminars for teachers, designed to address areas found to need attention, and to show how to develop communicative teaching skills. A new dedicated website was set up to disseminate exam information and support materials. Saville (2009: 151–152) suggests that this shows washback from the education system to the exam provider and is evidence of examinations in schools being part of a complex system.

Progetto Lingue Impact Study Revisit (PLISR)

In 2011–2012 Cambridge, now known as Cambridge English Language Assessment, returned to the same schools to conduct an iterative impact study, the Progetto Lingue Impact Study Revisit (PLISR). The research questions and the observation schedule remained the same as in the first

study and similar informant groups were involved. One difference was in the organizational design. This time qualitative data from interviews and class observations was obtained and analyzed and only then were questionnaires designed, based on evidence from the qualitative data. In this way it was possible to validate findings more effectively (Ellis, Hawkey & Docherty, 2014: 16). The questionnaires were disseminated more widely than in the initial study, covering 14 regions as opposed to 3, distributed online and involving a substantially larger sample of teachers (n=225 as opposed to 11 in 2001).

Ellis et al. (2014: 11) classified activities listed in the survey as *potentially* communicative activities (e.g. small group work) and *potentially* non-communicative activities (e.g. students doing grammar activities) and found that while teachers claimed to use communicative tasks frequently (with the exception of role plays, which were less popular), some less communicative activities, such as listening to the teacher speaking to the whole class, or grammar exercises, were also very frequently chosen. However, the classroom observation (18 lessons) showed a variety of good examples of communicative activities in use and noted warm teacher-student relations. The authors sum up:

> There seems little doubt that the language-as-communication message of *PL2000* has been received and absorbed into teaching practice more strongly over time although teachers' efforts in this area have not always been fully achieved. (Ellis et al., 2014: 18)

Findings with regard to use of materials and ICT echoed those of 2002, with stronger presence of digital technology than formerly, as might be expected with the developments in this field. The authors observe that PL2000 continues to encourage teacher development, especially in this respect, with references to ICT commonly occurring in the interview data. For these initiatives to truly promote communication, however, depends on the pedagogy employed.

External examinations were seen to play a strong role in PL2000 and perceptions of the Cambridge exams remained positive. Cambridge noted a significant increase in the numbers of young people taking their exams in Italy between 2001 and 2010 with strong performances at levels B2-C2, where there were three times the number of candidates. This might suggest raising of standards of English in line with the Italian Ministry's aims.

In conclusion, trends observed in the first PL2000 impact study were found to have been maintained and to have strengthened in the subsequent 10 year period, indicating that positive washback on teaching methodology and materials has been effected. However, it should be remembered that these courses run parallel to the main stream curriculum and are optional, as are taking external certificates.

The fact that Cambridge conducted a follow-up study is of interest and is evidence of their intention to build impact, validation and regular monitoring of exams into the system of test development as claimed by Saville in his 2009 model.

Studies on the impact of international exams in various contexts

While Progetto Lingue 2000 involved international examinations (those described above were from Cambridge English Assessment, but other exam boards were also involved in the project, both for English and other languages) it was included in the previous section for the reason that it focused on the use of international examinations in an education system. The studies which will be outlined below focus on international examinations which are used for a variety of purposes, mainly selection (IELTS) or certification (TOEFL) and exist outside the education systems of most countries.

The IELTS impact studies

IELTS is a high-stakes test used for certifying level of proficiency in English of candidates for the purposes of entering university, migration to Australia, Canada, New Zealand and the UK, or employment. It is available in two versions: academic, for university entrance and general for other purposes. Scores are given in bands from one to nine, the highest, described as 'expert user'. Very high standards of security are maintained at test centres (IELTS website, accessed 19.08.16) to ensure reliability. In 2015 it was taken by approximately 2 million candidates.

In existence since the mid–1970s, when it was known as ELTS, it was revised and became IELTS in 1989. IELTS underwent further revision in 1995, at which time it was decided to study the implications of the changes. Originally a British Council initiative, UCLES (now Cambridge English Assessment) became involved in the early stages and it was in cooperation with Lancaster University, under the direction of Charles Alderson (who had been involved in the Sri Lanka 'O' level impact study), that changes were made to the exams and plans for an impact study drawn up. Now IELTS is run by an international consortium composed of the British Council, IDP, IELTS Australia and Cambridge English Assessment.

The decisions to trace the impact of IELTS came at a time when Cambridge had adopted a unitary view of validity (Messick 1996, see Chapter 2) and were making a commitment to including impact in their test development as a part of validation. At the beginning of the revision period, in the mid–1990s, the main focus was to study washback from the exam on teachers, teaching, learners and learning and Hughes' tripartite washback model was adopted, looking at participants, processes and products, which was described earlier. This model was, however, to evolve, as understanding of the mechanisms of washback became better understood, a process which was to culminate in Saville's 2009 'impact by design' model. The second phase of the IELTS

study focused on the influence of the exam on teaching materials. At the same time the sphere of interest began to expand beyond the classroom, or teaching institution, to the wider societal milieu. Areas of concern also underwent a change and so the third phase focused on the attitudes and opinions of various stakeholders, both those involved in the micro-context, for example, in centres where learners were preparing for the test, but also beyond that, in the centres which ran the examination and in places where IELTS certified candidates were admitted, such as universities. The final phase of the IELTS study looked at impact on the test population, studying psychological aspects such as motivation and test related anxiety and collecting demographic data to allow comparison of the scores of different sub-groups to evaluate test fairness.

Over the phases of the study careful production of a variety of instruments took place, each undergoing a rigorous trialling, followed by extensive discussions and deliberations, standardisation, validation and a series of revisions. These processes are meticulously documented in Hawkey (2006), which provides a useful insight into the production of reliable and valid research instruments. The result was a set of questionnaires, a classroom observation schedule, clear plans for analysis of qualitative data and how this could be condensed for reporting purposes, and finally, guidelines on how to index, catalogue, and store and retrieve the large amounts of data such a project would produce. These instruments would be used again in the Progetto Lingue 2000 studies, and in the many other impact studies in which Cambridge have been subsequently involved (see below).

Findings from the phases were fed back into the test production process and used to inform changes. For example, information that time pressure during the reading test was causing high levels of anxiety (registered on a Likert scale) obtained from a large scale questionnaire survey conducted in Phase 4, led to revision of the number and length of texts and time allocated for this paper.

In addition to the studies conducted by Cambridge, a number of other studies were commissioned, or funded by the British Council and IELTS (e.g. Hayes & Read, 2004). This is often research in micro-contexts and so more appropriately classified as studies of washback. Saville (2009: 112) explains this with the analogy of types and tokens (as in type/token analysis, usually associated with study of lexis, where types would be a lexical category and tokens are examples of the type – thus a type would be noun, and dog, cat etc. would be associated tokens). He sees small-scale studies in micro-contexts as tokens, which illustrate the type "context". This interpretation could be seen as specific to this type of international examination, which takes place worldwide in a myriad of different settings. However, in large countries, with great regional variety, such as China, it could also be applied to national examinations. In order for the test provider to understand the context and the impact of the test, Saville argues, it is not enough to have quantitative

data from user surveys. These need to be supplemented by richer views taken from many different micro-contexts. Taken together with the macro--survey data, the micro-context studies provide a more informative picture of test impact.

The idea of outsourcing of studies continues as an ongoing process, with a number of research grants being made available annually. Some of these are to investigate specified areas, whilst others are open to tender and may also include wider-scale studies. These studies are written up and reported in online reports on the IELTS website. Since 1995 some 90 studies have been conducted relating to IELTS, although impact and washback constitute only part of these. Areas researched under impact and washback include:

- impact theory and practice (Chappell & Bodis with Jackson, 2015)
- increases in test scores
- stakeholder attitudes and views (e.g. test takers, teachers, administrative staff) (Murray, Cross & Cruickshank, 2014)
- predictive validity (the extent to which candidates are really prepared for study or work in comparison with the test score they received) (Lloyd-Jones, Neame & Medaney, 2007)
- preparation for the test
- uses of test scores (Gribble, Blackmore, Morrissey & Capic, 2016)
- performance of candidates in relation to IELTS scores (e.g. number of hours of study and average score gains)
- (based on filters on the Research page of the IELT website and IELTS call for proposals for impact and washback studies, 2016)

The differing focus and scope of these studies serves as a good illustration of the concerns of impact research. Some aspects of this, such as the use of test scores, will be explored further in the next chapter.

The TOEFL impact study

The TOEFL (Test of English as a Foreign Language) exam originated in the early 1960s as a result of cooperation between numerous institutions preparing candidates from non-English speaking backgrounds for entry to US universities. Conceived at the height of behaviourist, discrete point approaches to testing, the exam consisted of closed multiple-choice items with an emphasis on receptive skills. Over time the Educational Testing Service (ETS) took greater responsibility for it, until today the TOEFL test is wholly owned and administered by them.

In the early 1990s, in view of changes of views on communicative competence (e.g. Bachman, 1990), it was felt that there was a need for direct testing of productive skills. This was reinforced by concerns that it was possible to receive a high score on the TOEFL test, but to be ill-equipped to cope with the speaking and writing demands of university courses (Jamieson, Jones, Kirsch, Mosenthal & Taylor, 2000). In other words, the test was felt

not to be adequately meeting its purpose as selecting candidates suited to university entrance. Further impetus for change came from teachers, who felt that the form of the test, with its use of multiple-choice items and focus on receptive skills, was having adverse effects on classroom teaching among those preparing candidates to take the test. (In the 'old' paper-based TOEFL, writing was not included in the main exam but could be taken as an additional, separate test.) The final factor driving towards test revision was the decision to move towards a computer-based test (ibid.). From this we see that impact issues were key in making the decisions for change, coming both from teacher stakeholders in test preparation settings and from end-users of the test, those in tertiary institutions who had questioned the predictive validity of the 'old' version of the test. The result of a long process of revision was the TOEFL iBT® test, which was launched in 2005–2006. Through changing the approach and including direct testing of speaking and writing, using tasks which integrate skills and so more closely reflect the nature of language use in university contexts (such as listening to a lecture and taking notes, or writing with the use of written source materials), it was anticipated that the test would have a more positive impact on teaching and learning in preparation courses (Wall & Horák, 2006: 1–12).

With the aim of investigating whether this hypothesized positive impact was in fact taking place, Wall and Horák, commissioned by ETS, designed a longitudinal study. Following the form of other work in which Wall had been involved (e.g. the Sri Lankan 'O' level impact study), the research began with a baseline study, aimed to record the situation in teaching contexts before news, or information, about the new form of the exam had reached the participants. This would be followed by a subsequent phase or phases, investigating the transitional stage as teachers became aware of the planned changes and prepared for them and a final phase, which would examine the situation post implementation of the new test. Information obtained from the study was intended to be fed to the test design team, with a view to improving test quality and avoiding negative consequences.

Based on Henrichsen's Diffusion of Change model (see description above) Phase One was designed to determine the Prior (Antecedent) conditions, Phase Two to investigate factors facilitating or hindering the change process and Phase Three to ascertain if the test designers' intentions to have positive impact on the teaching/learning situation in test preparation classrooms, had been achieved through the changes introduced to the test (Wall & Horák, 2006: 6).

The researchers' initial task was to understand the positive impact the test designers anticipated through analysis of documentation of the test construct framework and "interviews" with members of the design team, conducted in the form of open questions sent by email, with follow up questions for clarification.

Ten institutions across Central and Eastern Europe in Bulgaria, Croatia, Lithuania, Poland, Romania, and Slovakia were identified. This region was selected for the reason that as contacts with ETS had been limited it was felt it would make a good "test case for the extent of and barriers to the diffusion of knowledge about innovations in the test and implications for teaching" (Wang, Eignor, & Enright, 2008: 299). Difficulties were experienced in finding centres where TOEFL courses were being offered and a mixture of preparation courses, some for the paper-based (PB) version and some for the computer-based (CB) version of the test had to be settled on, rather than exclusively CB preparation courses as had been planned, as there were far fewer courses on offer than had been anticipated by ETS.

Instruments for data collection were created based on the researchers' prior experience of impact studies, on analysis of information and documents pertaining to the existing TOEFL test in both PB and CB versions, and on expected forms of impact from the new test. Structured interview protocols for teachers, students and directors of studies and an observation sheet and checklist of expected lesson features and areas of focal interest were prepared. The authors decided to gather rich, qualitative data given the small size of the sample. Observed lessons and interviews were audio-recorded and transcribed. To help identify which variables in lessons were teacher-related and which test-related, each teacher was observed twice, where possible once teaching a TOEFL preparation class, and once teaching a different class of similar language level. However, this was not possible for all of the teachers. Data was collected in 2003.

Using a coding system based on the one used by Wall (1999), but modified to deal with specific features of the new study, the interviews were coded individually by both researchers and then compared to ensure reliability. Each of the observed lessons was summarized using information from the observation sheets and the checklists and one summary comparing each teacher's test-preparation and non-preparation classes compiled.

Wall & Horák (2006) provide detailed information on different aspects of the TOEFL test, based on a synthesis of information obtained from all their sources during the baseline study. The listening part of the test for example, was perceived as problematic in several respects: the intonation and speed of delivery were felt to be somewhat unnatural; topics seemed to deal predominantly with campus life in North America, which was found to be irrelevant to many of the students who planned to study abroad, but in countries other than the US; there were memory issues with the CB version of the test, as students did not see the questions before listening, first having to listen to the text, and then hearing the questions without being allowed to take notes; this was also felt to be unrepresentative of how students would need to listen in real-life in a university; teachers found no help in how to prepare students for the listening test other than by doing many practice

examples with them. For information on this and other aspects of the test, see Wall & Horák (2006: 32–72).

In general, the test preparation materials were found to play a strong role in course planning (reflecting Alderson & Hamp-Lyons, 1996), and many of the teachers simply followed the book. Few of the teachers seemed to have a clear idea of what the TOEFL exam aimed to test, or how to prepare their students for the skills they needed in the examination, other than what was given in the course materials. This may have been exacerbated in some cases by lack of teaching experience, or of training in English for academic purposes. However, the researchers noted that they themselves felt that information on the skills needed and how to teach them was lacking in the TOEFL preparation materials and on the ETS website. The teacher-subjects all viewed vocabulary as being of key importance in success on the exam, despite the fact that it is not tested separately, and spent a sizeable proportion of time on it in class. Writing was perceived as the most difficult part of the test for students in the region and so received a lot of class time, with various approaches to teaching observed. Many teachers expressed doubts about their understanding of the TOEFL writing assessment criteria and scoring scheme, which was reflected in their attitudes to and practices in marking. Speaking was generally not found to be a major course component, which was mostly explained as being because it was not part of the test. This aside, English was found to be the main classroom language for communication and was used to answer student queries.

From the information obtained, the researchers predicted areas to focus on in the second phase of the study, by comparing impact foreseen by the test designers with what had been learnt. They were interested to see how speaking would be treated as information about the spoken test became available and whether there would be more activities designed to develop this skill. As the new test would no longer contain a separate test of grammar (structure), the test designers had predicted less overt grammar teaching would take place. The researchers, however, based on the fact that they had observed considerable time spent on teaching vocabulary in class, when this was no longer tested separately, predicted that grammar would continue to feature largely in lessons. It was expected that course books would continue to play a key role in teaching and learning, but that there would be more evidence of computers in class, due to the fact that the revised test is computer-based. However, as impact research has shown that links between tests and teaching are not causal or linear, Wall & Horák also planned to monitor closely communication about the test revisions in the region and how this would be responded to, for example, by management, in determining the resources available in their schools.

In the second phase, Wall and Horák focused in detail on case studies of 6 of the teachers who had been subjects in Phase One. This was simply the number of teachers who agreed to continue and no selection process

occurred. They came from five countries: Bulgaria (two teachers), Croatia, Lithuania, Poland, and Slovakia. The aim was to obtain quality data which would help explain the change process. The authors cite Chapman & Snyder (2000, see discussion of their model above) who had emphasized that teachers need to have full understanding of a new idea and be aware of how they can put it into practice in their work before they can implement it. Wall & Horák (2008: 5) felt that it was crucial to understand the process through which teachers became aware of the changes in the TOEFL test and to track how they adjusted to these and accommodated them in their teaching. A year had elapsed since the end of Phase One, during which the final version of the revised TOEFL had been announced and information about the changes had started to be disseminated when Phase Two began in January 2005. The new TOEFL test was planned to start in October of that year.

Each of the teachers was interviewed via computer, twice a month, over a period of five months. Each month they received a set of open core questions asking about information and news they had learnt about the new exam, with repetition of the questions allowing their knowledge base to be tracked. In the first month questions were also asked about what had been learnt in the interim period since the end of Phase One. Data was sent and collected by email and the teacher responses were then discussed in the first of the monthly interviews. In addition, teachers received a task to do each month, which required them to do a reflective written activity on focal areas of interest as follows:

- The nature of TOEFL classes
- Teachers' awareness of the TOEFL, both the current version and the new version
- Teachers' reaction to the integrated writing tasks
- Teachers' reaction to the speaking test
- Possible content and methodology of future TOEFL preparation classes
  (Wall & Horák, 2008: 13)

Some of the tasks included practical exercises based on input material, such as rating sample written work using the scoring criteria. (For a full description of the tasks see Wall & Horák, 2008: 11–18). Teachers returned their responses by email and these were then the subject of the second monthly interview. The design of this study is of particular interest to researchers wishing to thoroughly investigate impact issues through a qualitative approach and the way the tasks are designed gives a good example of how teacher cognition processes might be captured. It may, however, raise the question of whether this was not an intervention, rather than an 'observational', ethnographic study, as it could be argued that engaging in the tasks 'pushed' participant knowledge beyond what they may have learnt in the naturalistic context. Ethically, it can be argued that

this was acceptable, as the researchers' indirect intention was to support the teacher-participants in becoming aware of and having the opportunity to acquaint themselves with key information about the new exam and the training materials relating to it. This is borne out by some of the participants expressing their thanks and acknowledging that engaging in the research project had helped them learn things which they would have had to find out for themselves, a process they described as arduous and time-consuming. Wall & Horák themselves discuss their intervention as being an issue (2008: 70), but conclude that they tried to minimalize the effect.

Interviews were conducted either by synchronous messaging (texting via the internet) or by emails sent and received during a fixed 'time-slot'. This meant that the communication occurred with a slight delay and so was asynchronous. This type of 'interview' had the advantage of providing a ready-made written record, thus avoiding the time and expense involved in transcription of recorded talk.

Wall & Horák (2008) provide a meticulous analysis of the responses to the tasks, summing up that while it was clear that the teachers' knowledge about the new test increased considerably during the five months, their understanding of it varied greatly and was entirely individual (ibid.: 56). Of key concern to the researchers was the search for evidence of how the teachers were responding to the exam revisions in their work, Messick's (1996) 'evidential link' of test impact. As the start date of the new test was not far off, it was expected that teachers would show how they planned to change the upcoming test preparation courses in response to the revised exam. While their surface knowledge about the exam had changed, there was still little to show that deeper understanding had grown and there were indications of confusion in understanding the test construct and what it aimed to test.

A significant event occurred at this point (March 2005), which was an announcement that the date planned for the launch of the new test had been postponed and it was not clear exactly when it would begin in the different countries involved, as a staggered phase-in was envisaged. By the end of the Phase 2 data collection, commercial textbooks based on the new exam were not yet available and teachers were clearly influenced by this to some extent, as a considerable part of their planning was book-based (see Phase 1) and there were signs that their absence was the cause of some uncertainty. The strong role of course books in washback studies has been discussed by Andrews (1994b), Lam (1994), Read & Hayes (2003) and Spratt (2005), and Wall & Horák note that their study has the potential to examine the role the course book has in shaping teachers' understanding of a test and in determining how they prepare their learners for it. Obtaining more insight on this aspect was highlighted as a focal area for the next phase. Despite the absence of new books, it was found that the teachers were beginning to think and plan the *content* of new courses, but were less clear about how they intended to implement this. For example, the new exam allowed note-

taking throughout, a fact which the respondents referred to as an important change, but only one of the six mentioned how she intended to introduce and practice this skill.

Wall & Horák had decided that there would be evidence that positive impact was taking place if the impacts intended by the test designers were found in the classroom preparation courses for the revised TOEFL test, and they emphasize that it is the factors which facilitate or hinder this process which they are trying to uncover (2008: 73). One area they draw attention to is *channels of communication,* stressing their importance in the dissemination of information. These can be divided into two main types: mass media and interpersonal (Rogers & Shoemaker, 1971: 24), but remembering Henrichsen's model, these can also interact with each other, with the result that a message may either be clarified, through discussion with others, or clarified as a result of accessing information about it in multiple sources and undergoing a form of personal-mediation process. On the way it may become distorted if interpretations of others, obtained either through personal contacts, or in the media, are misleading, unclear or erroneous, or if the person themselves misinterprets something. Wall & Horák (2008: 66–71) point out, for example, that it was members of school management who usually attended conferences which disseminated information about the new exam and then transmitted this to their staff, with instances recorded where the information related was transformed, by 'selective-editing' (parts omitted), or by introducing emphasis which had not been present in the original information, thus offering a personal interpretation rather than presenting facts. If the manager is not an EFL expert, as was the case in this study in four of the five settings, then the potential for distortion of key information is greater. This has a potential cumulative effect, as it may result in teachers not understanding important points, or not being informed of key factors and so they may fail to fully understand the change (here the new exam), which is one factor which hinders the implementation of an innovation (Chapman & Snyder, 2000). The study also made clear the important role played by the test-provider's website (ETS) which was used by the teacher-participants as the primary source of information (Wall & Horák, 2008: 91–92). While the test-provider cannot control the content of other internet sites, providing clear and unambiguous information, sample tests and scoring rubrics, sample student answers with commentaries and training advice on their own website can serve to diffuse the potential negative effects of 'unauthorized' interpretation and speculation present elsewhere. In the final phase of the study the researchers return to this point (Wall & Horák, 2011: 133–134), adding that there is also a need for explication of the test construct, making clear to test users what the test intends to assess and how this is operationalized, in ways which are easily understood.

Subjects in the research study also point to factors in the wider context potentially influencing courses preparing for the new exam. As some of the countries involved were new entrants to the EU, this may mean that exams acceptable in European contexts, and particularly Britain and Ireland, edge out TOEFL courses, especially as the UK-focused Cambridge exams were already more popular in the region. Another contributory factor might also be test-taker perceptions. If candidates feel that the new test is harder than the existing one, and the inclusion of speaking as a compulsory part may have this effect, then test-takers may seek alternatives.

As there was a delay in start date of the new test, an additional phase was introduced in the study to capture more data about the change processes. In phase 3 of the study Wall & Horák (2011) focused on the new course books and their role in the teachers' planning process. The subjects included 4 teachers, 3 of whom had participated in both of the previous phases and a fourth, who was added at the request of ETS, had taken part in Phase 1 after submission of the Part 1 report, and who came from a Western European country. It was explained that the test providers wanted a point of comparison with the Eastern and Central Europe teachers. Data was collected by means of two applications of tracking questions which traced the participants' access and response to information about the test and test-related preparation materials and through two computer-mediated 'interviews' (see description above), based on two tasks where the teachers did an activity and returned a written response by email as in Phase 2. The first of these targeted attitudes to and use of the TOEFL CB version of the test, and the second attitudes to and use of TOEFL iBT course books (Wall & Horák, 2011: 27). Each of the tasks had several parts and was very detailed, in an attempt to compensate for the fact that classroom observation was not possible because of budget limitations. In addition to data from teachers, analysis of 14 course books (named as being used by the participants) was carried out, using specially created criteria. Course books included both those preparing students for the 'old' exam (8) and materials for the new TOEFL iBT (6). A focus of this was to try to ascertain the approach to teaching advocated in the books with an aim to seeing if this would be a contributory factor influencing the participants' methodology. The analysis found that changes were mostly in terms of content, although small changes were observed in two of the new books, which encouraged use of prediction and pre-listening and pre-reading tasks.

It was found that teachers were critical towards the new books and had taken great care in making their selections. They noted that there appeared to be some discrepancies between aspects in the new books and information or samples available on the ETS website. In teacher reports of planning and use of the materials in their lessons there was little information to suggest that their approach had changed much from Phase 1, with the exception of allocation of time to different skills, where a marked increase in time spent on speaking was identified.

The fourth and final phase of the study aimed to determine if the subjects' teaching on TOEFL preparation courses had changed between 2003 (when the 'old' TOEFL was being used) and 2007 (approximately one year after the launch of the new test in the region) and if so, whether these differences could be attributed to the changes made to the TOEFL, now iBT, test. The main subjects were three of the four teachers who had taken part in Phase 3, two from Eastern/Central Europe and one from Western Europe. Data collection took place in 2007 and was planned to follow the same organization as Phase 1, with semi-structured interviews and lesson observations, although this time informants were teachers, the director of studies in the three institutions, but students were omitted, as the focus of the study had been narrowed from Phase 2 onwards. In this phase there were no non-exam classes observed, as this was not practically possible. Before the on-site visit, questions tracking the teachers' involvement with the exam were sent by email and responses analyzed. The timing was that face-to-face teacher interviews would take place after at least one lesson had been observed so that both answers to the tracking questions and queries arising from observation could be responded to by the researchers.

Wall & Horák (2011: 63) once again emphasize the need to establish 'an evidential link'

> The first aspect to consider is whether the teachers were aware of how the TOEFL iBT differed from the CBT. If they were not aware of these differences then it would be difficult to attribute any changes in their teaching to the changes in the TOEFL.

(CBT is the computer based version of the 'old' TOEFL test). Evidence of washback in teaching would be manifested in the amount of time spent on the different skills in class, methods, materials and the presence of any information or tips relating to the exam in the lesson content (ibid.).

Evidence was found of change in the content of how the four skills were taught, which could be described as positive, as the content reflected that of the tasks in the TOEFL iBT. The proportion of time spent on development of speaking and writing had increased substantially, while time given to overt teaching of grammar and vocabulary had decreased to a mere trace. This was now dealt with as a focus on form following productive tasks, or in response to learners' needs, during activities, which could also be seen as positive and in keeping with the anticipated washback. All the teachers were using materials designed to prepare learners for the new exam. They also explained scoring rubrics from the test for both writing and speaking to their students and these were used by two of the three teachers when marking.

In terms of methodology, the situation was more complex. Change was observed in approaches to the teaching of writing from two of the teachers with, for example, brainstorming of ideas and some student interaction.

Integrated tasks were being used (as in the new exam) by all the teachers and they also offered guidance in how students could tackle the new tasks and gave written feedback. Two of the teachers were also attempting new approaches to teaching speaking, for example, by working on increasing learner confidence in different ways, with the aim of helping learners ready themselves to give longer responses to prompts and to deal with the pressure of there being a time limit for their answers in the test. To sum up, in two out of the three teachers there was a move away from lessons where students were merely doing practice test tasks towards lessons where the focus was on language and skills development.

The teachers were seen to use different approaches to each other. One used considerable amounts of student interaction, both in pairs and small groups. He used pre-reading and listening tasks and encouraged discussion. The researchers point out, however, that some of the influence in his approach was coming from the course books he used, which suggested use of pre-tasks. But the teacher also commented in several places that although this was a test preparation course, it was now possible to use the sort of approach that was prevalent in general classes, as this was now not at odds with the exam ethos. The second teacher, by contrast, was still dominant in the class, but had substantially increased the amount of student interaction in her lessons when compared with Phase 1. The third teacher had not changed her approach, which remained 'traditional', in that she gave input, learners worked individually on test tasks, and there was a feedback stage, usually based on written homework they had done. The difference was that students worked at computers, rather than with books, and used the self-checking facility to check their work. The teacher believed that her role was to give students as much practice as possible, as this was the way to increase their scores. In other words, methodology was being influenced by teacher beliefs and materials.

What is clear is that this study found evidence of change in methodology, in contrast to earlier studies (e.g. Cheng, 1997), and that beneficial washback could be described as taking place. That not all of the teachers changed in the same way is in keeping with earlier findings, such as Burrows (2004), who also found influence of the filtering effect of teacher beliefs. Also like other studies (e.g. Cheng, 2005) course books were found to play an important mediating role. The importance of this study in the literature on impact and language assessment is threefold. First, it provides an exemplary account of how research into test impact can be conducted, meticulously detailing the reasoning behind all the actions taken, which is invaluable in helping us understand how to deal in practice with the challenges of such research. Second, it shows innovation in Phases 2 and 3, as the first study which attempts to address how to demonstrate that changes are evidentially linked to the new exam. The design and explanation of the tasks set for the teachers in Phase 2 is in this respect particularly valuable. It appears that this study

has been prepared with a view to using it as one element in a validation argument for the TOEFL iBT (Wall & Horák, 2011: 137), an issue to which we will return in the next chapter. Finally, through following the evolution of this five-year study, we are given a deeper understanding of the character of the impact of a new exam. Assisted by application of Henrichsen's framework, through investigation of the different phases, prior to the new exam, during the change process and after its implementation, and through rich description, we come to have much fuller insight of the factors that 'hinder or facilitate' beneficial impact. Thanks to this the 'complexity' of impact referred to in so many studies both of washback and impact, begins to unravel and become clear. We begin to really understand the mechanisms at work.

Recent work by Cambridge English Assessment (formerly Cambridge ESOL) on the impact of their suite of exams

The descriptions of the Progetto Lingue 2000 impact studies and the IELTS impact study showed the early work of Cambridge ESOL in the forming of a model of impact and the creation of procedures, methodology and instruments for carrying out research on the consequences of their exams. Saville (2012) proposes the use of his meta-framework (the 'impact by design' model described above) together with the knowledge, expertise and set of instruments developed during these studies, referred to as the 'impact toolkit' (2012: 7) to anticipate impacts from Cambridge exams and work to improve any potential negative consequences and to work for positive outcomes. Cambridge adopt mixed methods approaches to impact research, combining large scale quantitative data from test results and surveys in the macro-context with rich qualitative data, derived from observation, questionnaires and interviews in a range of selected micro-contexts. These multi-probes of different micro-contexts complement, complete and validate the general trends shown in the macro-data. To avoid the danger of misunderstanding of local contexts, as this is a UK based testing agency working in numerous settings across the world, Cambridge work together with 'insider' researchers, who are nested in the micro-context and may act as expert-informants, resulting in a combined outsider-insider approach to impact studies. Saville stresses the importance of understanding the complex interplay between the macro- and micro-contexts: "Without such methods it is difficult to find out about and understand how the interaction of differing beliefs and attitudes can lead to consensus or to divergence and diversity" (2012: 7). The final aspect of the Cambridge approach is that monitoring of exams should be an iterative process, conducted over time, as we have seen was the case with PLISR (see above).

Issues of the Cambridge journal *Research Notes* in 2010 and 2012 were wholly dedicated to descriptions of such impact studies, or to pilot studies in progress. These cover a wide range of contexts and settings, from the

use of Cambridge assessments in schools (e.g. Spain, Vietnam, China) to lifelong learning (India, Egypt) or a specialist exam for teachers (Mexico, Uruguay). The number and variety of these studies demonstrates evidence of Cambridge's commitment to impact research.

These studies are summarized in the table 1.5.

Table 1.5. Recent impact studies conducted on Cambridge examinations

Country	Date	Focus of Impact Study	Authors
Greece	2006	Use of FCE in Greece	Tsagari *
China	2012	Introduction of Cambridge exams KET and PET: focus on parental perceptions	Gu & Saville
China	2014	Introduction of Cambridge exams KET and PET	Yan, Gu & Khalifa
Spain	2014	Use of Cambridge exams in single-sex primary and secondary private schools	Docherty, Gratacos Casucuberta, Rodriguez Pazoz, & Canosa
India	2014	Use of Cambridge exams in SCOPE -Society for Creation of Opportunities for Proficiency in English	Salamoura, ffrench & Emery
Egypt	2014	Use of Cambridge exams in workplace placement program	Khalifa, Khabbazbashi, Abdelsalam & Elmahdy Said
Uruguay	2008	Use of Teaching Knowledge Test	Valazza*
Mexico	2014	Use of Teaching Knowledge Test	Khalifa, Papp, Valero & Videl
Vietnam	2012	Use of Cambridge Young Learner exams in Intensive English Program, offered to selected schools. Focus on Grade 2 learners aged 7–8	Khalifa, Nguyen & Walker

* this study was conducted by an independent researcher, rather than in collaboration with Cambridge.

Cambridge English Language Assessment (2016: 34) also define a set of procedures for investigating impact which they include as part of their framework for *Principles of Good Practice:*

A Monitor who is taking the examination (i.e. profile the test takers).
B Carry out Differential Item Functioning analyses to identify potential bias.
C Monitor who is using the examination results and for what purpose.
D Monitor who is teaching towards the examination and under what circumstances, and what kinds of courses and materials are being designed and used to prepare test takers.
E Monitor what effect the examination has on public perceptions generally (e.g. regarding educational standards) and/or how the examination is viewed by

those directly involved in educational processes (e.g. by students, examination takers, teachers, parents, etc.) and/or how the examination is viewed by members of society (e.g. by politicians, businesspeople, etc.).

The combination of Saville's work on developing a model and meta-framework and the expertise grown during the early studies described above (Hawkey, 2006; Saville, 2009), taken together with subsequent research which applies these principles has made a considerable contribution to understanding the phenomenon of impact of language examinations in education systems and in society.

In this section we have shown how the notion of test impact is conceptualized in the literature on general education and applied linguistics. Next, we considered who (the stakeholders) may be affected by high-stakes tests and how high-stakes tests are used in educational systems. We looked at arguments which suggest that high-stakes tests may be used as a force for good, bringing about beneficial changes, and also gave an overview of the counter-arguments, which see high-stakes testing as having a variety of detrimental consequences. This was followed by description of several models which aim to show how impact works and explain the factors and filters which can enhance the influence of an examination, or educational innovation, or inhibit its effect. We examined Chapman & Snyder's "Conventional Wisdom" Model (2000), Henrichsen's Model of Diffusions of Innovations in ELT (as used by Wall, 2005), and Saville's model and meta-framework: Impact by design (2010). Next, there was a review of impact studies in language testing, beginning with studies of the impact of national examinations in Sri Lanka and Hong Kong, and followed by the Proguetto Lingue 2000 impact studies, which looked at the role of external international examinations used alongside the formal education system. The next studies were of the impact of international examinations in worldwide use and we reviewed the IELTS and TOEFL impact studies in detail, showing how they have both contributed to developments in the field. In conclusion, we offered a brief overview of the work Cambridge English Assessment studying the impact of their suite of different examinations, from general English examinations for various age groups, from young learners to adults, to examinations for specific purposes, such as the Teaching Knowledge Test.

We have aimed to show how, as with research on washback, understanding of the nature and mechanisms of the impact of examinations has grown. We have seen how findings from research on washback have fed into models of impact. We have noted that work from other fields, such as educational evaluation and social impact assessment (SIA), in addition to analysis of research on the impact of exams in national systems have contributed to understanding of the consequences of high-stakes examinations. We have observed how providers of international tests have become more conscious

of their relationship with the potential consequences of a test, and how they have accommodated for this in test design and validation procedures.

In the next chapter we will examine in more detail the role examinations play in education systems when they are used for the purposes of evaluation of quality and accountability. In order to fully understand the difficulties inherent in the uses of tests in this context, we must begin by deeper consideration of the notion of validity and what exactly a test score tells us.

CHAPTER 2

Tests in use

Tests and examinations are part of the fabric of our lives. At some time all of us have had first-hand experience of them and we are able to tell stories about how we felt about them. In some cases these feelings would be negative, perhaps relating to a sense of injustice, which resulted, for example, from a lack of objectivity on the part of the examiner, an unpleasant surprise in the content of the test, or concerns about the way it was conducted. Perhaps we felt we had been misinformed, misled or unfairly treated. These aspects relate to the quality of the test itself, the conditions in which it is carried out and the procedures used. Another factor of which we all have personal experience is test results. Most of us can remember waiting in trepidation to find out how we did in an important examination. The moments we remember are probably connected with results which were key in a decision-process. Passing this important, high-stakes, examination, or getting a required grade or a number of points, made the difference between being able to undertake a course of action or not. This may have been getting access to a university, obtaining a qualification, or being accepted for a job.

> Language assessments are used in the service of a variety of decisions, including student selection, certification, classification, tracking, promotion or retention in educational programs, and allocating resources to schools. In order to assure that the decisions that are made, at least in part on the basis of language assessments, are fair and equitable, we must consider the specific uses or decisions for which the test is intended and designed, and the consequences of these decisions for different groups of individuals. Equally important, we need to consider the quality (i.e., reliability, validity) of the information provided by the assessment, and the relevance of that information to the decision to be made. This inevitably leads to questions about what a particular language test measures and how useful the results are for informing the intended decision. (Bachman & Purpura, 2008: 456)

This chapter looks first at the qualities seen as necessary to make a test a good measure which can be trusted. (The term test will be used interchangeably with the word examination). We will consider the notions of *validity* and *reliability* within the context of high-stakes tests which are

used for making decisions about people's lives, and discuss how these notions can be related to the use which is made of a test and its scores. This will include the question of fair, equitable and ethical use of tests and test scores and the consequences a test or its scores may bring about.

2.1. Questions of Validity

While earlier validity was perceived as having a number of types: criterion--oriented validity, consisting of predictive and concurrent validity, content validity and construct validity (Cronbach & Meehl 1955: 281), Messick (1989: 20) proposed a 'progressive validity matrix', in which validity is viewed as having different facets, all of which are inter-related. The first line of the matrix concerns the evidence for the validity of a test which can be used to interpret its scores and justify its use. The second line considers the "consequential basis" for validity which involves the "value implications" of the use of a test and its "social consequences." However, we are instructed to read the matrix progressively, that is from top left to bottom right and to see each of the 'squares' as containing not only what is stated in them, but also everything which has preceded them in the matrix. Thus the notion of "social consequences" (the bottom right square) encompasses evidence from construct validity, the relevance and the utility of the test for the purposes in which it is used, in addition to the value implications of the consequences of the test. In this way, the consequences of a test become one aspect of a unified construct validity (Messick 1996: 251).

Construct validity is made up of six aspects: content, substantive, structural, generalizability, external and consequential (Messick, 1996: 247). Each of these aspects will be described briefly in turn, according to Messick (1996: 247–251). *Content* refers to the relationship between the content of the test and what it purports to be testing (content relevance), both in terms of whether the test content is representative of the curriculum or syllabus content and in terms of the quality of the test content. It is expected that a test will provide good cover of all aspects of a curriculum or syllabus and that the test will not contain material which is not contained in the programme (content irrelevance).

Substantive aspects of validity refer to the theoretical rationale which explain the choices behind the tasks and formats chosen for the test. Here are explanations of the skills and processes which completing the test demand, and information as to relationships between different parts of the test and different items. It is expected that the test writer can give evidence that the test-taker actually uses the processes hypothesized in the rationale. It is also expected that, in addition to offering cover of the content of a curriculum or syllabus, the test also covers the processes which that programme claims to develop and that these are covered in a representational

way. Thus, if the curriculum includes development of different ways of, for example, reading, such as reading for general understanding, reading to understand specific detail, understanding relationships between different parts of a text and inferring meaning from context, then all of these processes will be covered in the test content.

Structural aspects concern the way in which the score on the test is composed. It is expected that the type and number of items and the proportion of the score allocated to different tasks and parts of the test reflect the construct the test aims to be assessing. It is expected that the test speci-fications will set out what the test aims to assess, how this is to be done (operationalization) and explain the theoretical rationale for this, in this way making the test construct explicit. The criteria and rubrics used for rating performance should also be consistent with the construct as specified. Thus a model of spoken language would be explained in the specifications, operationalized in terms of the tasks through which it will be assessed and in how it will be rated on the basis of the rating criteria. If the model has more than one aspect of speaking in its construct, then it would be expected that the test include all these aspects of speaking (such as production and interaction) and that the proportion of points allocated for each set of speaking tasks would reflect the model of speaking specified.

Generalizability covers two aspects. First, that what is sampled in the test is sufficient to give a reliable picture of the test-taker's ability in contexts beyond the test. Does the score on a "reading test", for example, give enough information to be able to describe what the learner will be able to read, how they will be able to do this (the processes they are able to use) and how well, or does it simply tell us that the test-taker can do this particular type of test task, which would exclude the possibility for generalization. Secondly, does the test score allow us to make these assumptions for all test-takers in a variety of settings, or is it biased to favour a certain type of test-taker, or a particular context. It is expected that the selection of tasks in the test and the processes they require are sufficient to allow generalization and that this will apply to the whole sample population.

External aspects of validity concern the relationship between the score on a particular test and other measures, such as, for example, classroom assessment. These other measures could also include criteria-related performance, using bench-marks or scales (such as the Common European Framework of Reference for Modern Languages, CEFR, Council of Europe, 2001), or other standardized tests. Any relationships found between these different measures should be consistent with the test construct, which should adequately explain them.

Consequential aspects relate to the legitimate uses to which test scores can be put and how they can be interpreted in the light of the test construct. This aspect carries within it all of the other aspects already described, in that if there are issues with any of the other aspects, then the ways in

which test scores are put to use should take any limitations into account. In this way we see that the consequences of a test alone cannot be used to give evidence about validity, as this is just one aspect of it (Messick, 1996: 242). Messick also makes it clear that any claims that a test has negative consequences must be supported by clear evidence that these are in fact consequences of the introduction and implementation of the test alone and are not the result of other educational practices or policy (ibid.).

The view of validity as a unified concept "has become the accepted paradigm in psychological, educational and language testing" (Fulcher & Davidson, 2007: 14) and has been inherent in the American Educational Research Association (AERA) standards since 1985. On the basis of this description of unified construct validity Messick offers the following definitions:

> Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores and other modes of assessment (Messick, 1989: 13).

In this way we see that what is being proposed for validity is an evaluation argument supported by evidence. He continues:

> Validity is not a property of the test or assessment as such, but rather of the meaning of test scores. Hence what is to be validated is not the test…but rather the inferences derived from test scores…inferences about score meaning or interpretation and about the implications for action that the interpretation entails (ibid.)

Kane (2010) would only partially agree, seeing validation as having two aspects, that of validity concerns during test design and development, which he would see as a quality of a test itself and that of validation of test use, which coincides with that of Messick, as seen here. Bachman and Palmer (1996) take the notion of "construct validity", which they, similarly to Messick, see as referring to the "meaningfulness and appropriateness of the *interpretations* that we make on the basis of test scores" (Bachman & Palmer, 1996: 21), but place construct validity as one aspect within a model of the usefulness of a test, which they regard as being of prime importance. The other elements of the model are reliability, authenticity, interactiveness, impact and practicality (ibid.: 18).

Argument-based approaches to validity

> The person investigating a test concentrates on refuting the counter-hypotheses a critic could make plausible. The job of validation is not to support an interpretation, but to find out what might be wrong with it. A proposition deserves some degree of trust only after it has survived serious challenge. (Cronbach, 1984: 155)

Kane (1990), agreed with Messick's definition of validity as dealing with interpretations made from test scores, but felt that the ways in which these interpretations can be made, the evidence needed to support, or the argument needed to justify the interpretations needed to be more fully explicated (ibid.: 1). Validity depends on the quality of this argument, which should be "sound, reasonable, plausible" (ibid.). Such an argument can be used not only for evaluation of the uses of a test and interpretations made on the basis of its scores, but also to strengthen test design. In this way if the test itself is made more robust the conclusions drawn on the basis of its scores become more accurate.

According to psychometricians (Kane, Feldt & Brennan, 1989: 143; Kane, 1990: 3) validity received much less attention than reliability because it is much more difficult to demonstrate and no clear procedures for carrying out test validation had been outlined. Cronbach (1971: 483) suggested that validation of a test should start with stating the interpretation proposed and that evidence should be collected systematically to give weight to this argument (Kane, 1990: 4). Cronbach later (1989: 162) referred to this as a strong programme of construct validity. However, no clear information was given in either instance as to what kind of evidence should be collected. This problem is also found in the American Educational Research Association, American Psychological Association, National Council on Measurement in Education (AERA, APA, NCME) 1985 standards for testing, and both Messick (1988) and Kane (1990) point out that this failure may lead to use of limited evidence for validation arguments, which might additionally be restricted to data that is most readily available, as no clear indications are given about what data is the most relevant. While Cronbach (1971, 1998) and Messick (1988, 1989) suggest that most relevant is evidence in support of counter-hypotheses, that is argument against the interpretations of the test scores in use, no criteria are offered to judge the weight of these counter-arguments. Kane argues that being able to dismiss counter-arguments is not enough, evidence-based argument *for* the interpretation being made should be offered (1990: 8) and he proposes that it is "inferences and assumptions" that will form the basis of the argument (ibid.: 9). He describes the argument based approach in a seminal paper (1990), which will be discussed here in detail for the reason that more recent developments (Bachman, 2005; Bachman and Palmer, 2010; Kane, 2013) assume an understanding of the philosophy of this approach.

Kane (1990) proposes two types of argument, the *interpretive* argument and the *validity* argument. The interpretive argument sets out the "reasoning from the test scores to statements about some object of measurement and possibly to decisions" (ibid.: 9). It should also make clear any plausible challenges to the proposed interpretation which might be potential weaknesses in the argument. In this way the interpretive argument is clear, descriptive and logical. The validity argument, by contrast, offers evidence

to justify why we should accept the interpretive argument and gives a theoretical rationale to the interpretive reasoning (Kane, 1990: 10; ibid.: 44). The evidence given may range from common sense to empirical data and statistical analyses. The validity argument should reflect the shape of the interpretive argument and provide "a systematic evaluation" of it (ibid.: 10).

The *validity* argument has two stages, the formative stage and the summative stage. In the formative stage clear definition is made of the interpretive argument and the case for the proposed interpretation of scores is presented. In the summative stage this interpretation is 'cross-examined', whereby the challenges to it are expounded and examined. These challenges are empirical. The two stages are not clearly defined and the formative process is carried out with the intention of strengthening the plausibility of the final, summative, judgements (ibid.: 19).

Kane (1990) like Cronbach (1988: 4), and later Wall (1999), draws parallels between test validation and programme evaluation. The formative and summative stages of the validity argument also parallel the conjecture and refutation phases proposed by Popper (1965, 1968) for development of scientific argument. Thus during the formative phase of the validity interpretive argument conjectures are posed which are then subjected to "possible refutation by empirical evidence" (Kane, 1990: 21) during the summative phase.

The formative stage of developing the validity argument may take place during the design of the test itself and so the way the test is prepared aims to be coherent with the interpretations which are to be made on the basis of the test scores. For example, items may be piloted using think aloud protocols to ascertain the processes test takers use to do them, or focal learners may be selected to give immediate post-pilot test feedback on the cognitive processes they employed. Such information could be used as evidence that the processes it was hypothesized would be tested were in fact being used (ibid.: 23).

The argument for how the test scores will be interpreted is laid out in the test specifications. This should include not only the test content but also the processes it is expected test takers will use and the procedures for how the test is to be conducted. Careful analysis of the test specifications can therefore give evidence to support assumptions or inferences (ibid.). At the same time this analysis may also reveal potential challenges. Information obtained from piloting can be used as evidence (see above), while comparison of the taught syllabus and the test content could offer evidence for assumptions relating to a test of achievement.

In the formative phase a preliminary argument is proposed which is then tested by submitting it to "empirical challenge" in the summative phase (ibid.: 25). The focus should be *problematic assumptions*, those which are the weakest in the argument. This weakness may be that they are controversial, that other possible alternative interpretations exist, or that the ev-

idence offered in their support is weak. For example, as decisions about cut scores (i.e. the score determined as the pass mark) are often controversial this is a potential weakness in an argument for a test to be used for placement. Kane (1990: 26) suggests gathering evidence to demonstrate that learners performing below the cut score "tend not to succeed" in the course, while those performing above the cut score "tend to succeed." Such evidence would make the validity argument considerably stronger, despite the fact that the evidence given would refer to an earlier cohort of test takers and not to the cohort currently taking the test.

According to Kane (1990: 28), the interpretation made from test scores is an "artifact" which is created and can be placed on axes of, for example, "focus" and "level of abstraction." Thus a task which asks learners to answer questions based on a reading passage may be interpreted as an indicator of reading comprehension ability, or as one aspect of "verbal aptitude" (ibid.: 29). The more generalized the interpretation, the more complex the interpretive argument needs to be. To justify how such a task can represent reading ability in general we need a well-evidenced argument to support the interpretation. In order to evaluate the strength of the argument all the propositions inherent in it ("inferences and assumptions") need to be stated overtly. Interpretive arguments should be logically structured with the basic tenets stated clearly and all subsequent interpretations made should be coherent with these. Such arguments can therefore be discussed critically and parts of them rejected as illogical, in that some assumptions made do not arise logically out of others. Critique of interpretive argument is consequently critique of its structure in terms of its logic and the critique points out its weaknesses and how it needs to be strengthened.

Interpretive arguments respond to new evidence and so are fluid, that is they are able to expand or contract, allowing wider or narrower interpretations. New evidence may come from increased understanding of the theoretical construct being tested, or from new analysis of data on the use of the test scores. The basic form of the interpretive argument is formulated to apply to the average test-taker. If there is a need to extend the interpretation to particular types of test-takers, such as those with special needs, then new evidence and new argumentation will be needed. Interpretations are not mathematical calculations and as such are not tightly defined. The requirement is that they should be clear and logically argued and that evaluation is made on the basis of the plausibility of the argumentation (Kane, 1990: 29–34).

Bachman (2005: 5), rephrasing earlier questions posed by Spolsky (1981), echoes Kane's (1990) division of the validation argument into interpretive and validity aspects, asking: "How convincing is the argument for using assessment in this way?" and "How credible is the evidence that supports this argument?" His main tenet, however, is that argument based approaches to date had not adequately addressed test use or the consequences of a test, but

had focused rather on how test scores can be interpreted. He proposes an Assessment Use Argument (AUA) which, like Kane's validation argument, is comprised of two parts: "an assessment utilization argument, linking an interpretation to a decision, and an assessment validity argument, which links assessment performance to an interpretation" (2005: 1). This was subsequently extended into a fully expounded version in Bachman & Palmer (2010).

While Bachman's AUA has generated interest it, is Kane's validation argument that has been taken up by testing agencies such as ETS and by researchers (Banerjee & Tsagari, 2016: 2).

Factors of a test which may impact on consequential validity

Qualities of a test which may cause washback are "authenticity" and "directness" of tasks, where authenticity refers to tasks which as closely as possible mirror real world language use and the processes this involves (Messick, 1996: 242–243). Direct test tasks should assess only what is described in the test construct and nothing additional which might in any way affect this. Thus, for example, tasks which purport to be testing listening and involve the learner in writing a response are questionable, as the listening construct is being polluted by the need to transpose phonemes into graphemes. It is possible that the learner understands the information (listening), but is not able to render it into writing and so what is being scored is not a good measure of the listening skill. Messick cautions, however, that authentic and direct tasks are difficult to achieve in a testing context; firstly because of affective factors at play in performance testing (which is what real-life use of speaking implies, for example), and secondly because the ways in which such tasks are rated rarely reflects the processes involved in the real world (ibid.: 245). Engaging in a conversation with someone in another language our concerns are more for mutual comprehension and intelligibility than for rating our interlocutor's grammatical accuracy or range of lexis, for example, aspects commonly found in rating criteria for spoken interaction.

Two key areas of concern for Messick (1996: 244), which he feels are responsible for the negative consequences of tests, are those of *construct under-representation* and *construct irrelevant variance*. Where the test assesses only some aspects of the construct and omits others this is seen as construct under-representation, while construct irrelevance is where aspects are assessed that are not part of the construct. If a test lacks authenticity, this may contribute to construct under-representation, while if it is lacking in directness there may be construct irrelevance. Bachman (2005: 16), however, disagrees with Messick, arguing that negative consequences of a test can occur regardless of its validity. Bachman claims that test scores can be used inappropriately even if the scores themselves are valid measures.

Issues with performance and validity

Uttering a "sentence" is an action in itself, not a description of the action contained in the "sentence", nor an affirmation of doing the action. Austin names this a performative sentence or utterance, or a performative (1962: 6–7). Judith Butler (1990, 1993) takes performatives as a metaphor for "the way in which our actions may construct a sense of our inner being or sense of self" (McNamara, 2001: 338). She sees gender identity as constructed through social interaction, in that one's sense of who one is, is partly constructed by experiencing it in society. We may not, however, be aware of this process, believing that persona to be our individuality. Butler holds that this sense of individuality is "performed" in that we try to understand our inner selves through "performativity", actions which are conditioned by society and our experience. Thus "performance" is not a realization of the inner self but an act of creation, the creation of a fiction, and as the process is not conscious we remain unaware of the origins of our gender identity in our social context.

McNamara (2001: 339) uses Butler's ideas to provide a new insight into performance testing, questioning whether language ability and language proficiency really exist as constructs and suggesting that rather they may in fact be constructed through the act of language assessment itself. If, following Butler's argument, language proficiency only emerges through test performance, then it is not a construct which can be pre-conceived by a test designer, thus undermining notions of some aspects of validity. Followed through to assessment this poses new challenges:

> Generally speaking, we administer tests to, assign scores to, and make decisions about individuals for purposes such as selection, placement, assignment of grades/marks, and the like. If we view language as co-constructed, how can we disentangle an individual's contribution to a communicative exchange in order to provide a score or assess a candidate's merit for a potential position?
> Chalhoub-Deville and Deville (2005: 826)

The relationship between competence (the learner's potential language ability), and performance (their observed language behavior), is controversial, with questions raised as to whether performance on a test can be taken as manifestation of competence, particularly in the case of speaking, if the test includes an interlocutor and a rater. It can be argued that speech produced in this context is socially constructed and as such may not represent the learner's individual competence, but rather a mediated, and so enhanced, performance. Even if the human element is removed from the test by use of computerized elicitation, affective factors may influence the learner's performance, thus rendering it an inadequate representation of competence. Then there is the question of the tasks included in the test, the combination of which represent the test designer's view of what constitutes spoken competence. Choice of tasks will inevitably be limited

by practical constraints, such as cost or time, but what is chosen must provide an adequate sampling of learner speech from which their spoken competence can be inferred to contexts beyond those of the test. Making decisions as to which tasks allow us to make such generalizations is fraught with difficulty (Messick, 1996: 251). While some test tasks, such as elicited imitation, are seen as highly reliable and correlate well with other sub-tests of linguistic ability, (Yan, Maeda, Lv & Ginther, 2015) such tasks could be seen as having low face validity (how does the ability to repeat something someone says relate to the learner's ability to speak spontaneously?), or as having potential negative washback in the communicative classroom if they become mechanical practice activities. The final selection of test tasks needs therefore to combine a variety of tasks of varying formats and focus in an attempt to capture the elusive nature of spoken competence (Messick, 1996: 249).

To sum up, the prevailing view of validity in language assessment today is Messick's unified concept, within which he includes the consequences of a test. He argues that validity is the meaning of test scores, rather than a quality of the test. Kane, by contrast, argues that there are aspects of the test design and development which should be valid, in addition to the uses made of the test scores. Validation, both authors agree, should be approached as a logical argument to justify the interpretations made from test scores. Bachman & Palmer tend to a more pragmatic view, that the test designer should be able to show those concerned "that the intended uses of their assessments are justified" (2010: 2), focusing on the *uses* as opposed to the *interpretations of scores.*

We next looked at the factors of a test which may affect consequential validity, considering in particular key issues described by Messick as *construct under-representation* and *construct irrelevant variance*. These were examined in relation to the assessment of speaking, which can be seen to create challenges. While on one hand there are arguments that tasks should be direct and authentic, there are many counter-arguments suggesting that direct, authentic tests of speaking do not necessarily tap the construct they attempt to test. In other words, the case of how speaking can be tested adequately is controversial and open to question.

2.2. The power of tests

The title of this section is taken from the book of the same name by Elana Shohamy (2001), one of the key researchers who has helped to open the eyes of those involved in testing to the social consequences of tests and the potential for their misuse, which may not at first be apparent. Foucault (1995: 184) sees tests as having a:

> normalizing gaze, a surveillance that makes it possible to qualify, to classify and to punish. It establishes over individuals a visibility through which one differentiates them and judges them.

Echoing this, Shohamy argues that tests are powerful tools which are viewed as having authority, and as a result can be used to persuade people to behave in various ways. Spolsky (2012: 495) warns that "both tests and guns are potentially so powerful as to be commonly misused." In this section we look at the deeper issues behind the ways tests are used.

Language tests can be used to implement government or state language policy in covert ways (Shohamy, 2003, 2006a), for example by imposing a requirement to take an examination in a certain language in order to gain entrance to a further education institution. The fact that one language has been chosen rather than any other gives the chosen language status and importance, which in turn suggests that studying it in school is desirable and so it should be given priority within the education system. Language tests are just one of a series of mechanisms which are used to create language policy (Shohamy, 2007).

Shohamy (2007) demonstrates that language tests can result in "determining prestige, status and hierarchy of language, suppressing diversity and standardizing and perpetuating language correctness and homogeneity" (ibid.: 122). She cites the example of Hebrew, which is the language of entry tests for universities in Israel, despite the fact that ethnically Arabic learners are educated in Arabic, thus implicitly downgrading the importance of that language. The case of English being used as the language to measure achievement for learners in the US under the No Child Left Behind (NCLB) programme disregarded the fact that there are immersion education programmes in heritage languages (e.g. Navajo in Arizona), or that immigrants require time to develop the necessary English language skills to enable them to demonstrate knowledge of content subjects, thus sending implicit messages about the inferiority of languages other than English (Byrnes, 2005). International examinations, such as TOEIC, establish norms for acceptability in language which may run counter to what native speakers might say or write (Young, 2012: 187). These norms are based on a standardized version of English from Britain and North America and suppress other varieties.

Shohamy warns that tests may promote the belief that "languages are uniform, standard and follow the same written norms" (2007: 124) through use of standardized items of a correct-incorrect format, or through use of a marking key which only permits certain language uses as correct, thus over-simplifying the complexity of authentic communication (Milroy & Milroy, 1999: 142). Scales such as the Common European Framework of Reference (CEFR) have a strong impact on both language policy and testing (Morrow, 2004; Fulcher, 2004), giving the impression that language develops in a predictable, hierarchical progression for all learners, as shown in the

descriptors. Despite the fact that the underpinning of the CEFR scales is still largely theoretical, rather than soundly empirical, CEFR scales are widely used both as the basis for curricula and for rating proficiency in tests. McNamara (2011) argues that CEFR is now strongly influential in educational and assessment policy in Europe and beyond, although the process needed to calibrate tests to the scales is extremely complex and determining equivalence of levels far from decisive. Where those tests are destined to be used to make important decisions about people's lives, such as in employment, university entrance, or obtaining funding, statement of the level of the test on the CEFR scales is of crucial importance.

McNamara also draws attention to the potential hegemony of CEFR if adopted as a test construct, because it is culture-bound, coming from a learner-centred, communicative language teaching setting, and that the norms and values inherent in its descriptors of speaking, for example, are strongly Western-European. Adopting CEFR as the basis for Europe-wide policy carries with it implicit acceptance of the superiority of these practices, regardless of whether they are in sympathy with the prevailing culture. Language assessment based on CEFR is being used to administer European language policy, even though the policy inherent in it may not be compatible with the contexts in which it is subsequently applied. Fulcher (2009) sees CEFR being used as a tool to enforce the will of a collective (here the EU) on member states, where the drive for standardization embodied in the Bologna declaration imposes one set of values on Europe as a whole. The fact that tests are now expected to be linked to CEFR "is to demonstrate compliance with a mandate" (Kaftandjieva, 2007: 35), and could be interpreted as an attempt to control both test constructs and content.

2.2.1. Use of tests for selection

Messick (1989: 85) discusses the consequences that tests may have in society when they are implemented for the purposes of selection:

> The central question is whether the proposed testing should serve as the means to the intended end, in light of other ends it might inadvertently serve and in consideration of the place of the intended end in the pluralistic framework of social choices

The potential pitfalls of selection are to do with social justice, as, for example, the act of choosing some people from a population on the basic of a test automatically excludes others, who may, as a result, feel they have been treated unjustly.

> In selection systems, we are thus faced with multiple sources of potential injustice – injustice in values, of rules, or implementation, and of decision making procedures – any combination of which may be salient in a particular selection setting.
> Messick (1989: 86)

The aspect of social justice is most acute when language tests are used as criteria for decision-making in immigration, a function known as 'gate-keeping' (Shohamy, 2007: 126). Tests may be used, for example, to verify identity, in the attempt to exclude those travelling on false documents, although this is problematic and highly controversial, as the following example shows. In a 2012 case involving the Irish airline Aer Lingus, a female Irish resident, travelling on a Greek biometric passport, was asked to complete language tests in Greek and English before being allowed to check in for a flight to Dublin from Barcelona (Coulter, 2012). When she protested on the grounds that holding a passport does not legally entail being able to speak the language of that country, an airline official told her "many people from your country travel on false documents", ignoring the fact that biometric passports are reported to be secure. After being refused a copy of the test to take away, she photographed it, which led to airline staff threatening her. On the insistence of the woman and her husband (an Irish national), the police were called and the incident reported. The woman lodged an official complaint on the grounds of equality, which led to widespread media coverage. This revealed that the test being used was also highly dubious, containing items such as "Please show me any cash you have in your possession." In addition there was discrepancy between the Greek and English versions, with the Greek version asking "Where are you going to in England?" (for a passenger checking in on a flight to Ireland), while the English asked "What is your travel destination?", indicating that the document had not even been checked for accuracy. Aer Lingus later reversed their policy as a result of this incident.

This use of testing was inappropriate for several reasons. First, the construct of the test is unclear and we do not know what it intended to test. Second, it contained ethically unsound items; third, its administration was unfair, as it took place publically, with no warning, against the will of the test taker in a situation of stress and anxiety. Nor was it not clear how the airline official was intended to check the passenger's responses, or verify their appropriateness. Nor do we know what were the criteria for a pass. Even if the test had been well-constructed, the conditions under which it was administered would have rendered it invalid, as under such circumstances it would not elicit a fair representation of the test taker's language competence. Or perhaps the intention was to design a test that it would be extremely difficult to 'pass' and so restrict travel for certain groups? Unfortunately, this is not an alien notion, as we will see in the next examples.

Hawthorne (1997), discussing the gatekeeping use of tests, cites the example of the *access:test* designed to test skilled workers seeking to migrate to Australia. In 1996, when the government needed to reduce migration, the number of points needed to pass was increased. In short, the language test was used for selection purposes in a way that is questionable on the grounds of social justice.

The level of proficiency required, the type and content of the tests which receive official approval and test administration procedures are all the subject of much debate. In 2010 the UK introduced a language requirement for those from outside the European Economic Area, (EEA) who are partners of British nationals, or people settled with the 'indefinite leave to remain', seeking to obtain or extend visas. Initially this was to be documented by any test at level A1 (CEFR). However, the uncovering of abuse in the administration of one widely used international examination in 2014, where "professional" test takers replaced the bona fide candidates and took the tests in their place for a fee (a practice known as "ghosting") led to a tightening of regulations. Now the requirement is for certification obtained only from an examination provider on a government approved list, in addition taken at an approved test centre, in response to allegations of malpractice in some test centres outside the UK. Spolsky (2012: 498) refers to a similar 'ghosting' incident which took place in the US in 2010 and alleges that this can be found wherever high-stakes tests are mandated for immigration purposes. The requirement for A1 language certification now extends to "parents of British citizens and persons settled in the UK" (introduced in 2012) for those from outside the EEA. From 2013 adult migrants seeking residency in the UK are required to provide certification of speaking and listening skills at level B1 and to take a new test, "Life in the UK." From 2016 those from outside the EEA entering the UK to join members of family already resident there will be required to obtain certification of speaking and listening skills at level A2 after two and a half years in Britain, in order to be allowed to remain and complete the five years in the country they require to qualify to apply for the right to settle.

> The new A2 requirement delivers the government's manifesto commitment to ensure that those coming to the UK on a family visa with only basic English will become more fluent over time. It will mean that the person can better engage in everyday conversation and thereby better participate and integrate in everyday life in the community.
> (Home Office website, 2016)

David Cameron's explanation (in January 2016 when he was UK Prime Minister) that the need to improve the level of English particularly applied to the large number of Muslim women in the UK who had low language skills and were isolated in society because of the "backward attitudes" of their husbands, provoked a media storm (Mason & Sherwood, 2016a). Further comments, alleged to Cameron, reported in the Telegraph newspaper, about young Muslim men being ripe for radicalization because the "traditional submissiveness of Muslim women" means they do not respond to prevent it, led to a protest action "#traditionally submissive" by educated female British Muslims on social media, where women posted photos of themselves holding up cards on which they listed (in English) their achievements.

More than 30,000 women responded within three hours. Author Shelina Janmohamed, instigator of the action explained:

> That's just one stereotype about Muslim women, it's not how we all are. We are vibrant, diverse, we're talented and we have opinions. The prime minister is always saying we need to take up British values, so I responded in the most British way I could – with sarcasm.
> (BBC Trending 26 January 2016)

Cameron tempered his message, explaining that having limited language skills might lead to feelings of alienation, which might in turn lead to a young person being more susceptible to extremism (Mason & Sherwood, 2016b). The act of legislating a requirement for language certification appears to be seen as a panacea for healing divisiveness in British society and indirectly as a means to combat the alienation of some young British Muslims. Baroness Sayeeda Warsi, Muslim member of the House of Lords, endorsed the idea of encouraging language skills, but disputed any possible links with extremism.

Legislation on language certification has been made with the specific intent of coercing members of British society to improve their skills. While some may argue that this is justified, as the apparent aims of deepening integration and empowering women are noble, we also need to consider the darker issues. If those in the UK on family visas do not manage to pass an A2 level certificate within two and a half years, they face deportation. This affects spouses or parents, with all of the associated trauma of family breakup. Funding of ESOL courses in the UK has been cut, reducing the number of places available. The A2 tests approved are general 7 minute oral tests conducted in a one-to-one interview with an examiner. No provision appears to be made for the fact that for women of some cultures being alone with a member of the opposite sex who is not a member of her immediate family is not usual. In addition, in the sample film available on the test provider's website, the test begins with the examiner (male) shaking hands with the candidate (female), an act which is not usual in some cultures. Put simply, passing this approved test entails far more than developing language skills. The performance conditions under which the test is conducted are an embodiment of British culture, an example of Foucault's 'normalizing' power in action.

McNamara & Roever give several instances where tests are used as "weapons within situations of inter-group competition and conflict" (2006: 196), such as during World War 2 on the Bataan peninsula where American soldiers were under attack by the Japanese. Understanding that the enemy had pronunciation difficulties with the sound /l/ the Americans deliberately chose passwords which contained multiple examples of this, such as 'lollapalooza'. In the dark, anyone challenged who began the password with 'rorra…' was shot at once. This pronunciation test is a modern version of

a test to distinguish friend from foe based on the same principles, described in the Old Testament, where members of the enemy tribe of Ephraim were identified by the defending followers of Gilead by their inability to pronounce the initial sound in the Hebrew word 'shibboleth' which was used as a password. Testing, in both these examples, had dire consequences.

To conclude this section, any test which is used for gate-keeping purposes, even within a country as a threshold test for passing from one stage of the school system to the next, should carefully consider if certain groups are not being "favoured" over others, and whether the consequences of the test are not resulting in social injustice.

2.2.2. Test results become labels

Another fear about the widespread use of standardized testing is that the test scores become a label which are used to define the test-taker. Bourdieu (1991) sees the test as an instrument of power, which guards the borders of a group to which membership is desired, for the benefits it will bring. In order to gain access the petitioner has to become 'instituted', that is accepted by the existing members, which involves displaying similar traits and behaviour. This carries the sense of there being a norm to which members of this group adhere and as a result those whose traits or behaviour do not conform to the norm are excluded. Intelligence tests (IQ) serve as an example of this thinking. Binet, who developed the first intelligence test (1905), did so with the aim of identifying children who were 'subnormal' in order to allocate them to special schools. In the UK, intelligence tests, alongside tests of English and maths (known as the '11+' test), together with teacher reports, were used to decide which young people (aged 11–12) could attend grammar schools, and so follow an academic track bound for institutions of higher education, and which students were deemed non-academic and so destined for either technical vocational education, or general ('secondary modern') schools and entry to the job market at 15. This resulted from the 1944 Education Act, which introduced mandatory free secondary education for all and raised the school leaving age from 14 to 15. Devised for the purposes of selecting young people who had potential for success and in whom it was worth investing scarce resources, the examination quickly became competitive, as the number of places at the prestigious grammar schools was limited to approximately one third of the cohort. Only a minority (around 5%), deigned 'manually dexterous', were selected for technical schools, while the remainder went to secondary moderns. There is evidence that the exam favoured those from middle class backgrounds, related not only to socio-economic effects but also to the cultural capital of the home (Bourdieu & Passeron, 1976), and parental aspirations (Hart, Moro & Roberts, 2012).

The stigma of 'failing' the 11 plus proved highly divisive in society, with the outcomes used to 'label' young people, even amongst themselves. This

was emphasized by the fact that grammar school children wore uniforms, while this was not usually the case for other types of school. Consequently, social inequality deepened, because adolescents attending grammar schools tended not to mix with others, as the focus of their worlds was so different. Once on the non-academic track it was only possible to gain admission to further education by studying for academic examinations (then GCSE 'O' and 'A' levels) part-time, at 'night school', a course which was followed only by those who were exceptionally determined, as the majority of those attending secondary moderns left school at 15 with no qualifications. The tripartite secondary system predominated until 1964 when comprehensive secondary schools started to become more common. The 11+ examination was abolished as mandatory in England and Wales in 1976, although it remains until today in some local authorities which still have grammar schools. The current prime minister, Theresa May, has promised to consider the re-introduction of selective secondary education later in 2016, which is creating strong debate in the media, particularly with regard to the aspect of measuring IQ. Selection of children for special education is no longer done purely on the basis of tests of IQ. Intelligence has been found to change as a result of environmental factors, through contact with others and through programmes such as Feuerstein's Instrumental Enrichment. IQ tests have been demonstrated to be biased towards the culture of the test compiler and so to unfairly assess members of ethnic minorities (Gipps, 1999: 361), calling their validity into question.

Broadfoot (1979: 44) argued that tests are used to give an illusion of equity in selection processes. The fact that everyone can compete, apparently on the same terms, leads those who do not succeed to take responsibility for the failure upon themselves, believing it results from their own lack of ability. However, as we have seen in the case of IQ tests and the 11+ exam, there are other factors at play which bias the outcomes and mean that the stigma which results from failure is unjust. A letter sent by the school principal of a UK primary school to accompany the results of a standardized test (KS2), seems to voice similar disquiet:

> we are concerned that these tests do not always assess all of what it is that make each of you special and unique. The people who create these tests and score them do not know each of you – the way your teachers do, the way I hope to, and certainly not the way your families do

The letter goes on to lists in detail all the things the test results do not tell about a pupil's abilities, and ends

> … the scores you get will tell you something, but they will not tell you everything. So enjoy your results and be very proud of these but remember there are many ways of being smart.

The fact that this letter went viral on the internet (Huffington Post 15.07.2014) appears to indicate that the sentiments it expresses touched a chord, indicating an awareness of the potential negative social consequences of tests used as measures of control.

2.2.3. Tests as a force for social good

Testing can, however, also be viewed as a force for social good. Eckstein & Noah (1993) remind us that one of the oldest known tests, for entry to government service in ancient China, was introduced in order to break the prevailing culture of nepotism and patronage and to open opportunities for all on the basis of merit, thus promoting social justice. In more modern times, entrance tests for employment in positions in government offices were introduced in the US in the 1870s. In similar ways tests can be used to reduce corruption by providing a transparent means for deciding who will benefit where there is a scarcity of resources to be divided, such as allocation of places at university, or even secondary school, in societies where further education is still the privilege of a few, rather than a universal right. Testing may also serve to limit the advantages of wealth, privilege, or membership of an elite and ensure greater equity when a selection process is implemented. The introduction of entrance examinations at the universities of Oxford and Cambridge in the mid nineteenth century, soon followed by Durham and London, helped to open the doors of institutions which had been previously accessible only to those of 'appropriate' background (Gipps, 1999: 357).

If we acknowledge the power of tests, and there seems to be adequate evidence that we should do so, then the question remains as to how this power can be used equitably and in ways which ensure social justice. This leads us to the question of ethics in language assessment, which we consider in the next section.

2.3. Ethical issues in the use of test scores

In this section we will consider issues concerning ethics in testing and fairness.

2.3.1. Ethics and responsibility

Language testing is constantly changing and developing and one of the ways this can be seen is in the debate about the social consequences of tests, which considers to what extent the consequences of a test are the responsibility of those who design and produce them. Hamp-Lyons (2000) makes a distinction between the *critical language testing* movement, which she interprets as being concerned with monitoring how test scores are

interpreted by stakeholders and the uses which are made of scores, which she sees as mainly the province of parents and learners, although also of concern to teachers and those responsible for tests (ibid.: 579), and *ethical language testing* (McNamara, 1999), which she interprets as the insider view of those who produce tests.

Shohamy (2001b: 376) explains that "Critical language testing refers to the need to question the uses of tests as tools of power and to examine their uses in education and society." This is done by critically examining the ways tests are used, what consequences they may have, both intentional and unintentional, close analysis of test content to see what is included and what is not covered and the reasons given for this, making transparent the decisions which are made on the basis of a test with the aim of informing test takers and test users, and monitoring the impact of the test with a view to minimizing any negative effect (Shohamy, 2001b: 376–377; Shohamy, 2001a: 131). Within this framework consideration is also given to the scores, whether these compare test taker performance with that of other test takers (normative), or whether they report test taker performance against a checklist or set of standards (criterion-referenced), what the test score is taken to mean in the context and whether this interpretation is "open to discussion and interpretation" (Shohamy, 1998: 333).

By contrast, "Ethical language testing…simply asks all language testers [i.e. those producing tests] to set themselves high standards and take every step to ensure that they are upheld" (Hamp-Lyons, 2000: 586). This stance has come about through a growing understanding of the need for the test provider to be accountable, aware that tests are used in many situations, beyond their control, which have social impact.

Ethical issues also extend to the approach taken to a test in the classroom by teachers and learners. Messick (1996: 241–2) argues that for the effects of a test to be judged as beneficial "there should be little if any difference between activities involved in learning the language and activities involved in preparing for the test." Mehrens and Kaminsky (1989) and Popham (1991) (in Hamp-Lyons 1998: 334) both offer checklists for judging the appropriateness of test preparation materials in this respect, on scales ranging from unacceptable to acceptable. *Unacceptable* preparation materials are seen as those designed with the intention of increasing the learner's score through practice designed to improve their test taking ability, with no aim of developing the skill which the test is assessing. Other materials may be seen as *ethically questionable*, as they could be interpreted as having some overall benefit in terms of language or skill development, but appear to be mainly aimed at score improvement. Materials deemed to be *unacceptable* or *dubious* contribute to 'score pollution' (Haladyna et al., 1991), as they inflate test results without actually developing the learner's competence. Through repeated practice learners improve their performance on specific test tasks. In this way the reliability of the test itself is compromised, as

it ceases to be a good indicator of what the learner is able to do with the language. This in turn limits the effectiveness of the uses of test scores, as they are no longer reliable measures from which learner ability can be generalised. Such inflated scores may lead a test taker, for example, to achieve a place on a course requiring a level of language which they do not in fact have.

Consideration of responsible test preparation led Alderson & Hamp-Lyons (1996: 295) to ask a series of questions which have contributed to a wider view of language testing and opened a new debate. They considered whether the test could be held responsible for how teachers prepare their learners for it, or whether this responsibility lies with the teacher, the materials writer, publisher, or the learners who require their instructor to adopt a particular approach. Acknowledging that there is an ethically acceptable way to teach learners on courses which will culminate in high-stakes tests, they wonder where the responsibility lies for training teachers in this kind of approach: with the test provider, materials writer, the publisher, or the institution which provides the teaching qualification.

While Hamp-Lyons (1997b: 302; 2000: 587) affirms that the language tester should take complete responsibility for the consequences of a test, Davies (1997: 33) argues that although the ethical testing perspective is morally correct, the tester cannot be held responsible, because the consequences of test use cannot be fully predicted. However, he agrees that the language testing profession does have a role to play in setting out guidelines for fair test use and codes of ethics, or good practice. This has come about with creation of documents such as the International Language Testing Association (ILTA) Code of Ethics (2000) and the European Association for Language Testing and Assessment (EALTA) Guidelines for Good Practice in Language Testing and Assessment (2006).

Consideration of what is ethical, is, however, not straightforward, as the contexts in which a test is implemented are not uniform. "What is considered ethically acceptable varies from country to country, culture to culture." (Hamp-Lyons, 2000: 589). This may lead the tester to a moral dilemma, if they find that the requirements imposed by a particular context run counter to their personal convictions. What is important is that each language tester should have their personal moral code with regards to ethical testing, on the basis of which they make decisions and that they should be prepared to decline to work for bodies with whose values, or practices, they cannot agree.

The consequences of a test also vary depending on the perspective we adapt when considering them. As different stakeholders have differing areas of concern (see Chapter 1), the ethical testing approach urges us to examine the consequences of a test critically for all parties involved.

Fulcher & Davidson (2007: xix) bring some of these issues together in creating what they call *effect-driven testing*, which they describe as "a new approach that we believe brings together testing practice, theory, ethics

and philosophy." Acknowledging that tests bring about consequences in society they urge

> it should be these test effects that drive the final design decisions about crafting particular items and tasks. We believe in "effect-driven testing": test creation in which the ultimate test design decisions are driven by the impacts that the test will have on stakeholders.
> (Fulcher and Davidson 2007: 50–51)

As it is not possible for a test to assess all aspects of the test-taker's language, it is necessary that the test delivers a sample which provides enough information to enable the user of the test score to be able to generalize about how this test-taker will perform in other contexts. Fulcher & Davidson argue, therefore, that the starting point for designing a test is to consider "the performance conditions under which the test-takers are capable of operating" (ibid.) and use these to build the test specifications. The test designers should additionally consider models of communication to inform the process. The language use situation thus becomes the main driver in the test design process, rather than taking a theoretical model of communicative competence as the starting point and then considering how this can be operationalized. In this way the designer begins by considering the test-taker and how they need to use the language in the real world. This determines the use to which test scores can be put, as the test specifications will clearly state for whom the test is intended and in what situations it should be used. Using it in other situations would be considered unethical and invalid, as this would involve generalizing beyond the domains or purposes for which it was designed.

In short, the question of ethics affects both those designing and preparing language tests, but also teachers who are preparing their learners to take tests. The test designer needs to consider the possible consequences the test may have on the test-taker, and to clearly define ways in which they intend the scores of the test should be used. The teacher needs to be aware that inappropriate use of test preparation materials may distort the impression given by their learners' performance on the test, a fact which leads not only to problems with test reliability, but also to negative outcomes for the learner as they find themselves, for example, wrongly placed in an ability group. It would seem that all parties need to consider the test in the context in which it is intended to be used, to have a clear picture of what they expect the test-taker to be able to do with the language in the real world, and to ensure that the test gathers sufficient information to be able to gauge whether in fact the learner is able to evidence this.

2.3.2. Fairness

According to the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 1988, hereafter referred to as the *Code*)

fairness relates to ensuring that tests "are as fair as possible for test-takers of different races, gender, ethnic backgrounds, or handicapping conditions." For those producing the test this means monitoring the content of test tasks for topics, language or illustrations which could be discriminatory; conducting statistical analysis of the performance of sub-groups to check for bias (such as differential item functioning, DIF); working to eliminate any factors in the test not related to the test-takers' skills which might be causing difference in scores; and producing modified forms of the test for groups with specific learning needs, such as, for example, those with visual impairment (ibid.: 2–3). Fairness is thus seen as an important quality of a good test, and guidelines for how it should be ensured are offered.

Kunnan (2000: 3), in the same spirit, added the question of access to the test to these areas. This covers issues such as the affordability of the test fee, the location of test centres, physical access to and facilities available in test centres for those with disabilities, whether test-takers have been able to familiarize themselves with the test conditions and with equipment to be used in the case of multi-media testing. In addition he included the category of educational opportunity:

> opportunity to learn plays a major role in test-takers' success on tests when test-takers have had the opportunity to learn the material on which they are assessed. Further if test-taker groups have differential opportunities to learn, then group performance on a test will most certainly differ significantly. (Kunnan, 2000: 4)

Beyond the test itself, and the conditions in which it is conducted, lie the social consequences of the way the test and its results are used. Kunnan (2000: 4–5) included the justice of test use in his definition of fairness, with concern that no group should be treated differently from any other on the basis of their test results. Where the test has a gatekeeping or certification function it should be clear that all test-takers have equal opportunity and that the test is not being used as a mechanism to favour some groups over others.

Willingham & Cole (1997), in contrast to the conceptualization inherent in the 1988 *Code,* considered fairness an aspect of validity, on the grounds that if a test is unfair this reduces its validity. They describe fairness as comparable validity, where the test should be equally valid for all test-takers (Willingham & Cole, 1997: 6–7). Willingham (1999: 11) thus considers that a test is fair when there is

> Comparability of opportunity for examinees to demonstrate relevant proficiency, comparable assessment exercises (tasks) and scores, and comparable treatment of examinees in test interpretation and use.

Kunnan (2004) drew up a Test Fairness Framework as a model for a fair testing system. This developed his earlier ideas and drew on the *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 1999,

hereafter referred to as *Standards*) which set out twelve standards for test fairness. Underlying the framework is a rationale based on ethics which is composed of two main principles, of Justice and of Beneficence. These can be summarized as follows:

Principle of Justice

All should be treated equally and the test should "be fair to all test-takers" (Kunnan, 2004: 33) This entails the principles that scores on the tests can be interpreted similarly on comparable constructs for all test-takers and that no group is subject to bias "by assessing construct-irrelevant matters." (ibid.)

Principle of Beneficence

The test should have positive effects in society. This entails the principle that information obtained on the basis of the test and its social consequence have positive effects. It also assumes that this information is accurate and that it can therefore be put reliably to use (ibid.).

The model includes five areas:

- Validity
- Absence of bias
- Access
- Administration
- Social consequences

Kunnan (2004: 39) suggests that the fairness of a test can be evaluated using an argument-based approach similar to Kane's (1992) validity argument.

In a 2008 paper Kunnan emphasizes the centrality of validity in the model of fairness, representing fairness as a circle which contains a flower-like system of interlocking circles with validity in the centre, overlying other 'petals' which contain the other aspects listed above (Kunnan, 2008: 236). By setting fairness as paramount he differs from the Willingham & Cole (1997) view and that of the 1999 *Standards*, where fairness was seen as an aspect of validity, and also from the 1988 *Code* which saw fairness as an aspect of test quality.

The test producer ETS published their own *Standards for Fairness and Quality* in 2002 and these appear to follow the guidelines for fair testing practice approach found in the 1998 *Code*, although the ETS standards declare that if difference is found in the functioning of scores in specific groups taking a test, then the scores should be studied in order to exclude the possibility of construct-irrelevant or construct under-representation, which are aspects of Messick's (1989) unified concept of validity (Xi, 2010: 150).

Kunnan's model of fairness is questioned by Xi (2010) on the grounds that fairness may be addressed within a validity argument (Kane, 1992; Kane et al., 1999; Kane, 2001, 2002, 2004, 2006, 2014; Chappelle et al., 2008), or within an assessment-use argument (Bachman, 2005, 2010) which are

better equipped to evaluate aspects such as the validity and social consequences of the test. The 1999 *Standards* suggest that each of the aspects of validity it describes (test content, score structure, rating etc.) should be examined for how they affect specific sub-groups taking the test in question, thus indicating a structure both for evaluating fairness and for planning research. While Kunnan's work has helped to bring fairness issues into focus, the model does not make it clear how fairness should be investigated, or how a fairness argument should be constructed. Xi concludes, however, that the approach and framework suggested by Willingham & Cole (1997) and Willingham (1999) "provide the most lucid conceptualization of fairness" (2010: 153) in its view of fairness as comparable validity. She builds on this notion, extending their definition of fairness

> as comparable validity for *identifiable* and *relevant* groups across all stages of assessment, from assessment conceptualization to the use of assessment results. This conceptualization of fairness implies that a test has to be fair to be valid. Anything that weakens fairness compromises the validity of test score interpretation and use.
> (Xi: 2010: 154)

In the same paper she proceeds to demonstrate how a fairness argument for the TOEFL iBT test might be constructed.

Kane (2010), responding to Xi, argues that the relationship between fairness and validity is not clear cut, expressing reluctance to place fairness within validity and preferring to see the concepts as separate, although intertwined. He justifies this by breaking fairness down into different concepts, based on the concepts of due process in American law, those of *procedural* and *substantive* fairness. Procedural fairness, put simply, relates to ensuring that all test-takers are treated equally or comparably, or that accommodations are made where necessary. Thus it is viewed as a requisite for a test to be considered valid. By contrast, substantive fairness is a more difficult concept, based on the decisions and interpretations made on the basis of test scores. These should be "reasonable and appropriate" (2010: 178) and "equally appropriate for all test-takers (at least roughly)" (2010: 179). While procedural fairness can be to a great extent controlled by the test producer and administrator and can be planned for in advance, substantive fairness concerns the way a test is used in a system, and how results function for different groups, and so can be evaluated only once the test has taken place. Unlike procedural issues, Kane holds that the use and interpretation of scores is beyond the test producer's remit. Clearly, however, if analysis of the performance of specific groups reveals the effect of construct-irrelevant factors, then it is a question of validity and the responsibility of the test producer to remedy. Decisions and interpretations are, by contrast, a separate case. Problematic is equality of "opportunity to learn" and here Kane separates tests used to measure achievement at

the end of a programme (where differing opportunities are a question of validity), from proficiency tests designed for certification, or the testing of specific skills in a specific domain for a qualification, where he argues opportunity does not necessarily relate to either fairness or validity. He cites the example given in Shepard (1993) of a test designed to determine whether children were ready to enter kindergarten. Found to function well as a predictor of the ability of children to cope in kindergarten, and deemed sound in terms of procedures, the test could be considered valid and fair. However, if a child scores low the result may have been caused by lack of opportunity in the home, for reasons of socio-economic deprivation, or dysfunction, and so the decision not to allow the child to attend pre-school would most likely be detrimental to the child. Thus, substantive fairness is much more difficult to evaluate than procedural, but, as illustrated, of key importance.

Kane (2010: 180–182) upholds Xi's use of a fairness argument as part of validation, but draws attention to the need to include it in the interpretative argument, with clear descriptions of which interpretations of scores, for which groups, the test is designed to allow.

In this section we have examined questions of ethics relating to language assessment. We considered issues of responsibility resulting from testing and examined arguments for the extent to which it is the test designer, or provider who can be held to account, versus the arguments that the uses made of a test cease to be their responsibility. We noted that codes of practice and guidelines for test use have been created in response to this widened interpretation of assessment. We then turned to consider the question of fairness in testing, with aspects of equity and inclusion raised. Finally we saw how the concept of fairness relates to validity.

2.4. Uses of tests and test scores

Bachman & Purpura (2008: 458) offer a useful analysis for classifying uses made of test results: "Language assessments are used in the service of a variety of decisions, including student selection, certification, classification, tracking, promotion or retention in educational programs, and allocating resources to schools." This allocation of resources comes about through comparison of the performance of schools, and measuring them against targets. At a system-wide level, other than for the allocation of resources, testing is also used for monitoring whether educational policy aims are being met and curriculum standards being achieved.

Fulcher (2009: 9) takes a different perspective, seeing the ways tests are used as reflecting the political philosophy of a society:

> Modern collectives use testing as a "scientific and technical tool" to control the educational system with the intention of (1) creating or reinforcing the identity

of the state, (2) increasing the sense of belonging to the state, (3) selecting and allocating individuals to roles or tasks that benefit the collective, and (4) introducing hyperaccountability to ensure uniformity and standardization.

Hamp-Lyons (2007), however, places assessment on a continuum of what she describes as two cultures: the culture of learning and exam culture. While, within the exam culture, focus is placed on examinations which describe norms in the population and provide a picture of language proficiency at a particular moment in time, within the culture of learning, focus is placed on the progress an individual learner makes over time as they follow a course. In the exam culture, test scores are used to show patterns across a large scale system, while in the culture of learning, the scores used are those from assessment created by a teacher in their classroom, which is often criterion-referenced and is designed to show what an individual is able to do, how well they can do this, and what they need to improve (Black & Wiliam, 1998). These scores will be used by the teacher to give feedback to the learner to inform them what they need to work on and how to do this.

Teasdale & Leung (2000) similarly point to two types of assessment: for learning and for measurement, where assessment for measurement corresponds to what is described in Hamp-Lyons' exam culture. Black & Wiliam (1998) also distinguish assessment of learning from assessment *for* learning (AfL), where AfL is part of the teaching and learning cycle in the classroom and is designed to promote effective learning.

The existence of these very different types of assessment leads to a tension between what a teacher feels they should be doing. On one hand, there is the need for classroom based assessment with its focus on individual growth and the generation of useful information, while at the same time there are the requirements of external, national tests, where scores are used for quite different purposes. It is most likely the fact of being caught between these two systems, with their very different demands, that leads some teachers to have an ambivalent attitude towards large scale testing. Where the large-scale test results start to infringe on classroom teaching, through demands for accountability, these attitudes may become negative, as teachers begin to feel that the importance of learning has been relegated below that of test outcomes. Voices of dissent from teachers and teacher associations may be one force calling for educational reform.

Another important group of stakeholders in national testing programmes are parents. Here again there is potential tension between what a parent wants from a test score and what a national test can realistically deliver. Parents what to know how their children are doing. They want to know whether their children are able to do certain things and how well they are able to do them. They want to know what help their children need and how this can be given. Yet, at the same time, they want to know how their children are performing relative to others, in their group, or class, or year.

In addition, parents appreciate the assurance of knowing how the school their child attends is performing in comparison with others, so that they may be sure that what appears to be a satisfactory result, is in fact satisfactory with relation to the larger population. They then feel reassured that the assessment of their children is reliable and is not inflated or deflated by the effects of the group, or the teacher. In other words, they expect far more than national test scores can adequately supply, with the demand for information on individual attainment of curriculum criteria. While they welcome the normative information on mean scores and reporting that enables them to compare their children's outcome with others, they also require a more detailed and personalized profile on the basis of the test scores. Once again there is a tension, as it is unlikely that national test scores will provide this kind of information, because it is not their primary aim. Frustration caused by the apparent dearth of information supplied by large-scale tests may lead to parents questioning whether national testing is really necessary and to voices calling for reform.

In response to concern from parents and teachers in 2007 in the UK, a House of Commons Select Committee was set up to investigate if there was really a need for national testing. After a careful enquiry process a report and evidence of testimony by witnesses were issued. One of these witnesses was the Department for Education and Science (DfES), who claimed that national standardized testing which aimed to measure attainment of curriculum targets had had many positive effects. These included an increase in teacher expectations and pupil aspirations. They found that parents are now more fully informed about their children's progress and about performance of schools. The testing system was felt to have placed an emphasis on achieving standards of educational attainment, which in turn has caused teachers to work for improved learner outcomes, a process which they consider is reported with transparency. While the committee found some acknowledgement that testing could have adverse effects, such as narrowing of the curriculum, rather than ensuring a broad education, it was generally felt that having a programme of national tests was better than what had preceded it, when both the content and quality of what was taught varied considerably from school to school. The report sums up

> We consider that the weight of evidence in favour of the need for a system of national testing is persuasive and we are content that the principle of national testing is sound. Appropriate testing can help ensure that teachers focus on achievement and often that has meant excellent teaching, which is very welcome.
> (House of Commons, 2008: 27)

The same commission, however, expressed concern about the importance placed on scores by Ofsted (the Office for Standards in Education) when evaluating schools, summing up "national test data are evidence of only

a very limited amount of the important and wide-ranging work that schools do" (2008: 39). They also found that Key Stage tests, which although not actually high-stakes, could be considered as such, on the grounds that they have considerable consequence for teachers and schools. The pressure this places on educators may be communicated to pupils and contribute to the children also perceiving the tests as high-stakes. The strong emphasis on accountability was reported to be having negative effects on teaching. Witnesses from the tertiary sector reported that students arrived unprepared for university, having been focused on rote learning at school, and so lacked the critical skills necessary for independent work (ibid.: 49), supporting claims that test-focused shallow learning had become the main goal in schools.

The House of Commons Commission called for the testing of children to be removed from school accountability, as there was considerable evidence to suggest including test results as part of the accountability system was having a negative impact on education (ibid.: 66).

Accountability forms the topic of the next section.

2.4.1. Tests and Accountability

In accountability a 'principal' sees an 'agent' as being responsible for some kind of performance, for which the 'agent' is required to provide a description (an 'account') to the 'principal' (Jacob & Kirst, 1999). When accountability is part of a system it is used to answer four key questions: "*Who* is accountable? *To whom* are they accountable? *For what* are they accountable? And *with what consequences*?" (O'Day, 2004: 21). A school is held accountable by an outside body, or a stakeholder, to implement educational policy and legislation, through appropriate processes and procedures, and also to deliver results in the form of outcomes of learning from their students, with various possible consequences. The theory behind such accountability systems is that supplying an institution with accurate, reliable information, validly interpreted, about what it has done will motivate it to work for continued improvement and to develop areas found wanting (Baker & Linn, 2004: 47–48).

Accountability may work in different ways within systems. Harris & Herrington (2006) distinguish government-based and market-based accountability. Where schools are viewed as public institutions which are run at the cost of the tax payer, there is a concern that the school provides a good service and functions as mandated in government legislation. It is therefore the responsibility of the state, or local educational authority, to assure quality. Schools can be held to account by offering incentives, such as additional funding for good performance, or by disciplinary action if they fail to come up to standard. This may be for example, withholding of funds, changes in school leadership, or increased monitoring. By contrast, where parents have the freedom to decide to which school they will send

their children, information about schools functions in a similar way to information about products for sale, and forces similar to those in market economics come into play. This process depends on information, on the basis of which parents can make their choices, being readily available. Schools, like goods, need to be competitive in order to attract 'customers' (parents) through offering attractions such as innovative programs, rich resources, excellent staff, small groups, or a record of good examination results. In these circumstances the school needs to be accountable to the parents, who may behave like dissatisfied clients if the school is not to their liking, by taking their business (i.e. their children) elsewhere. As the funding of a state school very often depends on the number of young people enrolled in it, the school authorities will do everything in their power to attract and keep pupils, thus pandering to the 'market'.

Accountability can thus be seen as a control mechanism, designed to effect targeted change in schools. Central authorities decide on the changes they wish to see, design and set curriculum standards to operationalize these, and create an examination system which tests the extent to which these goals have been met. The mechanisms used to bring this about are tests (where results are either expected to reach certain levels, or where results are compared with others), reporting systems, inspectorates, incentives and sanctions. The thinking is that differences in test results are the effect of the schooling test-takers have received. Hence, the school can be held accountable for poorer outcomes. Such thinking is, however, flawed, as it fails to take into account contextual factors which have been shown empirically to bias test results. Correlations have been found between socio-economic home backgrounds and results; between place of residence and results; between home language and results and between ethnicity and results. In short, it is not the contribution of the school alone which is responsible for test outcomes.

Accountability can be to an external 'principal', as illustrated above, but the 'principal' may also be within the institution itself, so, for example, a head of a subject specialization can hold teachers in their team to account for the outcomes of learners in their subject. The focus in the research described in the second part of this book is on external accountability, but we will see that within-school accountability also comes into play in the Polish context. Mechanisms in true internal accountability are different from those where accountability is to an outside body, as they assume self-regulation based, for example, on codes of professional conduct and performance benchmarks.

Testing and test results may play an important role in school accountability. We can place this on a continuum from systems where "the process of evaluating school performance [is] on the basis of student performance measures" (Figlio & Loeb, 2011: 384), as in many places in the US under No Child Left Behind (NCLB), (2002–2011, discussed below), to systems

where external examination results are one of several indicators used for quality assurance, such as in Poland. Tests used for accountability are usually summative assessments, and come at the end of an educational stage. Tests in this process are usually closely related to a state or national curriculum, which describes standards, targets, or behaviours, which should be attained by students at different stages of their education.

> The objective of standards-based reform is to identify a set of clear, measurable, and ambitious performance standards for students across a number of core subject areas, to align curriculum to these standards, and to expect students to meet these high standards. A central component of standards-based reform is the assessment of students to ensure that they are meeting the expectations set out for them, to identify the schools that have students who are relatively successful (or unsuccessful) in meeting these expectations, and to encourage schools to improve student outcomes.
> (Figlio & Loeb 2011: 386)

Such tests are standardized, in the sense that all learners receive the same test, or validated parallel versions of it, and that the test is administered in such a way as to ensure that the performance conditions are the same for all test-takers. The marking, moderating or rating procedures also follow carefully regulated procedures, using standardized marking schemes and rating criteria. As a result, standardized tests allow comparison to be made between the scores of different test-takers and decisions to be made on the basis of these.

Linn (2000), reviewing the role of tests in educational reforms in the US during the previous 50 years, gives four reasons why assessment is included in attempts to improve education and for accountability. These are that as a system solution for raising quality testing is a comparatively cheap quality assurance measure (as opposed to increasing staff-student ratios, increasing contact time, changing programs which involve new materials and teacher education etc.). Secondly, it is easier to impose a test on an education system than a new way of teaching. Thirdly, tests can be introduced quickly and, finally, results can be seen by all and can be expected to show improvements in the short term (e.g., Linn, Graue, & Sanders, 1990). Linn argues (ibid.), from a political perspective, demonstrating that tests can be used to show improvement in education within one term of office of a government, and suggests this explains their widespread use.

Accountability mechanisms can use test results as a 'status' measure, where the test scores are used to calculate the percentage of students attaining a specified target level (Figlio & Loeb, 2011: 392). This approach encourages schools to concentrate on getting the largest number of students to achieve the target, which tends to focus attention and effort on lower achievers, with the aim of getting their performance up to the desired level.

Alternatively, an accountability programme can use measures of 'growth', where testing is used to measure levels at a starting point and then repeated at a later point, with the difference in scores calculated so as to obtain a measure of 'gain'. The length of the period of schooling between the tests varies, but in England, for example, is one school year. This approach takes into consideration the fact that students enter a school, or programme, at different levels of attainment, and also recognizes the influence contextual factors, such as home background, have on learning outcomes. A growth measure encourages the school to design teaching so that each individual, regardless of their ability or circumstances, shows improvement, which is often perceived as fairer than the status approach. 'Status' measures are popular as they are easy to implement. One standardized test is administered for the whole student population. This was the approach taken, for example, by NCLB until 2014. 'Growth' measures are more complex, as they require a series of tests which are empirically linked on one scale. The algorithms used to calculate gain for individual students usually allow for factors such as poverty, language learning background (e.g. English is not the home language), ethnicity, and special needs. Growth for a year group (cohort) is measured by aggregating measures of individual growth. The result, referred to as *value-added*, purports to show the school's contribution to learning outcomes. The different approaches to measuring learner achievement may produce different results for schools, with a school found under-achieving on a status measure having positive growth effects or vice versa (Clotfelter & Ladd, 1996; Ladd & Walsh, 2002; Kane & Staiger, 2002; and Stiefel et al., 2005).

National examination results can be reported in various ways to enable valid comparison of the performance of one school with others. Commonly used forms are standard nine scales, and percentile rank scales. The standard nine scale shows the distribution of a school's results compared to the national mean score, expressed in terms of the number of standard deviations they lie away from the mean, and place a school in one of the different levels. Performance definitely cannot be expressed in raw scores, which show only the number of points attained and does not allow comparison with other tests, or with populations. To compare a school with others information is needed about how that school performed relative to the others, which depends on the level of the learners in the school and the school's performance in comparison with the whole population of schools in the sample. When comparing schools' outcomes another possible way is to convert mean scores to percentile ranks. These indicate "the percentage of test-takers in a reference group whose scores are equal to or below a particular score" (Bachman, 2004: 96). So if a school's mean score is placed at the 60th percentile, this means that 60% of all schools in the set scored the same as this school or less. Consequently the higher the percentile rank (i.e. the larger the number) the better the school's standing in comparison with

others. Low rankings indicate that there are few schools who scored at this level or lower. While such information may be useful for educational policy makers, it provides only minimal information about the school and nothing about individual achievement. As a result questions can be raised about this kind of information being used for accountability.

Wiliam (2010) points out some of the difficulties of using high-stakes test results for accountability. Using the illustration of the 2003 maths scores in the US on the international PISA tests (OECD, 2004), he draws attention to the fact that the variance in scores between schools (approximately 25%) is much less than the variance within a school (approximately 75%). Of this intra-school variance, 69% could be attributed to socio-economic factors, leaving a mere 8% as the possible effect of the school itself. Similar results were noted for science results (OECD, 2007). Like Hattie (2015), Wiliam suggests that the focus for increasing effectiveness of schooling should not be between-school measures, but rather in-school measures, as the aim is for improved learning for all. He concludes:

> Because differences between schools account for only a small proportion of the variance in student scores (in most countries less than 10%), standardized tests are rather inappropriate tools with which to hold districts, schools and teachers accountable.
> (2010: 42)

Wiliam (ibid.) hypothesizes that as different learners develop at different rates they achieve targets at different times. On vertically linked tests, as in the English system, some learners will reach a set target in year 3, while others will not reach the same target until year 7. Taking a measure of growth over a year for a cohort does not adequately capture individual growth, with annual growth in achievement for a cohort averaging between 0.25 and 0.4 of a standard deviation on test scores (Rodriguez, 2004). Kane & Staiger (2002) support this view, suggesting that review periods of greater than one year are needed to reduce measurement error.

Despite having misgivings, Wiliam (2010: 34) acknowledges that there is some evidence that accountability systems do impact on student achievement and raise scores. Hanushek & Raymond (2005) found that strict accountability (where severe sanctions applied if targets were not met) raised scores more than less rigorous systems, but that the raising of scores was not uniform and the attainment gap between Caucasian and African-American students widened. Carnoy & Loeb (2002) by contrast, found that accountability raised scores for all students. Bishop (2001a,b) found that external tests based on curriculum had greater positive impact on student test scores than other types of test. We will explore findings on negative aspects of the use of high-stakes testing for accountability in the sections on the US and England below.

The question therefore remains as to how testing can be used *reasonably* within an accountability system, as it appears it may have a potentially beneficial role. In acknowledgement of the limitations of restricting accountability of schools solely to the outcomes of learning as measured on standardized tests, many educational systems use a variety of indicators, one of which may be test results. The 2015 Eurydice report on school evaluation in Europe, which reviewed all 28 EU member states, plus Iceland, Norway, Turkey and Macedonia, reports that "in most cases external evaluation focuses on a broad range of school activities, encompassing educational and management tasks, student outcomes, as well as compliance with regulations" (European Commission /EACEA /Eurydice, 2015: 8). However, it also found that test scores have increased in importance for evaluation of the quality of schools over the past decade, with approximately 75% of the systems under review now using them as an indicator, as opposed to 25% in 2004. Reporting systems which allow the possibility to compare school results with those of similar schools, or with national averages, are now the second most common indicator for internal evaluation (ibid.: 11).

The 2013 OECD report *Synergies for learning* also shows that there is now a wider focus for accountability, which it attributes to individual schools having greater autonomy, and so the use of internal evaluation is encouraged. There is a recognition of the key role of the school principal in leading this process (Fullan, 1991). Like the Eurydice report, *Synergies* found widespread use of standardized testing linked to curricular standards, to assess learner achievement for accountability purposes. A new trend noted, however, is the growing awareness of the need for formative assessment, based on evidence of learner performance, and the emergence of the belief that *both* formative and summative assessment data should be used to give information about the teaching and learning process in school and in the education system as a whole. Separate schemes for the evaluation of teachers are also beginning to appear, rather than basing teacher appraisal solely on learner outcomes.

In the reporting of learning outcomes of individual schools it is now more common for contextual data to be taken into account and value-added measures are increasingly used. More developed technology now allows for more sophisticated reporting of results of standardized tests, including individual profiles.

2.4.2. Testing and Accountability in the US

In the US, concerns about the ability of the country to maintain its strong global position led to the introduction of minimum competency testing in schools, with the aim of raising the educational level of citizens and so increasing human capital potential (Madaus, 1985: 614). Although at first improvements were observed, these soon levelled off, leading to a belief that

such a testing programme was encouraging mediocrity and was lowering the levels in schools by not providing adequate challenge. In 1983 the report *A Nation at Risk* recommended a more rigorous approach in education, including the introduction of high-stakes tests based on standards, and for schools to be held accountable for their students' learning outcomes (Amrein & Berliner, 2002: 4). This marked a change, as prior to this students themselves were viewed as accountable for their test results (Chaloub-Deville & Deville, 2006: 511). The philosophy of teacher and school accountability continued with the passing of the act to implement the No Child Left Behind (NCLB) programme in 2002. Under this, each state was responsible for instigating annual testing of reading and mathematics in grades 3 to 8 primary and at least once in secondary school and for setting targets to be achieved each year. These targets specified the percentage of children in a given school who had to reach the desired level for reading and maths in that year. Annual targets were also specified for sub-groups classified for race, socio-economic status and special needs. The results for English Language Learners (ELLs) and special needs students were, however, included in the school's aggregated results. The goal was for all children to achieve the desired targets by 2014, that is 100% in all schools. The programme functioned by offering incentives or sanctions depending on whether the targets were attained. These measures varied from state to state, but included pay related bonuses to teachers and school principals for attaining targets, or achieving results beyond the set target, and a range of sanctions from publication of the results of under-achieving schools ('naming and shaming'), the requirement to offer free tutoring to learners in poorly performing schools, allowing parents to transfer their children to other schools, forced changes in school leadership, increased external supervision, or, in the case of five consecutive 'failing' years, closure of the school.

As the NCLB programme continued, it became increasingly unpopular, with anecdotal evidence and later a large body of research, outlining negative effects. The tests and findings attracted considerable attention in the media. Resistance to the tests also began to be noted, in the form of popular movements such as "Just say no to the test", promoted by the organization Fairtest.

Research on the impact of NCLB

This sections offers a brief overview of the extensive research associated with NCLB. It is in no way intended to be exhaustive, for reasons of space.

Blake (2012: 8) critically reviews the growth of the importance of high-stakes testing within NCLB and summarizes:

> While the initial intent of NCLB was to set educational standards, improve the educational learning opportunities for all students and thereby raise achievement scores of students, the current outcomes of NCLB appear to have cast

> a normalizing gaze, homogenizing and mandating a standardized and highly test based educational setting focused on achievement and conformity.

The evidence that NCLB initiatives have in fact increased teacher effectiveness and increased student achievement is scant. Nichols & Berliner (2008) stated that there was no important statistically significant data to support the claim. Kim & Sunderman (2005), Lee (2006), Nichols (2007), and Nichols, Glass & Berliner (2006) report that NCLB has either had no significant positive effect, or that student achievement has fallen. Koretz (2008) and Nichols & Berliner (2007) describe the negative effects NCLB has had on schools, teachers and school districts. These include practices such as "gaming the system", where schools selectively enter students for the tests, holding back less able learners to repeat a year, or even removing them to special schools, in order to raise the school average score. Teachers have been found to focus attention on so-called "bubble kids", those learners whose scores are just failing, in order to increase their scores and raise class averages, to the detriment of more able learners, who are ignored (Booher-Jennings, 2005; Madaus & Russell, 2010). Such practices result in artificially inflated scores, an effect first noted in the 1980s when all states in the US reported above average scores. Known as the Lake Wobegone Effect, after Keillor's book describing a town in which "all the children are above average" (Cannell, 1985, 1987), the effect means that test scores cannot be trusted, as they have been inflated by various forms of malpractice. Darling-Hammond (2007), Hursch (2007), and Lapayese (2007) claim that rather than improving education NCLB has in fact increased the attainment gap between different groups of learners.

Au (2007) carried out a meta-synthesis of 49 qualitative research studies conducted on relationships between curriculum and high-stakes testing in the US. He re-analysed data from these studies for evidence of alignment of the course content to the test, for narrowing of the course content (e.g. reducing time spent on non-tested subjects), for increase in time spent on or expansion of content of tested subjects, for evidence that knowledge was being structured differently (either in a more fragmented way, or more holistically with focus on higher-order thinking skills) and for evidence that teachers were changing the way they taught in response to the test, with particular attention to whether there was an increase in teacher-fronted classes, or in student-led interactive tasks. He found that there was significant evidence of test impact in three areas: content of the curriculum, structuring of knowledge, and teaching approaches. Curricula were found to align to the test. Knowledge was generally found to be structured in smaller, more fragmented pieces, oriented towards the test and taught in test-based contexts. However, there were a few instances where knowledge was found to be more integrated. There was a notable trend towards increased teacher-fronted transmission of facts. The qualitative data allow a more detailed picture to emerge and appear to show that the situation is

actually rather more complex than previous research indicates. There are indications that response to curriculum content change may depend on the subjects taught. For example, teachers of social studies in secondary school *added* reading training related to the test to their programmes in order to expand time spent on literacy skills, an area strictly speaking belonging to English Language Arts, thus indicating that hypothesizing a contraction/expansion of course content within a subject maybe an over-simplification. Changes in teaching approaches also appear to be sensitive to the format of the test task. In a few cases, where the test required learners to write essays based on authentic source texts (e.g. history in New York), social studies teachers were found to respond by increasing teaching of critical thinking skills and the amount of work in pairs. The quality of the test task, and the way it is rated, appear to influence how teachers adapt their teaching. Social studies was noted as an area which had the most disparate findings, with two authors claiming the test had no effects on teaching, others claiming curriculum alignment and others, as we have seen, reporting an increase in student-led activities and the introduction of work on higher-order skills. Au concludes that high-stakes tests have a significant impact on content, form and teaching in US school classrooms. What he does not do is to claim that this has increased either teacher effectiveness or student achievement.

Plank & Condliffe (2013), in a longitudinal study using classroom observations in addition to achievement data, found that in those classes where the teacher and school were required to show the greatest improvements, the quality of work in the classroom was compromised, further indicating negative effects of the use of high-stakes scores as a measure in accountability.

The Obama administration produced *A Blueprint for Reform* (2010), which aimed to encourage states to adopt the Common Core Standards (CCS)[1] (2010). These have been widely adopted across the country. Success in receiving federal grants, such as *Race to the Top*, are dependent on applicants creating a consortium of at least 15 states and on each of these having adopted no less than 85% of the CCS (Chaloub-Deville & Deville, 2011), thus levering states to introduce the CCS. The *Blueprint* (which continues the use of high-stakes testing to measure accountability) claims that research finds show that the accountability programme improves school results and raises learner outcomes and national standards, a fact which is questioned by Ravitch & Mathis (2010). Mathis (2010: i) states:

> US states with high academic standards fare no better (or worse) than those identified as having low academic standards. Research support for standards-driven, test-based accountability systems is similarly weak. And nations with centralized standards generally tend to perform no better (or worse) on international tests than those without.

1 Under US law the government cannot mandate national standards or a national curriculum.

To conclude this brief overview of research on NCLB and the follow-up programme, it would appear that little has changed since these remarks made in 1996:

> Despite the long history of assessment-based accountability, hard evidence about its effects is surprising sparse, and the little evidence that is available is not encouraging. …The large positive effects assumed by advocates...are often not substantiated by hard evidence…
> (Koretz, 1996: 172)

However, we should remember that the tests per se are not the cause of the negative effects reported, but rather the *stakes* which are associated with them. Accountability based solely on results of high-stakes testing can be considered detrimental (Madaus & Russell, 2010).

2.4.3. Testing and accountability in England[2]

Testing in England and Wales began as a national initiative alongside the introduction of a national curriculum in 1987. Testing for children aged 7, 11 and 14 was announced, as an addition to the existing GCSE examinations at 16 (the end of compulsory education), as a measure to improve levels of education, give information about learner progress and allow those having difficulty to be singled out for further support. The resulting Key Stage Tests were felt to be an objective and standardized measure of performance and a means to ensure comparability of education across the country.

From the outset, test results were made publically available, both to give information to parents and to act as an incentive to teachers to improve performance (House of Commons, 2008: 9). They are published in the form of performance tables, recording the achievement and attainment of schools. Initially based on aggregated data from each Key Stage Test, average improvement for a school was calculated and a target was set for the school to achieve in terms of the percentage of pupils expected to reach a certain level. This has since been revised, with the 'target' now known as a 'floor standard' and a measure representing progress introduced in addition (Hutchings, 2015). Attainment for a school is reported using both value added (a measure of growth between two tests which adjusts for the starting level) and contextual value added (CVA), which additionally includes adjustment for socio-economic factors. The media then transform these performance tables into rankings, popularly known as 'league tables'.

In 2007 the House of Commons convened the Children, Schools and Families Committee on Testing and Assessment to consider whether there is a need for national testing at the current level. After hearing a large number

2 In the UK, England and Wales have an education system separate from Scotland or Northern Ireland. The accountability programme described here, however, covers only England and not Wales.

of witnesses and reviewing written evidence submitted, the Committee concluded that there was a case for national testing, but found that the way test results were being used was causing "tensions in the system leading to undesirable consequences" (House of Commons, 2008: 20). They also found that the test results

> do not necessarily provide an accurate or complete picture of the performance of schools and teachers, yet are relied on by the Government, the QCA [the Qualifications and Curriculum Authority] and Ofsted [the Office for Standards in Education, Children's Services and Skills] to make important decisions affecting the education system in general and individual schools, teachers and pupils in particular.
> (ibid.)

The National Association of Head Teachers (NAHT) similarly acknowledged the validity of having a national testing programme, but questioned the uses made of the results in particular in performance tables and in evaluation of schools (ibid.: 41).

Published in 2006 the government document *Making Good Progress* claimed that the introduction of the new system of testing had had a positive impact on standards of performance (ibid.: 2). This assertion is, however, questioned by the National Association of Schoolmasters Union of Women Teachers (NASUWT), (House of Commons, 2008: 27), the National Union of Teachers (NUT), who question whether the higher scores in fact reflect improvements in learning, and the Association of School and College Leaders (ASCL), who suggest that the improved performance results from increased resources being available in education following increased investment in the sector. Witnesses called to the House of Commons Committee reported performance targets were causing a focus on borderline pupils, with the aim of getting them through the set Key Stage Test. This is reported to lead to "shallow" learning focused only on the test, with the result that the learner may find themselves progressing to a level which is in fact beyond their scope at that time and which cannot be maintained (Smithers, 2007). There were also charges that teaching to the test and similar practices was leading to score inflation, which may appear to be supported by the fact that apparently higher rates of success on national tests (GCSE and A levels) are not reflected in improved scores on international tests, such as PISA.

While initially, in the 1980s, targets were set for the average learner, these subsequently became minimum attainment targets to be achieved by all, which has led to increased pressure on the less academic and on pupils with Special Educational Needs (SEN). The expectation that all pupils should achieve the same target at the same time has been strongly criticized (e.g. by NAHT, and the Association of Teachers and Lecturers) as not reflecting child development and as having negative consequences for

pupils. If the target is too high the child may become stressed, frustrated and demotivated, while if it is too low the pupil may become bored and disenchanted. High early school leaving rates are seen as a direct result of this policy (House of Commons, 2008: 32). Also criticized is the tone of the *Making Good Progress* document, which views those who do not achieve targets on time as 'failures'.

Performance tables are the subject of considerable controversy. While the government defends their production on the grounds of providing a transparent reporting system, it is widely acknowledged that the information they contain is so limited that it provides only "a distorted snapshot of the work of a vibrant and organic community" (NAHT, House of Commons, 2008: 34). This fact has a negative impact on the work both of teachers and the school as a whole (p. 40). Small schools fare disproportionately badly on performance tables, as the percentage of the aggregated school score contributed by a single pupil is much higher than in larger schools. Thus one low score pulls the school average down far more than is the case in a larger institution. A 2005 MORI (Market & Opinion Research International) poll showed that parents paid little attention to the performance tables when making decisions about what school to choose on the grounds that the information in them was unclear. CVA in the tables was particularly criticized as requiring expert knowledge to understand. The reason one school has a lower CVA score than another may be attributed to a number of factors, such as the number of SEN pupils, in addition to the number of pupils from lower income homes, and unless this information is also made available it is unlikely an uninformed person will be able to make sense of it. The House of Commons Committee recommended that performance tables should be read "in conjunction with the relevant Ofsted report[3] in order to get a more rounded view of a school's performance" (2008: 38).

To sum up, the use of test results for judging the quality of education in a school and for holding the school to account is controversial. In the US claims for the beneficial effects of school accountability appear to be, as yet, inadequately supported by empirical evidence, while there is some, albeit debatable, evidence that accountability promotes questionable practices and may lead to negative consequences. In England anecdotal evidence from a variety of witnesses representing reputable educational bodies claims that accountability measures, and performance tables in particular, lead to an over-focus on exam results, which are felt to inadequately represent what schools contribute to a child's education. In turn, exam results are also felt to provide only limited information about a child's achievement.

3 Ofsted inspect individual schools and report on the quality and effectiveness of teaching and learning there.

2.5. The role of external examinations in the Polish educational system

In Poland national external examinations were introduced for the first time in 2002 (this will be explored in detail in Chapter 3). The extensive educational reform of 1999 introduced a national core curriculum and this was then supplemented by Standards of Attainment [pol. *standardy wymagań egzaminacyjnych*], setting out what would be assessed. Each school was required to produce a coherent program [pol. *program szkoły*] composed of subject programs and a plan for pastoral care and support. The school program was required to ensure complete cover of all aspects of the core curriculum and take into consideration the relevant standards of attainment. The external examinations, prepared by the Central Examination Board, were based on the core curriculum and the standards of attainment. Legislation on the education system requires schools to cover the core curriculum and so, as the examinations are based on the core, results of these examinations are seen as a means of measuring the effectiveness of the work of a school. This is expressed clearly in the following quotation (2008: 1) by the then director of the Central Examination Board, the late Mirosław Sawicki, in the introduction to a new quarterly publication designed for educators:

> We are aware that to a great extent the quality of education and the success of pupils depends on the school principals and teachers guiding the learning process of their students. Ever greater knowledge about examinations and their relationship to teaching will guarantee that you will be better able to plan the teaching process. This will ensure increasingly better cooperation between the examination system and schools.
>
> We all know that examination results are not the aim of schooling, education has much wider significance and more far-reaching aims, but it is examination results that are the most objective information about the effectiveness of your work. They can provide valuable information for you, without the understanding of which it would be difficult for you to plan your teaching in school. This information takes on meaning as regards content thanks to working together with the examination system. Raising the standard of this cooperation is the intention of the publishers of this quarterly. Only through the close working together of all those engaged in the education of the younger generation can we ensure better schooling for them. (own translation)

From the outset it was perceived that examination results would have a role in evaluation and quality control in schools (Ministry of Education and Sport Regulation on pedagogical supervision, 2004). With this in mind, legislation from 2004 onwards has required schools to analyze the results of external examinations, with the aim of improving the quality of the education provided. It must, however, be stressed that analysis of examination results is only one of a series of requirements defined for this purpose. In 2009 the Regulation

was revised and formed part of a new programme for evaluation of schools, which aimed to use the legislation to lever and motivate the development and improvement of certain practices in education (Mazurkiewicz, 2011: 311) with beneficial effects anticipated. The document is intended primarily for the school to draw up a plan, which aims to help learners develop their full potential. This plan is consulted with all stakeholders and agrees targets for the school to work towards with the engagement of all parties (Kowalczyk-Rumak, 2012). The descriptors in the document are also used as criteria for external evaluation.

The 2009 document explains that schools may be awarded grades from A (the highest) to E in seventeen different categories, of which the one relating to examinations is given first. Where an E grade is awarded, the school is required to prepare a corrective programme, implement, monitor and report on it to the local education authority within an agreed time frame. The Regulation sets out descriptors for levels D (the minimum considered acceptable) and B. The descriptor for level D reads: "results of [external] examinations are analyzed with the aim of improving quality of the work of the school... Findings from the analysis are acted upon in school" (Ministry of National Education (MEN) 2009: 1.1). This was the legislation in effect at the time of the research which will be described in Part Two.

Since the time when the research was conducted, the legislation has been revised and modified to read: "The school…in planning the educational process takes into consideration conclusions from the analysis of the results of the …*gimnazjum* examination… and other external and internal studies." (Journal of Laws of 2013, item 560: 12) The descriptors for levels D and B have also been changed. At the lower level the requirement is that the school

> carries out analysis which leads to conclusions and recommendations on the basis of which teachers plan and take action. Action taken by the school…is monitoring and analysis, and in case of need, modification. In the school… results of the examinations are analyzed and [also] the results of external and internal assessment
> (ibid.)

Level B now reads

> In the school…the results of external studies [pol. *badań*] are used and internal studies are carried out, as appropriate to the needs of the school…, which include studies of learner achievement and tracer studies of school leavers.
> (ibid.)

Interestingly, this requirement, now one of 12, rather than 17 as previously, now features as number 11, while in the 2009 document it was listed first. More significantly, if an E grade is awarded for this point, on account of the results of external assessment, the requirement to prepare an official

corrective programme no longer applies, although the school is still expected to "undertake activity to improve the situation so that the requirement is met" (paragraph 10.4).

Analysis of external exam results is most often conducted by teams of teachers in school specially convened for this purpose, or by a team of subject teachers. Analysis includes comparison of the school's mean score with the national average on the standard nine (stanine) scale, then looking at the spread of student results measured against the stanine scale and noting the number of those with lowest results. (Stanine scales were replaced by percentile scales in 2012). Examination reports produced by the Central or Regional Exam Boards are used to mark which parts of the test, or which items, representing different skills, or sub-skills in the core curriculum, were found most difficult nationally, and then comparing the school results against this. Research suggests that more qualitative use of test results is challenging for schools (Ligęza, 2013; Ligęza & Franczak, n.d; Milecka 2014a), with teachers experiencing difficulty with understanding item facility values and in identifying the sub-skills being tested by an item. After being analyzed, results are presented at a meeting of the whole school staff [pol. *rada pedagogiczna*] and discussed, (according to Milecka (2014a), this referred to 68% of schools studied).

That exam results are analyzed in Polish schools is beyond question, but how they are in fact used is not clear, with external evaluator reports suggesting that this "use" of the analysis may in fact be superficial (Stożek, 2010). Teachers seem to be concerned more with how to improve exam results in the following year than in how to improve the quality of their teaching. Actions taken could be characterized as "teaching to the test", or "narrowing the curriculum", by increasing the time spent on "problem areas", rehearsing tasks found to have been done less well, or organizing additional lessons dedicated to exam preparation (Ligęza, 2013; Ligęza & Franczak, n.d.; Milecka 2014a, 2014b).

In a study conducted by the Educational Research Institute, Warsaw in 2012 school principals from a sample of 150 *gimnazja* were asked how they used the exam results. 97% said that the analysis formed part of the plan for supervision of their staff; 92% claimed that exam results formed the basis of their evaluation of the work of the school; 86% used them to select in-service training for teachers and 82% made use of the analysis in preparing or modifying the school development plan (Matuszczak & Wasilewska, 2015: 232).

Reporting on research conducted following the first implementations of external examinations between 2002 and 2005, Lisiecka (2005) described how schools at first used results to compare themselves with other schools, by carrying out simple quantitative analysis. This subsequently developed to more qualitative analysis, where the meaning of the results was extracted and interpreted. These interpretations were then communicated to the different stakeholders, who made use of the results in different ways. Lisiecka (ibid.)

categorizes how schools use results as depending on the management style, where "authoritarian" schools compare their results with others and search for the reasons for low results, while "learning community" schools favour qualitative approaches and aim to improve the effectiveness of their work, based on self-evaluation and discussion. However, it has been found that teachers have difficulty with interpreting exam results and seem to lack the education, or experience to apply them in their work (Lisiecka, 2005; Milecka, 2014a).

According to Milecka (2014a), results are used for both diagnostic and corrective purposes. Schools also use results for confirmation of the quality of their work through external validation. According to the international Teaching and Learning International Survey (TALIS) (2013) 91% of teachers surveyed in Polish lower secondary schools consider results of examinations obtained by their learners as the most important source of feedback on their work. This was second only to observation visits (Sitek, 2015). This would appear to confirm Milecka's finding. Exam results are also used by schools to check whether action taken has brought about positive effects, or are used to compare with in-school assessments (Milecka, 2014a). There were a small number of reports of results being used to evaluate the work of teachers. Schools also reported using results to motivate or inform. This includes using information to individualize work with learners, especially those with special needs, or to recruit parents to work in partnership on specific targets. Results are also used in some cases to analyze the cover and implementation of the core curriculum. Despite the fact that results are not comparable year on year (see chapter three), 41% of schools in the study checked their performance and compared it with past years (ibid.). The majority of schools were found only to conduct quantitative analysis.

A recent study on how schools in Poland use results suggests that the situation is dynamic with evidence that schools are in the process of learning how to respond constructively to them in their work (Stożek, Kędracka & Rappe, 2015). Particular difficulty seems to be experienced with trying to ascertain the reasons for results and deciding how to tackle the question of raising effectiveness. Schools seem to be beginning to work with individual students to help them achieve their potential. Problematic is also establishing milestones and criteria against which to measure and report progress and growth in the target areas (Milecka, 2014a), and so in demonstrating improvement. This is reported as a reason for the awarding of grades of C or D in external evaluation reports (ibid.).

Work to improve the teaching process was found to include increasing work to develop skills that are needed to do tasks that caused difficulty (91% of schools), increasing learner activity through changing teaching techniques (64%) and teaching test-taking strategies (56%). Attempting to increase learner motivation was also reported, as low motivation is considered to contribute to low outcomes (Milecka, 2014a).

External evaluator reports show a tendency to value actions taken by teachers which could increase test scores, finding them "most effective". A lack of congruence was noted between what schools reported as having positive effects and what was noted by evaluators (ibid.).

Trends noted in 2013–2014 in reports on school evaluation were that how directors reported the use of exam results was not necessarily consistent with reports from teachers. An increase was seen in the types of information that were being used to monitor progress in school, in addition to the use of external exam results. This included small-scale studies done by teachers in their own classes and tracer studies of school leavers. In a few schools, information from taking part in national research projects, or PISA tests, were also mentioned. Of techniques reported with the aim of raising quality of work in lower secondary schools the smallest number related to the work of the teacher, with most aimed at individualization, improving the formative assessment process and increasing extra-curricular lessons on offer.

In this section we have given an overview of the legislation on pedagogical supervision for schools in Poland, one of the requirements of which is that schools analyze the results of external examinations and take subsequent action. This was followed by an overview of some of the research done which indicates how schools have responded to the legislation, which shows some indication of action taken following analysis of results having negative consequences, which could be described as negative impact.

The aim of this chapter has been to examine the uses made of tests and test results in society. The chapter opened with a consideration of the relationships between the validity of a test, its use, and the consequences of that use. We saw that Messick's unified concept of validity (1989), which encompasses the social consequences and impact of a test, is now the prevailing view in language testing contexts. Contemporary approaches to the validation of language tests are based on logical argumentation, one aspect of which includes test impact or consequences. Particular attention is paid within the validation argument to the test construct, which should be made clear in the test specifications, and the relationship between the test and the curriculum. It should be ensured that the test content adequately represents the curriculum. Factors which may affect the validity of performance testing (speaking) were considered to see how what Messick (ibid.) describes as construct under-representation or construct irrelevant variance may be avoided. We return to many of these issues in analysis of an examination in the Polish context in the next chapter.

We then considered the power of tests from social, philosophical and ethical perspectives. In the next chapter we return particularly to the notion of an examination with a gate-keeping function when we analyze a test at the threshold between two levels of the Polish education system. We have seen that for a test to function well, issues of equitability and fairness should be taken into account. If an exam appears to favour one group of

test-takers over another, then its validity may be called into question, with implications for the uses that can be made of its scores. In the next chapter we will see that this has important significance in the Polish context.

The role of tests and test results in accountability was then investigated, first in terms of the intentions behind such use of examinations and then through an overview of accountability in three educational contexts. It was noted that the use of large-scale tests for accountability may cause tensions, as different stake-holders have different expectations of the type and amount of information an examination may give. In all three contexts suggestions were made that the use of exam results in accountability is problematic and may lead to actions on the part of educators which could be determined as having negative consequences and which may compromise test validity. These issues will be of key importance in the study of a national exam in English which is the subject of the second part of this publication.

CHAPTER 3

The external foreign language examination at the end of lower secondary school: Theory and Practice

Understanding an examination (we will use the words examination, exam and test interchangeably) means understanding theoretical aspects underlying it, its rationale, aims and construct, which we understand as what the test professes to measure and how. In the previous chapters we have seen that washback and impact are associated with the relationship between an exam and the curriculum with which it is connected, so in order to be able to evaluate the nature of this relationship thorough analysis is needed of both curriculum and the examination. Once we understand what the curriculum intends to be taught we need to move to consider information about the test content, format, mark scheme, allocation of points and weighting. In this way we learn how the test aims are operationalized and, through comparison with the curriculum, are able to assess if there are issues with construct irrelevant variance or under-representation (Messick, 1989). Only once an understanding of the design and intentions of an exam has been obtained can we proceed to a study of that test in use in the context for which it was intended. Understanding the test in use also entails knowledge of the specifics of that context, as these can affect how the test functions.

Thus, taking us a step nearer to the study of exam impact which is the subject of the second part of this book, this chapter introduces the context of Poland and its education system in a period between 1989 and 2012, focusing on educational reforms which were initiated in 1999 and 2009. We look in detail at the external examination at the end of Key Stage 3 (*gimnazjum*) and in particular at the foreign language (FL) component.

Next we describe the historical origins of the FL exam and the context into which it was to be introduced. This is followed by an analysis both of the examination and supporting documents, the national core curriculum for FLs and the standards of attainment for the exam. The next section presents information about the results of the first three administrations of the FL exam (2009–2011) and feedback received by teachers. The final part of the chapter explains revisions of the core curriculum in 2008

and subsequent changes to the external examinations in 2012, including revisions to the FL exams.

The aim of the chapter is to set the context for the research study, described in Part Two, which took place in Poland from 2008–2012.

3.1. The Educational Reform of 1999

In May 1998 the Ministry of National Education (MEN) produced a blueprint for the planned reform of education, popularly known as 'the orange book', inviting comment and criticism from stakeholders over the next four months. In this book was laid out the plan to change from a two-tier to a three-tier system, making primary school last six years (instead of eight) and introducing a new type of school, named '*gimnazjum*', to cover the three years of lower secondary education.

The aim of such a change was to raise the level of education in society, by making secondary and higher education more common; to increase equity in education, by offering more equal opportunities; and to raise the quality of education (MEN 1998: 10). The rationale for the introduction of the new lower secondary tier was to extend general education by a year, to the age of 16, thus giving young people a year longer before they had to make decisions about their futures. It also aimed to group together learners of a similar age and stage of development, in order to better cater for their needs (ibid.).

In addition to the proposed change in school structure, a new system of external examinations was planned. The reform promised a national system with examinations at the end of each tier (primary, lower secondary and upper secondary), run by a new external body, on the basis of nationwide standards of education (MEN 1998: 30). These were to be published in two sets of documents, the Core Curriculum [pol. *Podstawa Programowa*] (MEN, 1997) and the Examination Standards of Attainment [pol. *standardy wymagań edukacyjnych*] (MEN, 1997).

The aims of the new external examination system (MEN, 1998) were to bring about comparability in examination content across the country by means of the new examination standards, which would be known to all the stakeholders, and also to allow comparability of test scores through the application of assessment criteria based on these standards. To ensure objectivity and reliability of assessment the examinations would be rated by teams of trained examiners who were external to the school. The school leaving examination document (*matura*) would become a nationally recognised certificate.

Until 2002 the school leaving certificate examination [pol. *matura*], taken at age 19, was the only external examination in the general education system. It was organised by the local education authorities [pol. *kuratoria*]

in each voivodeship, according to their own regulations. In addition, marking of written papers and rating of oral tests was done, in school and unsupervised, by the same teachers who had taught the pupils, raising concerns about reliability. There was substantial variation between regions in terms of content, form and level of the tests. During the FL *matura* examination, which lasted five hours without a break, there were ethical issues concerning assistance from third parties (including teachers) and use of crib sheets. Test fairness was also called into question through charges of teacher favouritism, or discrimination, and lack of objectivity in oral assessment, or rating of written work, largely arising from the absence of criteria for assessment. Test security was also problematic, with cases of examination papers available for purchase on the black market in the weeks before the examination.

Attitudes towards examinations were very different to those prevailing in Britain at the time and there was a widespread practice of turning a blind eye to irregularities. There was also a sense that teachers were colluding with students against a common "enemy", the school leaving examination, where teachers felt duty-bound to do everything possible to ensure good results for their pupils, both for the sake of their own reputations and for that of their schools.

As Poland developed as a democratic state, enjoyed increased contact with other countries and aspired towards greater integration with Europe, it became clear that reform of the examination system was necessary. Moves to reform the *matura* began with the Know-How funded SMART programme of the early 1990s. This grew into the *Nowa Matura* [New *Matura*] programme, which began to train selected teachers in the theory and practice of educational assessment. In the case of English, the programme began to introduce changes to the existing school leaving examination. In the (then) Katowice voivodeship, criteria for assessment of writing were introduced in 1997, despite considerable resistance from teachers. (The author served on the commission responsible for the *matura* examination in English in the Katowice voivodeship from 1991 to 1999. Reform of administrative areas changed voivodeships from 1999.)

The first national external *matura* examinations were offered at the end of upper secondary school in 2002 and were to include three compulsory examinations: Polish, foreign language (FL) and mathematics, at a choice of two levels. Reform of the FL *matura* examination had taken considerable time, effort and resources because of the size of the undertaking (approximately 400,000 test takers annually). Training of FL examiners (both to rate written papers and create (2002 version), conduct and rate oral tests had lasted from 1999.

Many teachers found the concept of a standardised system alien and felt threatened by the changes. The prospect that the new *matura* would replace

university entrance examinations also caused considerable public debate. Teachers and academics who supplemented their income running courses and classes preparing for the entrance examinations were vociferously against the idea. The planned introduction of a compulsory examination in mathematics at *matura* led to fears of large numbers of failures. In the face of widespread public concern, one of the first actions of the newly elected SLD government in autumn 2001 was to declare the new *matura* examination optional for test takers for 2002, to postpone its official start until 2005 and to remove the requirement of taking the mathematics examination.

We must also bear in mind the socio-educational context of the time. In 1999, the year in which lower secondary schools came into being, there was still a shortage of FL teachers with full qualifications. Up until 1989 Russian had been compulsory as a FL in school. With the change of system came freedom of choice of FLs in school and first choice was overwhelmingly English. However, with only a limited number of universities offering places, and becoming a teacher low on the list of an English graduate's priorities for reasons of salary and prestige, the number of teachers with full qualifications available was insufficient to meet the considerable demand. An alternative, in the form of three year teacher training colleges for foreign languages, was found and in addition legislation was drawn up to regulate the process for graduates of other subjects wishing to change their qualifications and become teachers of English.

Although the situation had progressively improved, in 1999 the shortage of fully qualified English teachers was still being felt, particularly in rural areas. In primary schools, for example, 43.1% of pupils in urban areas were studying English, as opposed to 30.1% in rural areas. Russian, by contrast, was being taught to 4.7% of pupils in urban schools, but to 20.6% of those in rural areas (MEN, 2000: 14). With a view to supporting FL teachers in rural areas under a programme "*Pozyskiwanie deficitowej kadry nauczycielskiej na obszarach wiejskich*" [Reducing the deficit of teaching staff in rural areas], MEN offered funding to teachers from small towns (population to 5000) to raise, or change their qualifications and this included an initiative called "Year 0", which gave such candidates an intensive year of language courses in teacher training college, designed to get them up to entry standard. There were also initiatives with the aim of increasing access to English, so that it could be offered in every *gimnazjum,* and to this effect each local education authority [*kuratorium*] appointed a representative responsible for the promotion of the teaching of FLs (MEN 2000: 44–47). In the light of the current almost universal teaching of English in *gimnazja* it is necessary to remind ourselves of the conditions prevailing as the new schools opened for the first time.

3.1.2. The external examination at the end of lower secondary school

The external examinations at the end of the lower secondary stage of education were designed to have two functions. First, to give a picture of the skills and abilities the learners had at that point in time, after three years education in lower secondary school, but also to indicate the orientation of the learner's abilities and so serve to guide their choice of upper secondary school. The examinations would be used as the basis for selection of pupils by upper secondary schools, replacing any entrance tests. The first external examinations at the end of lower secondary took place at the end of the first cycle in spring 2002. At this stage they included two papers, one in arts subjects and the other in mathematics and natural science. The results from the arts and mathematics papers were used for the purposes of selection for upper secondary school from the first examination in 2002.

As the first administration of the external lower secondary examination (ELSE) took place in the same year as the first administration of the new *matura* examination, which required considerable effort and resources, it was hardly surprising that at this stage an FL examination was not included, especially as the reform of the FL examinations at *matura* level had proved such a major undertaking. In addition, as we have seen, there was still an issue with access to choice of FLs in *gimnazja* caused by a shortfall in the number of teachers of FLs with full qualifications.

3.2. Origins of the foreign language component of the external examination in lower secondary school

This section traces the history of how the external examination at the end of lower secondary school came into being.

The *Strategy for the Development of Education 2007–2013* (Ministry of National Education and Sport, August 2005) gives a wide reaching analysis of the situation existing in the first decade of this century in Poland, outlining prevailing socio-economic issues such as unemployment and demographics. The document takes a European perspective and refers to the Lisbon strategy (2000) to which Poland, as a member state of the EU, was bound to conform. The main aim in education for those years was seen as to strengthen the human potential in the country through raising the level of education of the population (ibid.: 26). The "Education and Training 2010" initiative, signed by EU ministers in 2002, among other areas marked for development, highlighted eight "key competences for lifelong learning" which are seen as fundamental to making Europe a competitor in world markets, by helping people adapt as employment patterns change and by promoting an integrated society, with access to education for all. One of these competences is the ability to communicate in foreign languages.

The Lisbon Strategy called for EU members to adopt these principles and implement them in national educational policy by 2010.

The Polish Strategy document (2005) warned that the quality of schooling was under threat, as greater numbers of young people were continuing to further levels of education than in the past. Although the school system had responded to the transformation to democracy in Poland, this response had been piecemeal, with no overall solution put in place. The results were that education was falling behind and was not meeting the challenges posed by the new order. These challenges were numerous. Changing demographics and increased mobility needed to be addressed. Changes had been observed in attitudes, parenting and openness to the world. Issues of social and economic difference were seen as key, because of the strong influence of the home on educational achievement. An increased number of young people needed a more personalized approach to education, due to personal and developmental difficulties. New skills were needed to deal with the rapidly changing world. These include foreign languages, IT and entrepreneurship. Lifelong continuous education is needed to meet these challenges and this concept was, at that time, as yet unfamiliar in the Polish culture (ibid.: 27).

In response to these challenges the Polish Strategy for 2007–2013 set out to:

- allow all students to reach their potential
- prepare all young people to take an active and responsible role in society both in Poland and beyond
- work successfully towards a policy of inclusion for all
- respond to global changes and developments in science, technology and learning,
- respond rapidly and flexibly to changes on the labour market (ibid.: 8).

Within school this was to be tackled by increasing equity, by aligning the content of what was to be taught to the needs of a changing world and by adapting that content to the individual capabilities of each child. Changes in the core curriculum were seen as necessary to bring schooling nearer to the changing demands of the new century. The key competences (including foreign languages) were to be given central position and stress placed throughout the curriculum content on the development of skills, rather than knowledge per se. A positive attitude towards lifelong learning was to be promoted. By the end of upper secondary learners should have reached an advanced level in one foreign language and an intermediate level in a second. Foreign language should begin with 5 and 6 year olds in pre-school and *the external examination at the end of lower secondary school should include it.* (ibid.: 40, my italics).

In the Report on Language Education Policy Profile (MEN, 2005: 57, own translation) we read

> It would be good to consider introducing the second foreign language earlier (already in *gimnazjum*) and including a test of foreign language competence after *gimnazjum*. This is, however, connected with [the necessity of ensuring] the continuation of the same foreign language from one educational stage to the next. This is not always possible because of shortage of teachers. At least it would be possible to prepare Standards of Attainment for the end of the *gimnazjum* stage of education.

Thus we see that the inclusion of the foreign language component in the *gimnazjum* examination owed a great deal to the Lisbon Strategy and also to Poland's implementation of the European Key Competences, making foreign languages central to the planned revision of the core curriculum. In practice, implementation might be found, however, to prove problematic for reasons of the teacher shortage suggested in the Policy report. We shall return to these difficulties later.

3.3. The 1997 Core Curriculum for Foreign Languages

The national core curriculum replaced the previous 'minimum programmes' and its implementation is a requirement in all state funded schools. The main aims for FLs for the third educational stage (*gimnazjum*) in the 1997 version, for those taking an FL in lower secondary as a continuation from primary school were "To reach a level in the language which allows reasonably competent communication in everyday situations" (MEN, 1997). The core curriculum describes standards for the skills of listening, speaking, reading and writing and also the ability to distinguish facts from opinions, select information, and use monolingual and bilingual dictionaries and other information sources. Building on a foundation laid in primary school, the core calls for development and extension of language functions needed for everyday communication and for the learning of more complex grammar structures to deal with the present, past and future. The lexical resource is also to be developed, taking into account the cultures of places where the target language is spoken and with the use of authentic materials. Learners are also to learn more about the foreign culture and everyday life and behavior. They are to be able to distinguish between formal and informal language, and, in speaking, to be able to use compensatory strategies. The development of individual strategies for learning are to be encouraged.

The team drawing up the core for FLs had consulted the documents *Threshold Level* and *Waystage 1990* when deciding on the content, together with the UK Modern Languages Curriculum for 14–16 year olds, in anticipation of Poland's greater integration with Europe and the need for comparability in levels of attainment (Osiecka, 1997: personal communication). The decision to place communication at the heart of the curriculum was influenced partly by these documents, but also by the national programmes for primary (1990) and secondary schools (1987) in use at the time.

The programme for English in upper secondary comprehensive school (WSiP, 1987) had included communication as one of its main aims, along with language competence. Communication was understood as "the skill of effectively making oneself understood, that is effectively communicating target content and intentions in a way appropriate for the situation in which the act of communication takes place and for the roles of the speaker and interlocutor" (ibid.: 6). The 1990 programme for English in primary school (MEN, 1990) also clearly set out communicative intentions. In the Rationale (MEN, 1990: 8) we read: "the language taught serves for communication, it is a tool for practical use. Particular emphasis is placed on learning the spoken language presented in communicative situations…" The 1997 core curriculum can therefore be considered a continuation of earlier thinking.

3.3.1. The language construct behind the core curriculum

Having established the antecedents of the 1997 core it is necessary to attempt to establish the model of language inherent in it, as an understanding of this is needed in interpretation of the construct of the external examination. References to communicative competence, which we have seen was stated as the main aim, place the language model as relating to that of Canale and Swain (1980) and Canale (1983). These models hypothesize communicative competence as comprising four areas described as grammatical, discourse, sociolinguistic, and strategic competences. All of these competences are referred to in the core. Overt reference is made to developing grammatical structures. Discourse competence can be observed in references to "formulating coherent speech"; "understanding longer and more complex texts" and "making short coherent notes on what someone says." Sociolinguistic competence is referred to in "Initiating, maintaining and ending simple conversations"; "making appropriate linguistic responses to the interlocutor"; and the ability to "distinguish formal and informal styles." Strategic competence is referred to in the ability to use compensatory strategies (Canale 1983: 339).

Although Canale and Swain's model has been called into question by research (Allen, Cummins, Mougeon & Swain, 1983; Swain, 1985; Harley, Allen, Cummins & Swain, 1990), which failed to find adequate evidence for the existence of all four competences, and for the fact that it does not offer adequate explanation of the relationship between the different competences (Bachman, 1990: 87), or adequate description of "the context in which language use occurs" (Bachman, 1990: 81), it remains highly influential. Bachman (1990: 87) reworked the model, including similar areas, but re-grouping them under 'organizational' (which encompasses grammatical and textual competence) and 'pragmatic' competences (which include illocutionary and sociolinguistic competences). This influence can

be traced to the Common European Framework (CEFR) (CoE, 2001: 1) whose main aim is to describe

> in a comprehensive way what learners have to learn in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively.

CEFR defines communicative competence as including three sets of competences, linguistic, sociolinguistic and pragmatic (ibid.: 13). Sociolinguistic competence, following Hymes (1972), concerns the ability to use language appropriately in different contexts and with different interlocutors, including those from other cultures. Pragmatic competence, with reference to the work of Levinson (1983) and Searle (1969), refers to the "functional use of linguistic resources (production of language functions, speech acts), drawing on scenarios or scripts of interactional exchanges," (CoE, 2001: 13). Proficient use of discourse is included as part of this competence. Thus we can see that the language construct of the CEFR shares many commonalities with Canale and Swain's model and is entirely consistent with the proposed construct of the core curriculum.

3.4. The rationale of the foreign language component of the external examination at the end of lower secondary school (*gimnazjum*) (2009)

Finding documentation of the rationale of the foreign language (FL) component in the external examination at the end of lower secondary school (hereafter referred to as ELSE) is not as straightforward as for other similarly important decisions in education. Unlike the 1999 reforms (as a result of which both lower secondary schools and the whole external examination system came into being), which were accompanied by a series of booklets published by the Ministry of National Education (MEN) clearly stating the decision process and the thinking behind the changes, uncovering the rationale of the new FL ELSE required considerable due diligence, as its documentation is much less complete. There appears to be no such policy document available in the Central Examination Board (CKE). As a result it has been necessary to piece together the rationale from a variety of sources. What follows is a personal interpretation.

Between 2007 and 2013, as one part of a larger project funded by European Structural Funds (EFS) under the title "Raising the Quality of the External Examination System" there was Project V.1 "Piloting the foreign language examination in lower secondary school" co-ordinated by Dorota Obidniak (CKE EFS website). In an article in *Języki Obce w Szkole* (2007) she describes the new initiative to introduce an external FL examination in

the context of "long-term national educational policy" (ibid.: 66) and offers what appears to be a rationale for the new examination.

> A well-formulated examination may serve to raise qualifications and the quality of teaching. Above all:
> 1. Impact on the raising of language awareness
> 2. Impact on the improvement of learner skills
> 3. Motivate learners to study because of its attractive form and comparability
> 4. Allow learners to compare their language skills with others internationally (Europass)
> 5. Be written in the Polish and international system of language examinations
> 6. Impact on the raising of teacher qualifications
> 7. Support new ways [nowoczesne formy] of teaching FLs (ibid. but numbering added)

Obidniak also considers the importance of the social function of the examination, which she says should be created in such as a way as to:

> 8. Cause an improvement in access to education and an increase in social congruity
> 9. Serve as a factor for integration with other peoples and cultures
> 10. Popularize the need to be able to communicate in FLs in the world today (ibid. but numbering added)

It should be noted that, while undoubtedly noble in intention, these aims are not very clearly defined. It is only through consultation with other documents that they can be more fully interpreted.

Jarząbek, a member of the Project V.1 team, expressed the hope (2007: 70) that the introduction of the FL examination would increase the importance of FLs in lower secondary school, where they had hitherto been seen as lower in rank than subjects which were included in the external examination. This goes some way to helping us understand point (3) above. We can infer that the FL examination is intended to raise the profile of FL learning, focus learners' attention on the importance of FLs and thus motivate them to learn them.

Obidniak explains (2007: 66) that in drawing up the test specifications, international documents were being referred to, in addition to information from the Polish context. She cites the Lisbon Strategy, documents from the European Commission and Parliament on the key competences, *Europass*, and the *European Qualifications Framework*. *The Common European Framework of Reference for Modern Languages* (CEFR) and the *European Language Portfolio* (ELP) are also mentioned as reference documents. Jarząbek (2007: 68) expands on how these last two publications are used to make the assessment process more objective and to increase validity and reliability. She offers a table comparing some of the operational aims taken from the Test Syllabus with descriptors from the CEFR (ibid.: 69).

Europass, the document referred to in point (4), corresponds to the "Language Passport" element of the *European Language Portfolio*. Designed to document the holder's experience and achievements in foreign languages, it serves as a record of levels attained, which are evidenced by recognized certificates and supplemented by descriptions of course content covered. This pass can then be used when applying for a university place in any country within the EU, or as part of the recruitment process when looking for work in one of the other member states. The document represents part of European policy on mobility and is designed to facilitate movement between different countries by simplifying the process of recognition of language competence. Replacing the 'old' notions of 'elementary' or 'intermediate' ability with standard levels from A1 to C2 and descriptors taken from the CEFR (2001), the *Europass* aims to give the holder a Europe-wide understood description of what they can do in foreign languages and how well they can do this.

In order for examination certificates to be able to be entered in the *Europass* the examination needs to have been linked to the CEFR via a process which is set out in the *Manual: Relating language examinations to the Common European Framework of Reference* (2003; 2009, CoE Language Policy Division). This is a two-stage process involving first descriptions of all aspects of the examination specifications (aims, tasks, text types, procedures, assessment criteria and the marking and scoring process), and then a further, more complex stage which calls for empirical verification of the examination. Ethically it is not possible to claim that an examination is at a certain level on CEFR until it can be demonstrated with evidence that both stages of the linking process have been completed. In practice, however, this is not always followed by examining bodies and this is a bone of contention in the testing community.

We can infer from the reference to *Europass* and from Obidniak's further references to CEFR and ELP that there appears to be an intention of linking the new FL examination to CEFR. This hypothesis may possibly be confirmed by the way Jarząbek (2007) demonstrates the relationship between the examination standards and CEFR. Głowacka (2008) criticizes the language in which the examination standards are written as being "from an institutional perspective" and claims that they do not give learners a clear idea as to what is required of them. She argues that the standards need to be carefully explained in concrete can-do statements, rather than in general abstract terms. To show this she provides a table where she takes examples from the examination standards, identifies the corresponding can-do's from the general CEFR descriptors, next takes the matching descriptors from the CEFR self-assessment grid and finally gives the corresponding descriptors form the Polish version of the ELP for 10–15 year olds (CODN, 2004) to illustrate these same aims in language accessible to the age of the target group of test takers. Both these examples (Jarząbek and Głowacka) serve to

show that clear linking and correspondences appear to be intended between the examination standards and both CEFR and the ELP and that this was most likely intentional from the outset, possibly to facilitate a subsequent linking process as outlined in the *Manual.*

Point (5) in Obidniak's list, allowing the examination to take its place in both the Polish and international system of FL examinations, seems in some ways to be an extension of the previous point. We can infer that if the examination standards (test syllabus) have links to CEFR, then most likely there is also an intention for further linking of the contents of the tests themselves. The list of topic areas in the Test Syllabus, for example, can be linked to the CEFR list of domains of use (2001: 48–49). The structural- lexical list of contents for the English examination can be traced to *Waystage 1990,* a Council of Europe document which pre-dates CEFR. In this can be found a language-specific inventory of grammar (van Ek & Trim, 1998: 75–94) and lexis (ibid.: 95–105), unlike CEFR which refers to all modern languages and so has no detailed inventories of lexical or grammatical content. *Waystage* has been placed at level A2-A2+ on the CEFR scale (CEFR, 2001: 23) We can therefore infer that there appears to be an intention to produce an examination in English at approximately level A2, which uses *Waystage* as a reference document for its contents and the CEFR for its specifications. In this way the new FL external examination appears to aim to make itself comparable to other recognized FL examinations, both by means of its specifications and content, and possibly also by means of the processes and procedures which may be used to ensure its validity and reliability.

3.5. Analysis of the ELSE FL exam

In this section we will study the documents which describe the test in English which forms part of the ELSE, try to understand the test construct, and compare the test with the FL core curriculum.

3.5.1. Test specifications

Test specifications are understood as a declaration of "about what the test tests and how it tests it" (Alderson, Clapham & Wall, 1995: 9). The specifications are needed by different groups of people for different purposes. First they are needed by the team who are to create the examination and so the document should include such aspects as the target audience of examination, what the test is to be used for, how it will assess what it aims to test, what it will contain, what form it will take, and how long it will be (Alderson et al., 1995: 19). The specifications are also needed for those who will use the tests, but these may be given in a different format than for those designing and creating the test. Teachers and learners are more concerned about the

test content, format, assessment procedures and criteria than about the theoretical concepts underlying the test. Alderson et al. (1995: 9) refer to this simplified form of the test specifications as the test syllabus.

In order to establish whether the test is valid, that is 'tests what it is supposed to test' (Alderson et al., 1995: 10), (see discussion in chapter two), a different set of specifications are needed which make clear the theoretical construct upon which the test is built "what theories of language proficiency underpin [it] and *why* the test is the way it is" (ibid.). It seems logical that in studying the impact of an examination it is this aspect of the specifications which are of key importance. We need to understand the theoretical model of language on which the test is built, what the test designers believe "language" is and how they operationalize this. In other words, what they believe language proficiency, or language competence, constitutes in terms of the examination and how they intend to capture this and evaluate it. Evaluation will include how language use is to be assessed, by means of what use of what criteria, and how it will be scored. Allocation of points awarded for each test item may serve to reveal the relevant importance attached to the different skills and abilities tested. If we are to study what impact the test has made, then we need a clear understanding of the test designers' beliefs and intentions in constructing the test.

Returning to the people for whom the test specifications are needed, we can identify further stakeholder groups, one of which are those who intend to use the test scores. They need to understand the purpose of the test, what it contains and how this is tested, in order to be able to judge the usefulness of the information given by the results for their purposes. This aspect of how the test scores are communicated and the uses to which they are put has been discussed in more detail in Chapter Two.

The final parties interested in the contents, form and procedures of the examination are course designers and publishers, who wish to develop syllabuses and material aimed to prepare learners for taking the examination. They may also want to write practice test materials and need to ensure that these are as close to the "real" examination as possible. We will return to the role of publishers and published materials in the foreign language ELSE later in this chapter.

3.5.2. The Test Syllabus

In the absence of any other documentation, a situation which is not uncommon, as more detailed specifications of high-stakes examinations are often confidential and only made available to item writers, we must deduce the construct of the FL examination from the Test Syllabus [*Informator*] (CKE, 2007). First published in August 2007, two years before the first administration of the new examination as required by Polish law (MEN,

2007), a test syllabus was offered for each of the six foreign languages offered for assessment (English, French, German, Russian, Italian, and Spanish). The reason for this was that the syllabus contained a language specific lists of standards of attainment [pol. *standardy wymagań*], in addition to information about the test format, procedures and sample tasks. The Test Syllabus makes a clear statement about the connection between the national core curriculum [pol. *podstawa programowa*] and the examination standards of attainment.

3.5.2.1. Interpreting standards of attainment for the FL examination in English

The first part of the Test Syllabus consists of a catalogue of the standards of attainment for the examination. These are divided into three areas: comprehension of spoken text; comprehension of written text; and interaction. Each area is broken down into sub-skills. For listening these are:

Stating the main idea of the text
Giving the context of the situation
Stating whether the text contains given information; finding or selecting information in the text

For reading these same sub-skills also appear but three more are added:

Stating the main idea of parts of the text
Stating the intention of the text/of the author of the text
Recognizing relationships between different parts of the text

The fact that there are twice as many sub-skills for reading as for listening signals that reading is considered more important in the test construct. This is confirmed by the allocation of points, where reading is given 20 out of the total of 50, as opposed to listening, which is awarded 10. The reasoning behind this disparity is not at first clear, but the answer may be found in the core curriculum, where the language in which listening comprehension is described differs somewhat from that used in the description of reading comprehension. For listening learners are expected to "understand the general sense and main points of dialogues and speech of native speakers"; "to understand the sense of simple texts…"; and to "understand the sense of a text containing elements which are not understood, but the meaning of which the learner can deduce form the context". The description of reading, by contrast, refers to "longer and more complex" texts; the ability to discern the "main sense of a lengthy text by skim reading" and the ability to "look for required information or details in a text which is in part not understood." We can see that there appears to be an expectation that the learner's understanding of written text will be in advance of their understanding of spoken text at this stage. This is because of the different natures of reading and listening.

> The understanding of orally produced texts is subject to time constraints: they are produced only once in a linear form which does not allow the listener to go back to earlier parts or reflect on the text as a whole as easily as may be done with written texts.
> (van Ek & Trim, 1998: 43)

In the light of this interpretation the test construct, which weights reading more heavily than listening, seems to be justified. This is not to say that reading is considered more important than listening, simply that at this stage it is expected that the learner will be more skilled at dealing with written text. It should be noted, however, that there is no explanation of the rationale behind the weighting difference given in the Test Syllabus, which means the situation is open to misinterpretation by teachers, students and their parents. What is allocated fewer points may easily be taken to be of less importance.

The third area in the Standards, Interaction, includes three sub-skills:

> The learner
> - reacts with linguistic appropriacy in given situations or contexts, particularly with the aim of seeking, giving or relaying information, or declining to give information; initiating, maintaining and closing conversations.
> - recognizes and uses appropriate lexico-grammatical structures necessary for effective communication
> - can convey in the FL the contents of a text read in Polish, or the contents of material presented graphically.
>
> (MEN, 1997)

However, when we study the sample tasks offered in the *Informator* we see that the intention is for interaction to be tested indirectly, that is not by means of a test of speaking per se, but through the medium of writing and response to aural prompts.

3.5.3. Comparison of the core curriculum and the FL examination standards of attainment

If we compare the examination standards with the core curriculum contents we see that there is only a partial match. Listening and reading are covered with close correspondence. We can infer that in order to be able to deal with the texts proposed for reading the learner will need to develop their vocabulary, at least for recognition. However, there are problems with the skills of speaking and writing, which are under-represented in the examination standards (Messick, 1996). The core curriculum standard "The ability to formulate short, fluent and coherent speech on given topics using appropriate grammatical forms to express the present, past and future," or the standard referring to the use of intelligible pronunciation are not included. Of the sub-skills of writing only the first, "formulating a piece

of writing with a reasonable range of language" and perhaps "making simple and coherent notes from what someone says", have the potential to be tested if we study the sample test tasks given in the Test Syllabus. For listening, sample tasks include completing notes (with one or two words, or numbers) on the basis of what is heard, while for writing the learner could be asked to write a sentence or two in response to questions about a picture, or to complete blanks in a text. These are, however, minimalistic ways of testing what is described in the core. In addition, as we shall see, these task types were only used in the mock examination in October 2007 and were subsequently removed.

To conclude, the skills of speaking are operationalized in the test construct as interaction and this is tested indirectly through a variety of recognition tasks. Productive tasks concerned with indirect testing of speaking, given as examples in the Test Syllabus, were extremely limited in scope and were subsequently not used in the main examinations. Charges can therefore be made of construct irrelevant variance (Messick, 1996), as information on speaking per se is not gathered by the test, but only indirectly through the medium of written text. Skills of writing are tested at only the level of completing single words in blanks, as the planned productive task, designed to elicit samples at sentence level, was also removed before the main examination. What remains is a construct of skills of listening and reading which follows the skills outlined in the core curriculum. This is a situation which has the potential for causing narrowing of the syllabus if teachers teach to the test. The construct of the test does not match the language construct modeled in the core curriculum.

Smolik (2008) writes highly critically about the proposed new examination. He questions the aims (which he also had difficulty finding), given on the Project V.1 website, on the grounds that any idea that the results of the examination will be comparable at a European level, such as through the use of *Europass*, is unrealistic in the medium term. Any such comparison would require the examination to be linked to CEFR, a process which had not yet been completed for the *matura* examination, due to constant changes in the examination procedures (Smolik, 2008: 186). Secondly, he argues that if the purpose of the examination is to give learners information about their language competence, then the form in which results are planned to be given is inappropriate. The learner will receive only a total number of points, or a percentage, the real purpose of which is to calculate descriptive statistics for the whole population (ibid.: 186) and this will not give them any detailed or helpful information.

Smolik also expresses doubt about setting the proposed level of the examination at A2 (CEFR) (ibid.: 187–188). He draws attention to the very mixed levels of language ability in lower secondary schools and questions the psychological impact of having an examination at a level which does not present a challenge for many learners, while at the same time being

very high for others. He suggests that a two-level examination, such as that offered at the school leaving certificate, would be more appropriate, thus allowing for the range of ability. We will return to this in discussion of the results of the first administration of the test.

He also has reservations about the form of the examination, particularly questioning the decision not to include a direct test of spoken ability, which he predicts will have negative washback. He questions the requirement in the planned productive writing task to write in complete sentences, which he claims runs counter to the notion of communicative competence (ibid.: 188). He concludes that limiting the examination to a test mainly of reading and listening poses the threat that practice in speaking and the teaching of writing may be relegated to second place, as teachers face the task of preparing learners for the examination (ibid.: 189). In short, he feels that the examination in its planned form does not augur for success and potentially may have negative consequences in the classroom.

3.6. Support for teachers available in the period leading up to the new examination

Between late spring 2008, and the first examination (April 2009), there was intense activity from publishers of English language course materials, with conferences, brochures, free sample pages and downloadable practice tests for teachers registered on their web sites. If we consider that in the year 2009, 347 888 students from lower secondary school (CKE, 2009) took the exam in English, we understand that there is considerable commercial interest in examination preparation materials. In addition, the Exam Syllabus (*Informator*), produced by the Central Exam Board (2007), although giving clear information about the format of the tests and examples of all possible task types, did not include sample practice tests. This led to speculation on the part of teachers and some concern, naturally caused by the fact that they had no clear picture of what exactly their learners could expect in the exam. The Central Exam Board ran a national mock exam in November 2009, for which all lower secondary schools were eligible, but, as teachers are required to prepare their schemes of work [pol. *rozkład materiału*] for the year before the start of school on 1st September, this could be regarded by some teachers as late for the purposes of planning. Any additional information prior to September 2008 was therefore considered of great value by teachers. I personally attended three conferences in Katowice, organized by different publishers in August 2008. Each was attended by around 350 teachers. Exam preparation books were creating great interest, with informal conversations in the break focusing on comparing the books available from different publishers and discussing which seemed the best value (data from field notes).

By contrast, official information on the examination given directly to teachers was at a premium. School principals attended conferences organized by the Regional Exam Boards about the new examination, and were then to convey this information to the language teachers in their schools. Such information centred on administrative and organizational matters, which are the main concern of school heads and contained little content information for the subject teacher (data obtained during piloting of school principal interview). Some regional conferences for language teachers were organized by the exam boards, but numbers were usually limited to one teacher per school. Although information was available on the Central Examination Board website, it would seem that not all teachers looked for this, preferring the information to be delivered to them. There also appeared to be teachers who found reading and interpreting official information challenging and who preferred to attend a conference presentation, where it was delivered to them in a more accessible form (personal communication with publisher's representative, based on teacher feedback forms). The form of the official information, the language of the discourse in which it is presented and its graphic format, is an area that the national and regional examination boards could work to improve. Teachers seem to need clear, condensed information, in simple language, presented in attractive visual format, rather than lengthy text with complex sentences. To sum up, although no empirical data is available on this, it would appear that the publishers' activity, in comparison with the form and amount of official information available, played a key role in informing teachers. As the publishers' agenda was not only to inform, but also to promote and sell their materials, the fact that teachers received information in this way may have contributed to a focus on exam-preparation books.

3.7. The ELSE in use: 2009–2011

Before the first test administration it was announced that there would be an interim period of three years, until the 2012 examination, when the points from the new FL of the ELSE would not be counted towards selection to upper secondary. The reasoning given was the varied situation regarding the teaching of FLs in lower secondary schools, where some children were beginning the FL, while many were continuing it from primary school. The situation in special schools, where children with slight intellectual impairment had only 6 hours of FL through the *gimnazjum* cycle (as opposed to 9 hours for children in mainstream schools), and the case of the school year 2007/2008 and earlier, when some children in special schools had not had FL lessons at all, were also cited. For these reasons it was felt that results of young people in class three *gimnazjum* would not be comparable and so the results should not be taken as a factor in deciding their futures (MEN, 2007).

3.7.1. Results of the first examinations 2009–2011

As Smolik (2008) predicted, the first examination proved very easy for some of the *gimnazjum* students and challenging for others, resulting in a bi-modal distribution. Out of a possible 50 points the mean score was 30.6. This is misleading, however, because of the abnormal distribution. The largest group of learners scored between 19 and 23 points, while a second group scored between 42 and 49 points (CKE Report, 2009).

Comparing results in the FL external examinations from one year to the next is problematic. To be able to write of changes or improvements in performance we need to be certain that the test tasks from one year to the next are comparable in all respects: format, content, level of difficulty, and number of tasks testing each standard. While the Test Syllabus offers a framework within which all test tasks must fit in terms of format and content, it does not ensure that from year to year the tasks will be comparable in other respects. The Syllabus allows for different types of tasks (multiple choice, multiple matching, and true-false) and states in which section of the test these may appear, but it does not say which task type is to be used to test which standard. It is conceivable, therefore that in one year the same standard may be tested by one task type and in the following year by a different one. Nor is it certain that the same sub-skills will all be tested, or that the number of items testing each sub-skill will be the same year on year. If the skill is represented in the test in a different way each year, because different sub-skills are included, or because different weighting is given to those sub-skills by their having more or fewer items, then we cannot compare results on that skill from year to year, as the skill itself is not operationalized in the same way. Table 3.1 compares the distribution of items testing different sub-skills in the first three years of the FL examination.

The problem of comparability as regards sub-skills is best illustrated by looking at how listening is operationalized (see Table 3.1). While there are few differences in the number of items testing listening for details, we see that as far as understanding the context is concerned the 2010 test contained 4 items, as opposed to only 1 in the other years. Similarly the testing of understanding of the main idea is not constant, being tested with 3 items in 2009, not at all in 2010 and by only one item in 2011. Testing of reading shows somewhat less variation, but in 2011 understanding the main idea of the text was not tested. In 2011 50% of the score for reading was comprised of reading for detail items as opposed to 30% in previous years.

There is a further problem where a sub-skill is tested by only one item, as the score is affected by standard error of measurement. This means that we are unable to draw reliable conclusions about performance on that sub-skill on the basis of scores on one item. Additionally, it is difficult to decide whether it is a feature of the *item* that determines the score, or the level of the test-takers' competence. In 2010, for example, the sole item testing

the ability to understand the main idea of the text had an average facility value (0.60), but this does not tell us whether it was simply just a good item, or whether the test population were reasonably good at this sub-skill. There is simply not enough information from a single item. In the 2011 exam, there were two items testing the ability to identify the context of the text, one of which had a high facility value (0.86) suggesting the item was very easy, while the other had a value of 0.57, suggesting it was reasonably difficult. In short, on the basis of a small number of items, we do not have enough reliable information to be able to describe the abilities of the test population with regard to some of the standards in the syllabus. If one of the aims of the examination is to give useful information to teachers then this is a serious issue.

Table 3.1. Distribution of number of items testing different sub-skills 2009–2011 (Data taken from CKE reports 2009, 2010, 2011)

Skill	Sub-skill	2009 No. of items	2010 No. of items	2011 No. of items
Listening	For details	6	6	**8**
	Context	1	**4**	1
	Main idea	3	**0**	**1**
Reading	Main idea of sections of text	5	4	4
	Specific information	6	6	**10**
	Context	2	3	2
	Author's intention	3	3	1
	Relationship between parts of text	4	3	3
	Main idea	0	1	0
Interaction	Chooses appropriate function	8	9	8
	Mediates	9	8	8
	Correct lexical-grammatical structures	3	3	4

The next problem for comparing performance year to year relates to the level of difficulty of the items in the test. To be certain that items are at the same level, they must be standardized following a rigorous procedure, using a panel of expert judges, and then piloted on a random sample, using a careful cross-sectional design to avoid bias, compared statistically for difficulty, using facility value and discrimination indices and preferably also analyzed using item response theory (IRT), which gives information about how different members of the test population scored on that item. We

expect that a difficult item will only be answered successfully by test-takers who have high scores on the whole test, while easy items should be done correctly by the whole test population. Using Rasch analysis items can then be placed on a scale of difficulty. In this way a bank of items of comparable level of difficulty can be created. This is a slow process, as it either means that a very large number of items must be written, standardized, piloted and analyzed initially, or that this happens successively over time, which involves the additional use of anchor items, items which would be used in all subsequent pilot tests to provide a fixed point for comparison. (The Central Examination Board is in the process of constructing an item bank at the time of writing).

Comparability in main tests can also be ensured through use of anchor items. These are items which are used in successive test administrations without change. Statistical comparison can then be made of how different test populations fared on the same items and reliable statements made about differences in performance. However, in the context of Polish external examinations this is not a viable option. Tradition has determined that examination papers are made public immediately after the test has taken place, together with the answer key and transcript of the recording. This therefore precludes any possibility of re-using any item in a later test, as having been made public they are automatically no longer reliable. Some test-takers may have seen them and tried them, while others have not, thus rendering the requirement that all test-takers have the same treatment invalid (Szaleniec, 2010). Anchor items across main tests would only be feasible if all the test papers remained secret.

The next problem when it comes to attempting to compare performance on the basis of test results from year to year is that the test populations are not the same. Clearly they are all learners in the final class of lower secondary school, but there the similarity ends. As every experienced teacher is aware, year groups differ widely in abilities, attitudes and motivation. As regards foreign language experience between 2009 and 2011 considerable changes were taking place in access to English lessons in lower secondary and primary schools. Changes were also taking place in teacher qualifications, as has been discussed earlier. Thus, to say that the test populations themselves are comparable is also questionable. In short, attempting to compare test results from year to year is a minefield.

What is it possible to say in general on the basis of the results of the first three administrations of the external FL examination? There are noticeable differences in performance depending on the location of the school. The results of the first test in 2009 showed that students in the countryside and in towns with populations of up to 20,000 scored lower on all parts of the examination. There was an average difference of 6.01 points between children attending schools in areas classified as '*wieś*' [village] and those

in cities with populations of over a hundred thousand. (It should be noted here that the classification '*wieś*' is administrative and does not relate to the size of the resident population, nor does it necessarily indicate that the area is entirely rural, as it can be subdivided into rural and urban administrations.) Professor Konarzewski, then director of the Central Examination Board (CKE), commenting in the press on the results of the first FL ELSE examination remarked:

> This is a huge difference. In Poland we have two kinds of English, urban and rural… These results are a warning for schools and local education authorities [*kuratoriów*] We have to find out who is teaching languages and with what methods…
> (Pezda, 10.06.2009)

Dorota Obidniak, in the same press article, explained that the results show variation as they represent three different populations: those who have only had three years of English in *gimnazjum*, those who have had six years, having started in primary class 4, and those who have had English from the first class of primary, or even from kindergarten. She states that parents in rural areas have less choice than those in cities, as in rural areas there is a general problem with shortage of kindergarten places, to say nothing of kindergartens offering foreign languages.

Conversely, the results for Russian showed dramatically better results in rural areas than in large cities, with a difference of 11.55 points. Both Konarzewski and Obidniak point to teacher availability and teacher qualifications as an explanation. They claim that shortage of teachers of English in rural areas meant that either Russian was being taught (and evidently taught well judging by the results), or that parental pressure had forced a switch to English, meaning that schools had employed partially qualified teachers. In other words, town and 'country' had different opportunities for learning (Kunnan, 2004), raising issues of fairness with relation to the test.

The trend for regional difference continued in results in both 2010 and 2011. The results for different locations are shown in Table 3.2. A report from the European Foundation for the Development of Rural Areas in Poland entitled "English – an opportunity for whom?" took up the issue of equity, underlining that other papers of the *gimnazjum* external examination do not show bias to urban areas. The results indicate that "English is the Achilles heel of rural schools" (*Rolniczy Magazyn Elektroniczny*, 2011) and the authors call for dialogue with and response from the Ministry of Education. Bearing in mind that from 2012 the points from the FL ELSE would count towards selection for upper secondary places we see that this bias against test-takers from some locations is a serious ethical issue which has the potential to compromise test validity.

Table 3.2. Results of the examination in English from 2009–2011 according to location

	Mean scores (standard deviations) of whole test		
	2009	2010	2011
National results	30.63 (10.84)	29.88 (11.68)	28.28 (12.50)
Location of school			
'wieś'	27.89 (10.49)	26.87 (10.84)	25.11 (11.29)
Population to 20,000	29.74 (10.89)	28.91 (11.42)	27.26 (12.12)
20,000–100,000	32.02 (10.89)	31.50 (11.63)	29.89 (12.58)
Above 100,000	33.90 (11.18)	33.84 (11.84)	32.01 (13.02)

A second trend which appears throughout the first three test administrations is for girls to score slightly higher than boys on all parts of the test, with mean scores showing between 1.5 and 2 point advantage for girls. Differences are marginal on the listening test, but larger for reading and interaction. The results are shown in Table 3.3.

Table 3.3. Results of examination in English 2009–2011 according to gender

		Mean scores (standard deviations) on sub-tests		
	Sub-tests (maximum no. of pts.)	2009	2010	2011
Whole population	Listening (10)	6.52 (2.16)	6.92 (2.55)	5.81 (2.68)
	Reading (20)	11.98 (5.45)	11.09 (5.10)	10.58 (5.60)
	Interaction (20)	12.14 (4.17)	11.87 (4.93)	11.89 (5.13)
Girls	Listening (10)	6.59 (2.11)	7.03 (2.51)	5.95 (2.66)
	Reading (20)	12.31 (5.41)	11.52 (5.07)	10.91 (5.65)
	Interaction (20)	12.52 (4.07)	12.23 (4.78)	12.50 (5.0)
Boys	Listening (10)	6.44 (2.20)	6.82 (2.58)	5.67 (2.69)
	Reading (20)	11.66 (5.48)	10.66 (5.09)	10.26 (5.53)
	Interaction (20)	11.77 (4.23)	11.53 (5.04)	11.30 (5.19)

3.8. Feedback information for teachers

After each test administration information is made publically available on the Central Examination Board (CKE) website. First preliminary results are published, indicating general trends, and then a full report which includes item analysis, showing the level of difficulty of each item for the whole test population. In addition to the national reports, each of

the eight regional examination centres publish reports about their areas. These give information such as mean scores for the voivodeship and for smaller administrative areas (*powiat, miasto*), as well as individual reports for each school. These school reports are general, giving mean results for the whole test and sub-tests. Both national and regional centres offer some interpretation of the results and comments. In some years conferences for school representatives (eg. school head and a teacher) are organized by the regional examination centres where the 'local' results are presented and interpreted and compared with national results.

Other than the official reports from the examination centres, feedback and information for teachers is often made available by publishers. Although the information part of such conferences is usually based on the national preliminary, or main report, and is presented reliably, the publishers have their own agenda for making this information available, in short, to promote sales of their books. This aside, such conferences, usually held in late August just before the start of the new school year are extremely popular with teachers. Conferences are usually organized in several locations within a voivodeship, on several dates, with information sent directly to teachers on the publishers' mailing list, in addition to announcements made on social media. At such events free materials are usually made available in the form of booklets or downloads. Macmillan, for example, in 2009 following the first results, published a booklet *Klucze suksesu* [Keys to success] (Piotrowska & Skulski, 2009) which summarized and interpreted the results and then made practical recommendations for teachers how to act to improve their school's results in the next examination. Useful ideas were included together with sample activities, which served a dual-role: to give teachers genuine suggestions how to develop their teaching, but also to promote sales of an examination preparation book.

3.9. Reforms of 2007–2008

As we have seen, the document which sets out the content of what is to be taught in school is the national core curriculum. Any national external examination must take this list of contents as its base, and consequently any revision to the core will entail changes in the examinations.

3.9.1. Revision of the Core Curriculum 2008

In 2008 substantial changes were made to the National Core Curriculum in all subjects. The reasons behind the reform lay in the changes made to the educational system in 1999 with the opening of lower secondary schools. These changes had helped to raise the level of achievement in core subjects (as seen in the results of the international PISA tests in 2000, 2003 and 2006)

and contributed to an increase in the number of young people taking the *matura* examination and entering higher education. In school this meant that a substantially higher number of learners were staying longer at school and choosing general education, rather than technical secondary schools. This change had led to a sense that educational standards were falling, as upper secondary education was having to cater not only for a larger number of academically inclined young people than in the past, but also for those who previously would have taken more vocationally oriented options. The authorities, determined to make higher education more universally accessible and so raise the standard of education of the general population (following the spirit of the Bologna Process), deemed that curriculum reform was needed to address the problem. Schools needed to change their approach to less academically inclined young people, and offer them the same opportunities as their more able counterparts. One way of doing this was through revising the content of what was to be taught.

One area identified as problematic in the school process was the trend for repetition of material in each of the different educational stages. Teachers of foreign languages, for example, routinely began again at elementary level in the first class of an educational stage (e.g. class one *gimnazjum*; class one *liceum*) regardless of the level of the pupils. While undoubtedly there were differences in level between students coming from different schools (caused by patchiness in teacher qualifications, frequent staff turnover, or absences resulting from teacher shortages, in addition to individual learner differences) some learners had managed to attain a level above elementary and were forced to tolerate repetition of an introduction to the language. The result was that rather than having six, or nine, years of foreign language the pupils had, in fact, experienced a three year course, repeated two or three times. The curriculum reform, therefore, aimed to ensure a continuous process of education, starting from a solid general foundation in primary school and then moving to a consolidated syllabus in *gimnazjum* and upper secondary, progressing to the school-leaving examination. Another new requirement of the reform for foreign languages was that learners in lower secondary school were given the possibility to follow on with the language they had studied in the earlier educational stage, thus ensuring continuity.

The second change in foreign language education, introduced with the reform, was the implementation of an early start in foreign language, changing the point of introduction of compulsory FL from class 4 to class 1 primary and thus increasing the length of exposure to the FL by three years, with consequent implications for later levels of achievement.

The next reform was lowering the age when the second foreign language was introduced. Previously this had taken place in class one upper secondary, but from 2009 onwards this took place in class one *gimnazjum*.

The next change in the modern languages curriculum was the alignment of its organization, so that it followed the sections set out in the *Common*

European Framework of Reference (CEFR, 2001). Thus the section entitled Contents in the earlier curriculum became Knowledge, Comprehension, Production, Interaction and Mediation, with sub-divisions in each of the latter parts for speech and writing (MEN, 2009: 64–65). Levels of attainment were also overtly stated in terms of CEFR, with A1 set as the level for the end of stage 2 (primary classes 4–6) and a level of A2+ (between A2 and B1) at the end of stage 3 for those continuing the FL in *gimnazjum* (ibid.: 69). Upper secondary targets were set at B1 for the basic level, B2 for the continuation programme and C1 for those following a bilingual course (ibid.: 42).

The final change in the new core curriculum was in the requirements, which are now expressed in operational aims, describing in concrete terms what the learner will be able to do after each educational stage. These subsume what had hitherto been only included in the Test Syllabus. In addition, the 2008 core curriculum also included a list of topic areas, similar to those in CEFR.

To conclude, although the revisions made to the modern languages curriculum were less radical in terms of content than those of other subjects, they introduced important changes in the organization of teaching the foreign language in school. Clear links were indicated between the national core curriculum and CEFR, and the required standards of attainment (previously described only in the Test Syllabus), were now included in the curriculum.

3.9.2. Revision of the external examination system

In 2007 the Ministry of National Education (MEN) published a document setting out the rationale of the proposed changes to the external examination system, which would result from the revision of the core curriculum. The first change in foreign languages was that the new system would include tests in foreign languages at three stages, primary, lower secondary and upper secondary, instead of two as previously. It would also be more coherent, involving similar types of tasks at all three educational stages, from the test of competence at the end of Key Stage 2 (primary class 6), through the lower secondary test at the end of Key Stage 3 (*gimnazjum* class 3, Year 9), to the school-leaving certificate test at the end of Key Stage 4 (upper secondary class 3, Year 12). Tests for the basic levels at Key Stage 2 and at the basic level (III.0) at Key Stage 3 would be composed of closed items only. All learners in lower secondary would take the basic level test in FL and those who were continuing the language from primary school would, in addition, take the extended level test. However, it was planned that until the year 2018 the results from this second part would not be taken into consideration for selection to upper secondary. In 2018 those learners who had commenced obligatory foreign language from class one primary would end stage 3 and

this was intended to be the first cohort for whom the extended level test would be used for selection. Until that time results for the second part of the test were to be used only for the purposes of streaming pupils into groups according to levels of ability in upper secondary school.

The next major change would involve the way in which results of the tests were published. Pupils and their parents would now receive scores given as percentages and in a percentile, as opposed to a raw score in number of points as before. The percentile score would show that the test-taker's result was the same as, or x% better than, all others taking that part of the test. This would give clear information about the individual's level of achievement. Learners would receive separate results for each of the subjects in the test and at each of the levels which they attempted. This aimed to give more detailed information and allow for more effective use of the test results.

Local educational authorities would receive information about each test task, broken down into mean results on each task for each school in their jurisdiction, contrasted with national and regional results and those for smaller administrative districts. Schools would receive similar information, but further broken down to give individualized scores for each learner. The general public would have access to mean results on the whole test on a national scale, as well as mean results at the level of smaller administrative districts.

3.9.3. Changes to the FL examination at the end of *gimnazjum*

The revision of the core curriculum in 2008 caused a need for the introduction of tests at two different levels, assessing the contents of the standards outlined as III.0 (for those commencing the FL in lower secondary) and III.1 (for those continuing the language from primary school.) The fact that this structural change necessitated new examinations also gave the opportunity to make other revisions, taking in the findings of research on the 'old' examinations (eg. Lewkowicz & Lewandowska-Kittell, 2008, 2011; Poszytek, 2008; Smolik 2006, 2008) and recent developments in applied linguistics research (Smolik, 2012: 24). Unlike the earlier version of the FL ELSE, a clear attempt was made to share the rationale of the revised tests with a wider audience, also through the pages of the journal *Języki Obce w szkole*. Taking the view that the changes to the core curriculum can be considered an educational innovation which were to take place over the period 2009–2015, Smolik (2012: 23) places the revision of the external examination system within this process. By means of this innovation, as we have seen, the Ministry of National Education intended to improve the quality of education.

The aims of the revised FL ELSE were now overtly stated as: "diagnosing the skills of individual learners, with the possibility of using the results to place learners in appropriate groups according to language level at the next educational stage" (Smolik 2012: 25). In addition "each of the examinations… should have characteristics which, at least in their intention, will bring about positive washback [pol. *pozytywny efekt zwrotny*] in the teaching and learning process." (ibid.) Citing Bachman and Palmer (1996: 17), Smolik argues that "the most important characteristic of a test is its usefulness", where usefulness is interpreted in the context of how well the test functions for the purpose and in the context for which it was intended (ibid.). (This notion has been discussed more fully in Chapter Two.) Although conscious that the absence of a test of oral skills "may potentially have a negative impact on what happens in school" (2012: 29) Smolik would have us understand that the *gimnazjum* FL examination is only one of a suite of national examinations which serve one million test-takers annually and that consequently practical constraints force a difficult compromise. However, we are to believe that all efforts have been made to ensure the reliability, validity and authenticity of the tests is of the highest order (ibid.: 29–30).

This seems a problematic argument. The stated aim of the test is to diagnose learners' language skills. In the same article we are told that the revised core curriculum (2008) brings together the examination standards of attainment and the curriculum targets in one document. Therefore, if the main goal of Stage 3 is to achieve "effective communication in the foreign language" (MEN, 2009: 42) this implies that what is to be diagnosed by the external examination is the extent to which that goal has been achieved. It is extremely hard to justify that an examination which focuses on comprehension of oral and written text (comprising 60% of the points allocated at the basic level), and tests oral skills in *response* in interaction only indirectly through closed receptive items (where only four points are given to responding to a recorded prompt and the remaining six items are selecting a response to written prompts), is a valid construct of effective communication. Response in interaction is allocated 25% of the possible points. At the extended level no attempt is made to test oral skills indirectly. In the extended test reading and listening are allocated 50% of the points, use of English 25% with the remaining 25% testing written production. In other words, a very small proportion of the combined examination (basic plus extended) indirectly assesses whether a learner can *react* in a 'spoken' interaction and the authenticity of some of these items can be viewed as questionable. In natural conversation we have no time to stop and think about what our interlocutor has said, but have to respond at once. How authentic then is it to give a written prompt and written responses to choose from and consider that these represent "interaction"? Even within the understandable constraints it ought to be possible to include more tasks which require

response to aural prompts. Or failing this, to increase the number of items that test interaction and to include them at both levels of the examination?

In the previous version (2009–2011) the Test Syllabus alleged that 40% of the points were allocated to "responding", thus placing it on a par with reading, while 10% was allocated to listening. Actually some of the "responding" points were for tasks testing "use of lexical-grammatical structures needed for effective communication", but the sample task illustrating this in the Test Syllabus was based on the transcript of a dialogue, arguably lacking in authenticity, but at least making an attempt at face validity. In fact, the test papers in 2009 and 2010 allocated 13 points (26% of the total of 50 points) to indirect speaking tasks, and 12 points (24%) in 2011. What appears to have happened in the revised examination is that "communication" has been divided into two, with the indirect test of 'speaking' in the basic level and the productive test of writing in the extended level, but the total allocation of points (25%) remains similar to the previous version.

What is new in the construct of the revised examination is the inclusion of a test of "use of language", which has apparently been included at the behest of teachers who expressed "frustration resulting from the indifferent attitude learners showed towards those things that are not assessed in the examination", with the inclusion of new tasks attempting to give both teachers and students a more positive attitude towards the test (Smolik, 2012: 26). A further argument given is that "use of language" tasks work well for the purposes of diagnosis, which is seen as the aim of the examination (ibid.).

Within the design of the tests themselves changes were made. With the aim of increasing the scope of what is tested, and attempting to increase reliability by testing skills more comprehensively, a greater number and variety of texts were introduced for the reading and listening parts. Shorter and longer texts are included and the graphic representation of the texts improved to make them look more authentic. Texts such as menus, text messages, postcards, posters, advertisements, leaflets and announcements have been added. In response to the increased importance of digital literacy, tests of finding specific information in longer texts were added, as the ability to search for and select relevant information is seen as a much needed skill.

The provision of tests at two different levels is another change. We have already seen that the allocation of points and the tasks allegedly testing "communication" are different in the two levels. Other differences are that the basic level tasks test receptive knowledge of structures and lexis, while at the extended level tasks also test productive knowledge and cover a wider range of material. At the extended level texts are more varied, more complex and longer, and in the listening test read at a slightly faster tempo.

Thus, we can see that, while some attempts have been made to increase the validity of the examination by widening the construct of reading and listening to include more comprehensive testing of a greater variety of sub-skills in more different ways, the fundamental problems of construct and,

to some extent validity, remain unaddressed. Practical constraints are cited as the reason for this. The result is that the revised examination, with its new section on English in use, is a mainly a test of reading, listening and English in use, with 12.5% of the test given over to indirect testing of oral communication, which has been operationalized as choosing the correct response in a multiple choice task in answer to an oral or written prompt. A corresponding 12.5% is allocated to an open productive writing task. This task appears to be an attempt to widen the test content and decrease construct under-representation. It is now only spoken production in the core curriculum that is not covered at all.

This chapter has traced the history of the external examinations at the end of lower secondary school (ELSE), placing them in a wider socio-economic and educational context. The origin and creation of the first FL component of the ELSE has been described, together with an attempt to define its rationale and construct. The curriculum reform of 2008 is explained and the consequent changes made to the FL component of the ELSE are detailed. A brief critical analysis of the revised examination is given.

This chapter ends the theoretical part of this work and we will now move on to investigate the role of the external examination in English in lower secondary schools in Part Two.

CHAPTER 4

Investigating washback and impact from a national language examination at the end of lower secondary school

4.1. Rationale for the research

As we saw in Chapter Three, Poland underwent an educational reform in 1999 which changed the eight-year primary, four-year upper secondary system to include a new level, the lower secondary school, *gimnazjum*. Learners entered this school after six years of primary, stayed for three years and moved on to a choice of three-year upper secondary schools. We also saw how Poland underwent a change in foreign language policy, and how in response to EU agreements, changed the point at which the first foreign language began in primary school, and lowered the age at which the second foreign language was introduced. A new core curriculum was introduced in 1999 to coincide with the creation of *gimnazja*, and new national examinations at the end of the lower secondary stage in 2002, as the first cohort completed the new school. In 2009 foreign languages were added to the subjects learners had to take in these examinations.

The decision to introduce a mandatory foreign language examination in Year 9 (learners aged 15–16) was a signal of the importance of language learning. Poland's entry to the EU in 2004 had increased mobility at an unprecedented level, with large numbers moving to Britain for work (as Britain was one of the few EU countries which placed no restriction on Polish migration at this time). Knowledge of foreign languages, and of English in particular, could therefore open new horizons and offer new possibilities.

This climate of change created many exciting opportunities for research. A new national examination, to be taken by all in mainstream schools (with different versions for learners with special needs) was a potentially powerful agent for change. The questions were whether a change would come about, and if it did, what form would this take? From this seed of an idea grew the study of impact which is the subject of this work.

Aside from curiosity about the Polish context, it seemed that there was also a chance to discover more about the ways an examination effects change

and the mechanisms this involves. This study responded to McNamara's (1996) claim that "high priority needs to be given to the collection of evidence about the intended and unintended effects of assessments on the ways teacher and students spend their time and think about the goals of education" (ibid.: 22). It also anticipated Cheng's (2008) belief that future washback and impact studies should be "multiphase, multimethod and longitudinal in nature" (ibid.: 35), and the call of Cheng et al. (2015) for investigation of washback in context, from the perspectives of different stakeholders.

Of particular interest was the hypothesis posed by Alderson & Wall (1993: 121) which suggests that only some teachers will be affected by washback. This influenced the decision to collect and analyze qualitative data from teachers and search for indications in their responses which might explain why some are affected by washback, while others are not. Later Alderson (2004: ix) posed further questions: "…what does washback look like? What brings washback about? Why does washback exist?" Within the data from teachers would it be possible to find answers to these questions? In the same text, Alderson (2004: xi) goes on to stress the importance of understanding teachers' "*beliefs about teaching and learning*, the degree of their professionalism, the adequacy of their training *and of their understanding of the nature of and rationale for the test*" (italics added). This statement influenced the direction of the research and design of some of the instruments. It was hoped that by examining the data we would come to understand more about these, as yet not fully explained, phenomena.

Significance of this research

The research described in this part of the work contributes to the literature on washback and impact in several ways. First, unlike many studies conducted to date (e.g. Green, 2007; Saville, 2009; Wall & Horák, 2008, 2009, 2011), this is research on a mandatory national examination in state school. Secondly, unlike many other school-based studies which focus on state school school-leaving exams (e.g. Ferman, 2004; Gosa, 2004; Shohamy et al., 1996; Wall & Alderson, 1993), this research is on a national exam at the end of lower secondary school (currently Year 9). It is also an exam which is used for selection purposes, but for the choice of upper secondary school, and not university. The learner respondents are younger than those in many studies, between 13 and 16, depending of the phase of the study.

The next aspect of this study which is different from many others carried out to date, is that it is a longitudinal study, tracing the changes brought about over a period of five years in an education system. It should be stressed, however, that this is not a panel study which follows the same participants over that period, but three separate studies on different populations in the same context.

The final contribution which this research has to make is due to the size and nature of the sample from which data were drawn for Phase 3, the impact study, which forms the most important part of this work. Data was collected from three different groups of stakeholders potentially affected by the new examination: school principals, teachers and learners. The sample of school principals (115) is representative for lower secondary schools in Poland, and within these same schools teachers (301) were interviewed. The greatest emphasis has been placed on analysis of qualitative data obtained from interviews. The reason for this is to attempt to discover more about the washback process: "It is high time for more empirical studies involving all stakeholders in washback to explore how washback operates and functions as a comprehensive and interwoven entity…"(Cheng et al., 2015: 470). Most studies to date focus on a small number of teachers (e.g. Burrows (2004), 30 teachers interviewed and 4 teachers observed; Watanabe (2004), 5 teachers; Wall & Horák (2006), 10 teachers). The abundance of data available allows much greater insight into the processes at play which appear to determine *if* a teacher will be influenced by the exam, *why* they may be influenced, *how* they claim to be influenced and *to what extent.* This enables us to see more clearly trends and patterns which emerge, to offer verification of some hypotheses tentatively proposed by earlier authors (e.g. Watanabe, 2004; Green, 2007) and to contribute new hypotheses.

4.1.1. Selecting the research design

In this section we consider the factors which influenced the choice of research design and instrumentation.

Research methods used in studies of washback and impact

In the chapters on washback and impact we have seen that these are highly complex phenomena which take place dynamically over time. The key issue in any study of washback or impact is to find evidence that an examination is in fact causing changes which can, without question, be attributed to the introduction of the examination and not to other causes (Messick, 1996: 247). It is the challenge of establishing this evidential link (Alderson & Wall, 1993: 117) that makes the design of such studies difficult. In order to be able to state with any conviction that change has taken place, it is therefore necessary to start by investigating the situation before the new examination is introduced, so as to provide a point of comparison (Weir & Roberts, 1994: 46). There is, therefore, a consensus that first a *baseline study* should be conducted which provides a description of the context prior to the first test administration. This is a procedure followed in many influential washback studies (e.g. Wall & Alderson, 1993; Cheng, 1998; Wall & Horák, 2006).

Detailed analysis of the new test and documents supporting it are needed, in order to ascertain the purpose of the introduction of the examination and any agenda which the authorities have connected with its implementation. We have seen in Chapter 2 that some educational bodies intend new examinations as levers which they hope will effect desired outcomes, and so supporting documents need to be scrutinized with this in mind. Careful attention needs to be paid to the content of the new test, the types of task it includes, the ways in which these tasks will be assessed, according to what criteria, and the procedures the test administration will follow. Any, or all, of these can be expected to potentially be a cause of change in teaching, or learning. The examination also needs to be analyzed in relation to the mandatory curriculum, if one exists, as we have seen that mismatch between curriculum and examination content, or purpose, can cause changes in course planning, content or focus. We undertook such an investigation into the national external examination in foreign languages for lower secondary school in Poland in Chapter Three, concentrating on the examination for English for those continuing the language from primary school.

Expectations of evidence of washback or impact

In the review of literature on washback we saw that many different aspects of teaching and learning may be affected by the introduction of a new test, and that these include participants, processes and products (Hughes, 1993: 2). In order to focus a research study of washback, Alderson & Wall (1993: 127) state that

> it will be necessary to spell out in some detail what the predicted effects of the test are, and it is highly likely that this statement will have to take into account the nature of the test concerned, the educational context within which it is used, and the nature of the decisions that are made on the basis of the test results.

In the studies of impact reviewed we saw that Henrichsen's (1989) model of diffusion/implementation in ELT recommends considering what aspects of the process will be affected by the innovation, in order to plan the impact study. In Social Impact Assessment (SIA) (Vanclay et al., 2015), the initial phase includes predicting, analysing and assessing the ways in which the project may have impact, and taking pro-active measures to reduce possible negative outcomes. Thus, both methodology from both washback and impact indicate we should predict what may happen as a result of the introduction of a new examination. Then, by comparing what is expected to be found with the data obtained, we can draw conclusions about whether washback, or impact, is in fact taking place, and if so, in what areas, and to what extent. It was not possible for the writer to implement any pro-active measures before the introduction of the 2009 examination in Poland, as suggested by SIA, but it is to be hoped that the findings of this research may

suggest "prophylactic" action which could be taken before the introduction of other new examinations, which could go some way to reducing possible negative consequences.

One of the difficulties inherent in washback studies is how to be clear that any change noted is in fact change which is caused by the examination, and not by other factors. In the Polish context, when several major changes were happening at once, both at a societal level and with school and curriculum reform, this was a particular challenge. Watanabe's question (2004: 28) was found especially useful in this respect: "would teaching/learning become different if there were no exams?" and helped focus the research. A second question "How will we know change has taken place? What will we see?" was asked time and again during the design of the research.

4.1.2. Factors influencing the research design

The project was conceived in stages and was subject to practical constraints, which played a decisive role in determining the choice of methods and instruments. Phases I and II were conducted by the author, without funding and without the status of being a university researcher. Employed in a foreign language teacher training college, I was simply a senior teacher with an academic title, known to a large number of school teachers and principals through supervision of undergraduates on teaching practice, or through in-service workshops for teachers under the national INSETT programme. Being able to conduct research in schools, consequently depended a great deal on personal contacts and good will. It was therefore easier at the early stage of the project, in the baseline study, to start from investigating teachers, as this was where contacts were most open. In Phase III circumstances changed (as we will see below), and all the constraints present in the earlier parts of the study no longer applied. This meant there was much greater scope for verification of data from other respondents both in questionnaires and interviews. Here was a new concern, however, which was to ensure comparability of the data, which was gathered, not by the researcher, but by a large number of different people. Consequently, in the design, strong measures were put in place to ensure reliability, by establishing rigorous procedures, detailed instructions for application of instruments and standardisation sessions to train all those involved.

From a practical point of view learners in the Polish context are the most difficult respondents to access. Time during a language lesson spent completing a questionnaire has to be made up at another time, meaning schools are not willing to agree to this. As a result, access to learners is only possible during the hour that they have with their class teacher, which is logistically difficult to arrange. An outsider cannot enter a school without permission from the principal and cannot be left unaccompanied in the presence of young people. For learners who are below the age of majority,

written permission to take part in the study must be obtained for each learner from their parent or guardian.

Whenever young people are involved it is also important to consider the potential effect taking part in the research may have on the participant. If we are investigating learning in school, and in particular among younger respondents, then there are ethical and professional responsibilities to be considered. Asking learners to reflect on their learning and the teaching they experience promotes reflection which may lead them to question the status quo. This has implications for design of the instruments and procedures.

Obtaining information from teachers is challenging. They are very busy people and, while they will readily complete closed questions in a questionnaire, they tend to baulk at filling in longer open responses. Appearance is key, if it looks long and complicated, a survey is less likely to be returned. As a result, although allowing access to a large number of participants quickly, the kind of information which can be obtained from a teacher questionnaire tends to be limited. While an open question will potentially reveal more, it becomes necessary to offer a series of options from which the respondent may select, in order to obtain a response at all. This means that the instrument is potentially biased, as the options on offer are those the researcher predicts may be possible answers, when in fact the truth may lie elsewhere. This reminds us of the need to consult draft versions of instruments with individuals who belong to the target group, but will not be part of the final sample. Their comments provide valuable insights and improve the quality of the final instrument.

Planning for reliability

Asking teachers to complete questionnaires about what they do is fraught with difficulty, due to the Hawthorne Effect. Teachers have a tendency to respond as they think the researcher would like them to respond, or as they think they ought to respond. They are well-aware that certain practices are viewed as more desirable than others and so may suppress information if they feel that their behaviour runs counter to these. Depending exclusively on information obtained from a questionnaire, particularly in the case of teachers, is therefore unwise. Consequently, the information needs to be verified in some way. The choices are to interview a series of teachers who completed the questionnaires and ask them similar questions, to check if the answers correspond to what was given in the questionnaire (checking for concurrent validity, by measuring the same phenomenon in two ways and comparing them); or to ask similar questions of another group of respondents who can verify that what the teachers said is really the case. Thus, if we are interested to find out what teachers do in the classroom, we can ask them and we can also ask their learners the same question and check for correspondence. Ideally this is done by checking the same question from

three perspectives, or by three measures, known as triangulation. A third option is to observe lessons taught by some of the teachers who completed the questionnaire to see if what is claimed there is what in fact is taking place. In this way we see that the pragmatic, realistic stance leads us to adopt mixed methods in the research design.

Classroom observation, while potentially offering rich data, is also not without practical difficulty. Many teachers do not like being observed and will not agree to it. If they do agree, there is a danger that what is seen is a 'showpiece' lesson, which does not resemble their everyday teaching. As a result, the researcher needs to see more than one lesson taught by a teacher and preferably to offer them the opportunity to choose which groups are observed, so that they feel as comfortable as possible with a situation which is, undoubtedly, stressful. One final aspect in the Polish school is that teachers have a detailed plan, the scheme of work [pol. *rozkład materiału*], which determines what they will do in each lesson. It is planned a year in advance and allows for very little flexibility. If there is a certain aspect of teaching we would like to see, then we need to ask teachers when that will be possible. Again, logistic constraints come into play, as this might entail multiple visits to the school.

Issues concerning qualitative data

One of the instruments used in Phase III was the interview. This was used to collect rich data, with fuller and more detailed answers than could be obtained from questionnaires. This type of information was felt to be more likely to offer explanations of *why* certain actions were taken and also, possibly, to explain *relationships* between decisions or actions. We shall see that this type of information is key in exploring the mechanisms involved in washback and impact. Interview data was also, as indicated above, used as a measure to verify information obtained from other respondents, the same group of respondents, or that respondent themselves. Data obtained from interview is, however, not without problems, as we will see in the section below.

Interview data

Kvale (1996) points to two different views of data obtained from interviews, the *veridical* view, where what is said is interpreted as a true statement of the interviewee's opinion or belief and thus taken as reliable data, as opposed to the *symptomatic* view, where the data are seen as revealing the interviewee's stance on the topic in their particular context. From this perspective interview data may be found to contradict data from other sources, as they reveal the individual perspective and interpretation of a topic as viewed by this person through their own eyes. Rather than seeing this as invalidity, the symptomatic view uses the data to obtain a holistic picture of the relationships of the interviewee with others, or about their feelings and experiences (Block, 2000). Freeman (1996) echoes this

dichotomy, seeing data from interviews as either *representational* of facts or events, or *presentational*, as in revealing information about the interviewee as a person.

Interviews are interactions between the interviewer and the interviewee, where the interviewee responds to questions posed to them. Consequently, prevailing views see the response as co-constructed (Cicourel, 1964; Silverman, 1973; Holstein & Gubrium, 2003; Talmy, 2010) and so it may represent views different to those the interviewee might state independently of the interview situation.

Block (2000), drawing on Bahktin's idea of voice (1981), Lemke's concept of speaking as a member of a community (1996), and Gee's views of the individual conforming to a discourse they view as socially acceptable (1995), explains that the interviewee speaks with different voices within one interview, representing not only their own personal views, but those they perceive expected of them in their role as a member of a discourse community. In addition, they may speak from different perspectives within the same interview, as each of us simultaneously holds different roles (Holsten & Gubrium, 1995). Thus, the teacher, asked about aspects teaching in his/her school, may answer from the perspective of the teacher of class A (a cohesive high-performing group), or from the different perspective of teacher of class B (a problematic group with a high proportion of challenged learners), or may step back and respond as a member of the modern languages team, or may speak as a dissatisfied employee who has a poor relationship with the school head.

Potter & Hepburn (2011/2012) describe challenges which exist in the "design, conduct, analysis and reporting of qualitative interviews" (ibid.: 3), some of which refer to the way in which the interview is reported and others to the way in which the data is analysed.

In order to avoid common problems in reporting they make the following recommendations:

a) Clearly explain the context of the interview. This refers to the decisions behind the choice of interviewee, the information given to interviewees about the purpose and content of the interview and what they are asked to talk about.
b) Make clear the role of the interviewer. Data quoted from the interview should be given in context and if in fragments, this should be explained.
c) Report data verbatim, including all features of speech, rather than as a transcript from which features such as hesitation, false starts or repetition have been edited.
d) Refer to specific features of the interview in the analysis. In this way over-generalization, or under-analysis (Antaki et al., 2007) can be avoided.

Among challenges facing the researcher in analysis Potter & Hepburn (ibid.) refer to the problem of 'footing'. Goffman (1981) writes of speakers in interviews taking different 'footings' in how they relate to the topic and

each other. Thus, one interviewer may appear disinterested, objectively collecting data, while another may give verbal or non-verbal signals to suggest their reactions to what is said. Where these signals may be interpreted by the interviewee as suggesting approbation, this may influence what the interviewee says. It may also affect the length of the response, if the interviewee feels they are being encouraged.

Care should be taken in the way in which questions for interview are constructed, particularly in whether the interviewee is addressed directly about their own opinion about their own context, or whether they are asked to comment on a general statement about a phenomena. If the statements are non-specific, then extreme care must be taken about using this data to draw conclusions about the belief system of the individual in question (Potter & Hepburn, 2011: 28–30)

In interpreting qualitative data the researcher should make every attempt to avoid bias. "As researchers, we need to be aware that our previous training, experiences, and attitudes all contribute to the way we view the events we observe." (Allwright & Bailey, 1991: 74). Thus, it is advised that researcher make clear their "interests in the study, contexts of research and the implications for the interpretations of the data" (Luttrell, 2010a in Roulston & Shelton, 2015: 7). In analysis of qualitative data we should take care not to fall into the trap of confirmation bias, (Nickerson, 1998) whereby we extract from the data only information which supports the viewpoint we have already decided we wish to express.

This short overview highlights some of the issues which were taken into consideration in planning the design of the interviews, the procedures for conduct of interviewers and the analysis which was carried out.

In sum, when considering the design of the study, substantial thought was given to the logistics of how it would be conducted, in order to obtain informative and reliable data and yet cause minimal disruption within the schools which agreed to take part.

4.2. Design of the study

In the next section we set out a description of the project which was conducted.

Main aims of the research

The research was designed in order to investigate the impact of the introduction of a new mandatory national examination in foreign language on the teaching and learning of English in lower secondary schools (*gimnazja*) in Poland.

Research question: Will the introduction of a national examination in English have an effect on the teaching and learning of English in lower secondary school in Poland? If yes, what?

Conceiving the research design

Recognizing that any effects caused by the new examination would take place gradually, over time, this was conceived from the outset as a longitudinal study. Initially it was intended that the research would last about 4 years, commencing with a baseline study in the school year 2007–2008, prior to the first administration of the new examination in spring 2009, continuing in cycles to examine the effects immediately following the new examination in 2009 and then again later in 2010 and 2011. The first design was for the author to conduct the study individually.

Initial research design

PHASE I		PHASE II	FURTHER PHASES
May 2008	April 2009	May 2009	May 2010, May 2011
Baseline study	First exam	Initial washback study	Iterative studies planned

Phase I the Baseline study was conducted and Phase II the Initial Washback study immediately following the first test administration, were carried out as planned. These data were analyzed and reported in two papers, published in 2009 (see Ellis, 2009a, 2009b). However, at this point a key change took place in the author's circumstances which allowed the use of data from a much larger study.

As a result a change was made to the initial design of the impact project.

PHASE I		PHASE II	PHASE III
May 2008	April 2009	May 2009	March-May 2012
Baseline study	First exam	Initial washback study	Impact study

Data for Phase III of this study come from the first year of *Badanie Uczenia się i Nauczania Jezyków Obcych* (BUNJO) [Teaching and Learning Foreign Languages in lower secondary school], conducted throughout Poland in 2012. The research was undertaken by the Foreign Language (FL) section of the Educational Research Institute (ERI) in Warsaw, where the author was working at the time. Funding for the project was obtained under the European Social Fund Operational Programme Human Capital, Priority III- High Quality of Education, within the grant "Quality and effectiveness of education-strengthening of institutional research capacities" Section 3.1.1. Creating conditions and tools for monitoring, evaluating and researching the educational system. The research project was conceived and designed by the author, with some development and revision made by members of the FL section team, led by the late dr Magdalena Szpotowicz.

Year One of the BUNJO study had the specific aim of providing a full and detailed description of the process of teaching and learning English

in the context of lower secondary schools, with the aim of identifying factors which appeared to contribute to differential levels of achievement. Many factors were anticipated as potentially having an effect on teaching and learning, one of which (in the light of research already done by the author in Phases I and II of this study) was predicted to be the national foreign language examination. Thus, while use of the BUNJO data has the benefit of extending the scope of the current research (more details will be given below), the fact that BUNJO had a different overall aim is a distinct drawback in terms of the design of a study of the impact of the foreign language examination.

Changes in 2012

Another significant event, however, also took place in 2012, which was not predicted in the initial research design. Substantial changes were made to the form of the examination. From being one examination consisting of three papers, it changed to a two-tier examination, at basic and extended levels, with all learners continuing English from primary school being required to take both levels.

2009	2012
New FL exam in lower secondary	FL exam revised: 2 levels Basic and Extended

The extended level now included an open writing task (further information on changes to the exam were given in Chapter Three). At this point the question posed by Watanabe (2004: 29) was pertinent: "would teaching/learning become different if the exams were to change?" In effect, it could be argued that the whole research design was compromised by these changes and that the impact study should have been extended beyond 2012. This was, unfortunately, not feasible. An alternative is to view the three phases of data collection as having taken place at critical moments in the history of the examination: prior to its implementation in 2008, immediately following its first implementation in 2009 and at the time of, or following closely, the implementation of the revised examination in 2012. The 2012 examination was doubly significant, as this was the first administration when it became officially "high-stakes", with points from the basic level FL exam counting towards selection for upper secondary.

Predicting what changes the new exam might bring about

The project was conceived a longitudinal study. It was understood that washback is a complex and dynamic phenomenon and consequently the project was planned to be organic and responsive, with each subsequent phase growing out of the previous one(s), building on what had been discovered. It was anticipated on the basis of other studies (e.g. Cheng, 2005) that washback would take place and that it would bring about change.

First, however, it was necessary to predict what these changes might be. This needed to be done ahead of the baseline study to ensure that data on focal areas would be available for comparison once the examination had come into effect. However, in the event that other unforeseen consequences might come about, the baseline also needed to obtain enough information to draw up a detailed and broad picture.

Predictions about what areas could be affected by the exam were made based on Wall & Alderson's (1993) hypotheses of what could be affected by an examination, together with the author's extensive experience of observation of lessons in lower secondary schools as part of her work as supervisor of teaching practice of undergraduates. As reading and listening comprehension were to feature strongly in the new exam these were selected as initial areas of focus in the baseline study. The rationale was that if changes in teaching were to take place, then they might be most noticeable in these areas. Assessment practice was also selected, as it could be predicted that with reading and listening being important in the exam, then these might become more frequently assessed, or differently assessed after its introduction. The role of the course book was selected for focus, as previous studies of washback (e.g. Cheng, 2005) had found changes took place in teaching materials. Teachers would be asked what aspects of language were important in teaching and learning as their priorities might be affected by the examination.

Phase II took place one year later, very shortly after the first administration of the examination. The participants were learners who had taken the exam, their teachers, and other teachers who had taught similar learners in other schools. The focus was mainly retrospective and evaluative, asking for information and opinions about the examination.

Phase III took place three years after Phase II. In its design first the initial picture of the teaching/learning process which had been drawn up in the baseline study and information obtained from the Initial Washback Study was reviewed. Next, following the hypotheses of Alderson & Wall (1993) we predicted again what washback in the Polish classroom might involve and what could constitute evidence of this. The following predictions were made:

1. teachers would change the teaching material used for materials designed to prepare learners for the new examination. This would be evidenced by teachers' justification of the choice of course book and explanation of its role in the planning of their work.
2. teachers, in statements about their aims for teaching English in *gimnazjum*, would highlight the importance of the role of the examination in their teaching process.
3. teachers, when asked to explain their planning process would place high importance on the role of the examination.

4. teachers may express concerns over tensions between the core curriculum and the examination content. This is because speaking is not tested directly during the examination. This would be evidenced in statements about what is most important in teaching/learning at this educational stage, or in the teacher's aims for the course.
5. reading, listening, speaking and writing would be assessed in the classroom to reflect their role in the examination. This would be evidenced by grades being given for the different skills, and frequency of tests of skills, obtained from teacher/learner statements.
6. there would be frequent use of materials designed to prepare for the exam. This would be evidenced by teacher/learner statements.
7. learners in class one would have been given information about the tests, and how they are scored, indicating that the effects of the exam impact on all classes, not just the final class. This would be evidenced by statements from learners.
8. learners would view the FL examination as important for their futures. This would be evidenced by statements from learners.

Having drawn up the list of expected effects at the level of teachers, learners and the classroom, and how these would be operationalized, we then considered the wider context of the school and predicted what effects could be expected there.

Expectations of evidence of impact at school level

At the level of the school it was expected that it would be the *results* of the examination, that would be seen to have effect, rather than the content, or form of the examination.

The examination will be deemed to be having an impact at school level if:

1. the results are found to play a role in planning work in school. If this is the case the school principal will be able to articulate what role the examination results play in planning.
2. the results are found to effect changes in the organization of FL teaching in the school. This may be evidenced by actions such as an increase in the number of hours allocated, organization of groups (e.g. introduction of streaming), or allocation of resources (e.g. a dedicated classroom, new equipment).
3. the results are found to play a role in the internal evaluation process of the quality of work in school. If the FL examination is having an impact the school principal will be able to articulate what role the examination results play in monitoring work in school.
4. the results are found to play a role in evaluation of the effectiveness of teaching and learning. The role of the examination may be revealed by asking the principal to describe the ways in which foreign language teaching is evaluated in their school.

In each case, the examination can only be judged to have brought about these changes if the principal explicitly attributes these changes to it.

4.2.1. Aims and focus of each phase of the project

In this section the aims of each phase of the project are set out and the research questions listed.

Phase I the Baseline Study

Main aim

To obtain a picture of how English is taught in lower secondary school, with particular focus on the development of the skills of reading and listening.

Research questions

In lower secondary school:

1. What is the role of the course book in teaching?
2. Which aspects of language are the focus of teaching and learning?
3. How are reading and listening taught?
4. How frequently are reading and listening practised?
5. What is assessed?
6. What is the teachers' knowledge of and opinion on the new exam?

Phase II the Initial Washback Study

Main aims

To see if the introduction of a national examination in English had had an effect on the teaching and learning of English among teachers and learners in class 3 of selected lower secondary schools in Poland.

Research Questions

1. What aspects of language do learners consider important in learning English in *gimnazjum*? How does this compare with information from the baseline study?
2. According to learners, how frequently are reading and listening practised? How does this compare with information from the baseline study?
3. According to learners, what is assessed? How does this compare with information from the baseline study?
4. Is the exam important for the learners?
5. What (if any) changes did teachers make to their teaching in class 3 because of the exam?
6. What, according to teachers, is the learners' attitude to the exam?
7. What effects did teachers notice in their learners' behaviour which they attribute to the exam?
8. What is the teacher's attitude to the exam?

Phase III the Impact Study

Main aims

To examine for effects of the examination at two levels, the level of the classroom and the level of the school. At the classroom level it was assumed, on the basis of the Initial Washback Study, that effects of the examination would be seen. The primary aim therefore was to seek reasons why washback was taking effect and what caused it to come into effect. At the school level the aim was to find out whether the results of the FL exam were causing any effects, and if so, to look for reasons why this was the case and how impact from the exam came about.

Research questions

1. Is choice of teaching materials affected by the exam? If yes, in what ways? What reasons do teachers give for this?
2. What is the attitude of teachers as regards the exam in their teaching? What reasons do they give for this?
3. How do teachers view the relationship between the exam and the core curriculum?
4. Which aspects of language do teachers consider important in teaching and learning in *gimnazjum?*
5. Are all four skills assessed with similar frequency? If not, what reasons do teachers give for this? Is there evidence that this results from the exam?
6. Does the exam affect the planning of teachers' work? If yes, in what ways? What reasons do teachers give for this?
7. Which aspects of language do learners consider important in foreign language learning in *gimnazjum?*
8. According to learners are all four skills assessed with similar frequency?
9. What are learner attitudes towards the exam? What reasons do they give for this?
10. What do learners in class one *gimnazjum* know about the exam?
11. What role does the examination play in planning the work of the school?
12. What role does the examination play in evaluating the quality, or effectiveness of English teaching in lower secondary school?
13. Is there evidence of changes in the organization of English in lower secondary school which can be attributed to the exam or the exam results?

4.3. Method

In this section first participants who took part in the study and the process by which they were selected are described. Next, the instruments which were used in collecting the data are detailed.

4.3.1. Participants included in the project

At the outset, in the initial design, the participants were to be teachers and learners in lower secondary school. The reason for this was that the first conception was to examine effects on teaching and learning in the classroom, as a washback study. The wider scope of the BUNJO project allowed the area of interest in Phase III to be broadened and the concept of the study of the effect of the examination also. From being solely concerned with the classroom, it was extended to look at the wider context and became a study of the impact of the examination. For this reason the participants in the main study were extended to include school principals. Thus, across the three phases of the study different subjects were the focus of attention, as can be seen in the table below.

Table 4.1. Participants in different phases of the study

Year	Phase	Title	Teachers	Learners	Principals
2008	I	Baseline study	X		
2009	II	Initial washback study	X	X	
2012	III	Impact study	X	X	X

4.3.2. Sampling

Different approaches to selection of the sample were taken in each of the phases of the study. Phase I the Baseline Study aimed to include teachers from schools in a variety of locations, urban and rural, in a convenience sample. Schools for observation were selected because of contacts with teachers who were willing to be included. Phase II was planned initially as a cluster sample, including all the lower secondary schools in one city in the south of Poland, but constraints led to it becoming a convenience sample. Phase III, a large-scale funded project, allowed a random sample of schools, representative for Poland, to be selected. The sampling is summarised in Table 4.2 next to.

4.3.2.1. Description of participants in the study

Phase I: Baseline study

45 teachers of English, in a convenience sample of twenty five schools, from four cities and two villages in Silesia, completed a questionnaire. (Silesia is a voivodeship in the south of Poland with a population of 4.63 million in 2012). Information about gender and age was not collected.

Observations were carried out by the author in three selected schools in two cities in Silesia. Seven teachers (5 female and 2 male) were observed, teaching a total of fourteen lessons.

Table 4.2. Sampling in the three phases of the study

Phase	Type of sample	Number of schools	Participants	Gender, age
Phase I: Baseline study	Convenience	25	45 teachers 7 teachers observed in 3 schools (14 lessons)	–
Phase II: Initial washback study	Convenience	7 (learners) 7 + 11 (teachers)	232 learners 21 teachers (10 matched to learners, 11 volunteer)	Class 3
Phase III: Impact study	Random. Representative at level of school. Groups of learners nominated by school principal. Teachers of these groups. 4 learners for interview selected randomly from each school. All teachers interviewed	120	Learners 4343 questionnaires 480 interviews Teachers 380 questionnaires 301 interviews Principals 115 interviews	Learners: questionnaire F: 53.25% M: 46.75% Age: Class one 14 (95%) 13 (1%) 15–16 (4%) Learners: interview F: 56% M: 44% Teachers F: 86% M: 14% Age: 23–31 (34%) 32–40 (47%) 41–60 (19%) Principals F: 64% M: 36% Age: 31–35 (2%) 36–40 (6%) 41–45 (17%) 46–50 (27%) 51–55 (35%) 56–60 (12%) Above 60 (1%)

Phase II: Initial washback study

A convenience sample of 232 learners in class three, from seven lower secondary schools in two cities in Silesia was selected. The study was intended to include a larger number of schools, but it was found at the time of collection of data in May that in many cases class three no longer

attended school. 10 questionnaires were collected from their teachers. The resulting sample is uneven in terms of the number of learner respondents in each school. As the sample of teachers was so small, it was decided to supplement it, and additional questionnaires were completed by 11 teachers of class three *gimnazjum,* during workshops held by publishers in August, making a total of 21 teachers. Information about gender and age was not collected.

Phase III: Impact Study

The sample of schools was randomly selected from a national data base of public lower secondary schools in Poland from which special schools and schools located in other institutions, such as hospitals or remand centres, were excluded. For inclusion in the research, schools needed to offer English as a continuation from primary school and to have at least 50 pupils in year one (school year 2011/2012). This yielded a representative sample of 120 schools located throughout the country.

In each school the principal was asked to nominate three first year groups, continuing English from primary school, to take part in the study. As far as possible, these groups were of different levels of ability and taught by different teachers. This was the case for 105 schools. Some were small schools and had fewer groups, 14 schools had 2 groups and one school only one group. The final sample included 344 groups, and contained a final total of 4717 learners. Only learners whose parents, or guardians, had given written permission took part in the study. Learners with special needs, or a level of language which prevented them from being able to write the questionnaire, were excluded from the sample. 4343 took part in the questionnaire study. Four learners from each school were selected at random from the target groups, making a sample of 480 for the learner interviews. For the teacher questionnaire study, all teachers of English in the 120 schools were to take part (382), but data was collected from 380. The teachers who taught English to the selected groups formed the sample of teachers who were interviewed. This was 307 in total, but 301 interviews took place. In each school either the main director, or one of the deputy directors, was to be interviewed, but the final sample was 115.

4.3.3. Instruments

In this section instruments used in each phase of the project are described.

4.3.3.1. Instruments used in Phase I The Baseline Study

The baseline study included a questionnaire for teachers and classroom observations.

The questionnaire

The questionnaire was designed to find information about the teacher's beliefs about what aspects of language are important in teaching in lower secondary school and what aspects teachers think are important for their learners. It included 20 questions which asked about the course books and other materials used by teachers, and asked them to estimate the percentage of course time spent on the course book and accompanying materials (e.g. students' practice book, exercises from the Teacher's Book). Sections were included which investigated the teachers' approach to teaching reading and the frequency reading is the focus of attention in class, and a parallel section on the teaching of listening. Teachers were also asked about assessment, including a question about what aspects of language they give grades for and the sources of tests that they use. The final area of interest was the new examination. Teachers were asked if they were familiar with it, whether they had done any practice tests with their learners and were also requested to give their opinion on the difficulty of the new exam for their learners.

Teachers were also asked to give information about their qualifications, length of teaching experience, level on the scale of educational promotion [pol. *awans zawodowy*] and their opinion on teaching English in lower secondary school.

Question types included ranking, selecting from lists, nominal scales, and yes/no questions. Where respondents were asked to select from lists an option "Other? What?" was included. Some of the Yes/No question included an open follow up question. The questionnaire was given in English (see Appendix 1).

Observation

Lesson observations were unstructured and no pre-conceived observation scheme was used. The reason for this was that there was no specific focus, rather to obtain an overall impression of what might be considered a "typical" lesson at this educational stage. Reading and listening tasks were considered of key interest, but not exclusively. The author took detailed ethnographic notes which provide a chronological record of the lesson. They detail times, interaction patterns and materials used. Notes are made of what the teacher did and said and how the learners responded, both verbally and non-verbally. After the lesson a brief discussion was conducted with the teacher, when questions were asked about critical incidents or individual learners.

4.3.3.2. Instruments used in Phase II the Initial Washback study

The Initial Washback study included two questionnaires, one for learners in class three lower secondary who had recently taken the new examination and the second for teachers.

Learner questionnaires

The learner questionnaire was divided into three areas as follows: a) general information about the learner's attitude to, and aptitude for, learning English; b) the learner's perception of the frequency of occurrence of different types of activities in English lessons and of the assessment process; c) the learner's perception of the external examination in English. It was given in Polish to ensure the respondents would understand (see appendix 2, English versions of the questions are given in the report of the results in Chapter 5). These questions were planned to give information about teaching from the learner perspective, and through comparison with results from the baseline study on the frequency of occurrence of activities, look for possible indications of change; information about assessment was intended to check for consistency between what was being taught and what was being tested; and finally information about learner attitudes to learning English and motivation, aimed to look for potential positive, or negative washback from the examination.

Teacher questionnaire

Questions in the teacher questionnaire asked directly if they had changed their teaching as a result of the new examination and, if so, in what way. A second area asked about the teacher's perceptions of the effects of the examination during the year on different groups of learners they taught, based on their observations of learner behaviour. Finally, teachers were asked to give their reaction to the examination itself. In this way it was planned that some insights might be gained into whether teachers had changed the contents of their teaching, or how they taught, based on what they claim to have changed.

4.3.3.3. Instruments used in Phase III the Impact Study

Instruments used to collect data used in the Impact Study included:

- interviews conducted with learners, teachers, and school principals
- learner questionnaires
- teacher questionnaires
- field notes kept by the author during the piloting of instruments for this phase

As mentioned earlier, there were constraints on the design of Phase III of the study. The aims of the Impact Study are quite different from those of the first part of the BUNJO project. As a result, the instruments used in BUNJO were not designed exclusively to identify washback from, or impact of the FL exam, but had elements included in them which had this aim. As only small parts of BUNJO questionnaires and sections of interviews were used for this study the instruments are not included in the appendices. The parts used are described fully here. All instruments were in Polish.

Teacher Interviews

Only one section (two questions) in the BUNJO teacher interview refers to the exam directly. The whole interview covered six areas: course books used for teaching English in the school; focal areas for learners and for teachers in learning/teaching English in lower secondary school; ways in which level and effectiveness of teaching can be raised; individualization; the assessment process in year one; and the role of the FL *gimnazjum* exam in planning and the choice of materials for teaching English.

The questions directly relating to the exam were:

1. What role does the foreign language exam play in the planning and choice of materials in teaching English in your school?
2. Does this affect class three only, or all other classes too?

Teachers were also asked about the course books they used in *gimnazjum*, their opinion about them and the role the book plays in the teacher's planning in class one. Another section referring to the assessment system used in class one, asked teachers what they assess and how frequently. The interview included other sections, but the ones described were where the examination was found to be mentioned.

Teacher questionnaires

The whole BUNJO teacher questionnaire comprised 60 questions, 11 of which were relevant to the present research. Only one of these questions directly referred to the *gimnazjum* exam. Most of the questions in the questionnaire followed the same format, namely a statement, followed by a 5-point Likert scale asking for the degree of agreement with the statement, from *I totally agree*, to *I totally disagree*. The question on the *gimnazjum* exam followed this format, as did a series of questions on the frequency of assessment of different aspects of language learning. The last set of questions asked teachers to give their opinion on the importance of different aspects of language in teaching and learning foreign languages, to which teachers were to respond on a four-point scale from *not at all important*, to *very important*. It should be pointed out that the first two of these items were flawed, in that the first did not specify the group which the teacher should refer to when completing it and the second set did not specify the language. However, as these teachers had agreed to take part in the BUNJO project, had received briefing materials about it, as representatives from each school had attended a briefing conference, and as they were all aware the project in 2012 referred to learners in specific groups learning English in Year One, there is a chance that this was the group they had in mind when answering. The respondents were also all teachers of English, so it is to be expected that this is the language they were referring to in their answers. This is not, however, an excuse for these important oversights.

Learner questionnaires

The first questionnaire in many respects mirrored the one teachers completed and parallel questions to the ones described in the Teacher Questionnaire were extracted. In it there was one question directly about the exam. Learners were asked to respond to the statement "We do lots of exercises during class preparing for the *gimnazjum* exam in English" on a five-point Likert scale, from 1, *I totally disagree* to 5, *I totally agree*. The last set of questions described above, on learners' views of what is important in teaching and learning foreign languages, however, were not in this questionnaire, but formed part of a second questionnaire, which dealt mainly with issues such as access to English outside school and information about the learner's home background. The content of the set of questions was the same as in the Teacher Questionnaire.

Learner interviews

Interviews with the learners included several sections, only one of which concerned the exam.

The section of the interview relating to the exam was comprised of the following questions:

1. At the end of lower secondary school you will take the *gimnazjum* exam in foreign language. What do you think about it?
2. Do you know something about the *gimnazjum* exam in foreign language? What?
3. Is the *gimnazjum* exam in foreign language important for you?
4. Is it as important as the other parts of the exam?
5. Has the *gimnazjum* exam in foreign language been mentioned in school?

Interviews with school principals

The interviews with school principals were extensive and covered a number of topics, including, among others, organization of language teaching in the school, measures taken to allow all learners to develop their potential, evaluation of foreign language teaching in their school, monitoring of cover of the core curriculum, and ensuring the effectiveness of FL teaching. One question specifically referred to the exam:

What role do the results of the [national] exams in foreign languages play in planning and monitoring foreign language teaching in your school?

Ethical issues

All participants in the longitudinal study were assured that any information they contributed would be reported maintaining their anonymity. This applies to any identifying facts or features, including specific location of schools. Participants and schools are consequently identified by num-

bers, and any names of people or places have been removed. Participants were also assured during the data collection that no-one in their school, other than themselves, would have access to any data pertaining to them from the study. It was believed that giving such assurance would alleviate any concerns, and open the possibility for the participant to be critical if they wanted.

4.3.4. Focal areas across the three phases of the study

Table 4.3 on the next page aims to clarify the relationship between the three phases of the project and show how focal areas connect.

It can be seen that in Phases I and II variables include: areas considered of importance in teaching and learning, assessment and views on the exam. Frequency of testing of reading and listening featured in both of the phases. In Phase I the focus was teachers, while in Phase II the main focus was learners.

In Phases I and III there was a focus on views of the exam from teachers and learners. Assessment was also a variable in both these phases. Both teachers and learners were asked about areas of importance in teaching and learning.

In Phases I, II and III there was focus on areas considered of importance in teaching and learning and on assessment.

In this way it can be seen that the design of the project allowed for comparison across and between the different phases.

4.4. Procedure

Each phase of the study will be described in turn.

4.4.1. Phase I: Baseline Study

The study was conducted in May and June 2008, that is at the end of the school year prior to the first administration of the national external examination in foreign languages, which took place in April 2009. Permission was first obtained from the school principals of 25 schools for the research to take place. The questionnaire was delivered, with a covering letter, to one of the teachers of English in the school, with a verbal explanation and a request to distribute it to colleagues. A date and time were agreed for collection of the completed questionnaires. The researcher collected both completed and blank forms.

Table 4.3. Focal areas in each of the three phases of the impact study

	Research questions and focal areas			Instruments
	Teachers	Learners	School principals	
Phase I Baseline Study 2008	What aspects of language are important in teaching/learning in *gimnazjum*?			Questionnaires Observation
	What is assessed?			
	How frequently are reading/listening practised?			
Phase II Initial Washback Study 2009	What changes were made as a result of the exam?	What aspects of language are important in learning FLs in *gimnazjum*?		Questionnaires from learners and teachers
	Opinions on effect of exam on learners	What is assessed?		
		How frequently are reading/listening/ speaking/writing practised?		
		Is the FL exam important?		
Phase III Impact Study 2012	What is important in teaching?	Opinion on exam		Field notes during piloting of interviews Interviews with teachers; principals; learners Questionnaires from learners and teachers
	What are the aims of teaching?	Knowledge about exam		
	What is role of exam in planning and choice of materials?	Is the FL exam important?	What is the role of exam results in planning and monitoring FL teaching in your *gimnazjum*?	
	Opinion on whether exam practice exercises are done	Opinion on whether exam practice exercises are done	How is FL teaching evaluated in your *gimnazjum*?	
	What aspects of language are important in teaching/learning in *gimnazjum*?	What aspects of language are important in learning FLs in *gimnazjum*?		
	What is assessed?	What is assessed?		

Observation

Teachers known to the author were contacted and asked if they would agree to be observed. They were informed that the aim was to observe different classes in lower secondary school to get an overview of the level of learners in different groups and in different contexts. It was made clear that showpiece lessons were not the aim and they were asked to simply teach as they normally do. The reason for this was to attempt to capture as realistic a picture as possible of everyday classroom teaching and for teachers not to feel they themselves were being evaluated in any way. They were also asked if they could persuade their colleagues in school to take part in the study. If teachers agreed, formal permission was then obtained from the school principal during a personal visit from the researcher. As a result, observations were conducted in three lower secondary schools in two different cities in Silesia with seven different teachers. Where possible the same teacher was observed teaching two different groups, although this was not always the case because of timetabling and logistic constraints. This also meant that there was not an even distribution of observation of different year groups, with only one lesson observed in class 1, as opposed to six in class 2, and seven in class 3. The plan below shows the distribution of schools, teachers and lessons. Where two classes are listed this meant different groups of learners were observed, but where the note "double lesson" is included this meant the same group were observed in one lesson lasting 90 minutes, with a break. School 2 had one additional hour of English a week, with learners studying 4 hours a week in class 3, while the other two schools followed a standard 3 hours per week programme.

School 1
Teacher 1: Class 3, Class 2, Class 2
Teacher 2: Class 2, Class 2

School 2
Teacher 1: Class 3 double lesson
Teacher 2: Class 2, Class 1
Teacher 3: Class 3 double lesson

School 3
Teacher 1: Class 2
Teacher 2: Class 3 double lesson

Observation data

The researcher took detailed descriptive notes during the lesson. Immediately after the lesson there was a short feedback session where oral comments were given and questions asked for clarification. Any further information obtained was then added to the lesson notes. This mainly

referred to information given about individual learners who had been seen as either representing a higher, or lower level than the group, or who were not apparently engaged in the lesson. Comments were also made about materials used.

4.4.2. Phase II the Initial washback study

Before the study took place both the learner and teacher questionnaires were first consulted with teachers of English in School 2, where observations had taken place during the baseline study, and adjusted. They were subsequently piloted in April in one class 3 in that school and minor changes made. The final version of the questionnaires, which was for pen and paper, were then distributed in May 2009, approximately three weeks after the FL examination had taken place. The learners had all taken the new exam.

4.4.3. Phase III the Impact Study

Data for this phase was taken from BUNJO, a national research project described above. Due to its scale, the project was put up for public tender and the first year data collection was conducted by the firm Millward Brown SMG/KRC (which will be referred to as the contractor). The firm were supplied with the instruments to use, with materials for training interviewers and data collectors, and with full and detailed procedures to follow. This was done to ensure reliability of the data through consistent data collection techniques and procedures. The first year of the project was supervised jointly by the author and Anna Michałowska, FL section Educational Research Institute in Warsaw (ERI), in cooperation with a representative of the contractor. ERI were responsible for all decisions relating to the quality and content of the study, while the contractor undertook practical aspects of the data collection. The contractor functioned within strict guidelines, drawn up by the author and other members of the FL section, which governed all aspects of the operation.

Data collection procedures

Learner questionnaires

Learner questionnaires were administered in pen and paper versions by data collectors, recruited by the contractor to specifications given by ERI and who had attended a 135 minute training session. This was prepared and run by members of the FL team, using materials for which the author was partly responsible. A protocol, explaining the exact procedure to be followed, together with the questionnaires, were carefully explained. The data collection took place in a window from April to early May 2012 and appointments were agreed with each school. A second administration was arranged for

learners who had been absent on the day of the first administration. Learners completed the questionnaire in class in the presence of a teacher and the data collector. There were two different questionnaires, which were completed on two different dates. Each learner had a unique identity code.

Teacher questionnaires

As finding a time when all teachers in one school would be available at the same time was deemed impractical, it was decided to administer the teacher questionnaire online. This was prepared in stages. The contractor was given the paper version of the questionnaire and instructions as to how it should be converted to an online version. A draft of the online questionnaire was then made available for the author to check. Next, following specifications set out in the agreement with ERI, the contractor arranged a stress test, which simulates the maximum predicted number of respondents online simultaneously on the platform which they planned to use. The aim of this is to make sure that the system is capable of dealing with so many responses at the same time. A report on the stress test was delivered and checked by an IT specialist in ERI. Finally, codes for access were sent to ERI for the FL team to trial the questionnaire *in situ* on the dedicated platform. Final minor revisions were then made. Once the final version was ready, teacher participants in the study each received personal access codes, which allowed the contractor to check who had taken part. The questionnaire was available for a period of approximately one month. Reminders were sent to teachers as the window during which the questionnaire was available was coming to an end.

Interview

Interviews were conducted in Polish in the sample schools with the school principal (115), with teachers of the groups nominated by the school for the study (301), (more detail is found in the section on the sample) and with students in each school drawn at random from those groups (480). Teachers who were interviewed had earlier also completed BUNJO written online questionnaires. Learners interviewed should have already completed at least one of the tests included in BUNJO and two questionnaires. The interviews were conducted by 54 people recruited by the contractor, according to specifications given by ERI, who attended a training session, which will be described below.

Each interview followed an interviewer frame, which gave the interviewer the exact questions to use and full instructions. The frames had been piloted by the author and members of the ERI FL team. This involved discussion with the participant following the interview. Comments were written up as field notes. Minor revisions were subsequently made to the interlocutor frame and questions. Some of the piloted interviews were filmed for use in the training session. Interviewers (54) all took part in a training session

(7 hours), run by members of the FL section, using materials which were created together with the author. These covered the background to the project, interviewer behaviour, the questions, predicted problems and how to respond, and technical procedures such as use of recording equipment, identification coding of data and documentation. Interviewers were shown films of sample interviews for orientation. Role playing activities were used during the training to help the interviewers familiarize themselves with the materials and procedures.

Each interview in school was audio-recorded on two battery-operated digital recording devices, one main and one back-up. Both interviewer and interviewee had unique identification codes which were recorded at the start of the interview. The size of the data sets meant that scrupulous organization of the recordings and related documents was necessary. Interviews took place between 19 March and 15 May 2012, with separate sessions for the three groups of respondents.

Transcribing interview data

The contractor recruited 10 people who were responsible for transcription of the interview recordings. This was done using the programme f4. A training session (4 hours) was organised, attended by a member of the FL section, which included a practical transcribing exercise, where the work of the transcribers was checked for accuracy. Transcriptions were done verbatim, including false starts, hesitations and repetitions.

4.4.4. Coding procedures

In this section the various procedures used for coding different parts of the interview data will be described

BUNJO Coding of the interview data

The author worked together with Kamil Sijko from ERI, and his colleagues Karolina Malinowska and Iwona Babiak, to explain the research design and set the foundations for how the interviews could be analyzed. It was decided to use a grounded theory approach (Strauss & Corbin, 1998), whereby the interviews would first be read and then a series of codes created to describe and tag trends and patterns observed in the responses. This system is known as open coding. The ERI team worked first with 10 transcripts from each of the respondent groups to identify initial codes.

The contractor recruited a team of 24 people to code the transcripts, who attended a 3 day training session run by the ERI team, led by Kamil Sijko. The first part of the training, run by a representative from the contractor, was in how to use the programme Maxqda, selected by the author for the coding process. The group of coders then worked together with the ERI team, successively working on interviews from each of the groups of respondents,

school principals, teachers and students. This included working together on chosen fragments and creating codes, first a main code for a topic area and then breaking this down into more detailed axial codes to account for different aspects. A simple example of this is how the interview question to students "Has there been anything said about the *gimnazjum* exam in foreign language at school?" was coded. The main code was "*gimnazjum* exam" and this question then had three subcodes, "yes", "no" and "don't know". A somewhat more complex coding system can be illustrated by what was used for the question "Is the exam in foreign languages as important for you as other parts of the exam?" Here the main code was "importance of the exam" which was then subdivided into two branches: absolute [value] and relative [value]. The absolute value branch had three codes "important", "not important", and "unclear response" while the relative value branch was subdivided into "more important' , "less important" and "as important as". Coding in Maxqda includes highlighting the fragment of text which is to be coded and tagging it with a code. The programme then allows the researcher to access both a whole interview with codes visible, and coded fragments highlighted, but also to see all examples of text allocated a given code across the sample. Coding stopped at this level, with no further attempt to theorize. The aim for BUNJO was that the data could be used to provide a broad description of patterns and trends in the data on a wide number of topic areas.

The training session also included practical parts when the coders worked first individually on the same segments of interviews and then compared their coding with others, with the aim of reaching consistent use of the agreed codes. After the end of the session, the coders worked in their own homes on allocated numbers of transcripts.

During the coding itself an online platform was made available by the contractor, via which the coders could contact each other, and the team. An online forum, moderated by members of the ERI team, allowed for clarification of problems, posting of fragments of interviews which were difficult to code, and online discussion, to ensure codes were used as consistently as possible. Any new codes created during the process of analysis were shared and described, and all coders added them to their code trees. A window, during which new codes could be posted, was set to try to limit the creation of new codes, as each new code entailed every coder returning to what had already been coded and coding once again. Coders were required to follow the forum very closely during this time, and to regularly update their code trees.

4.4.4.1. Coding for the Impact Study

Learner interviews

Learners interviewed were aged approximately 14 at the time of the interview, which took place towards the end of their first year of study.

The age of the learners is important. Many young people of this age are not particularly forthcoming when asked for their opinions, either because they feel self-conscious, or because they are shy, or perhaps because they have not given the topic deep consideration and have no response to make. So to be candid, the parts of the learner interviews relating to the exam are disappointing. This may be the result of a problem with the design of the questions, which may have led to a predominance of single word answers. In the field trial, by contrast, longer responses were elicited, perhaps as a result of a longer wait time and use of signals of encouragement on the part of the author/interviewer. In the main study, however, many of the interviewees responded extremely briefly and few attempts were made by the interviewers to encourage them to expand their answers. Codes used therefore for analysis in this project are the codes created during the BUNJO coding, as it was felt there was little to be gained by attempting to re-code them. It was also decided, after due consideration, to limit reporting of the data from the learner interviews mainly to quantitative analysis, as it was felt that there was no deeper insight to be gained from attempting qualitative analysis.

Teacher interview data

This impact study sets out to investigate one part of the BUNJO teacher interview data set in greater detail, with the aim of coming to a deeper understanding of the role of the FL examination in the lower secondary school. The codes applied by the BUNJO coders were broadly descriptive and covered all of the areas in the interview, whereas, for this study of impact much more specific information was needed, relating to the areas predicted to show the effects of the exam. For this reason a different approach and different coding were needed.

The author undertook new coding of the teacher interviews using the following procedure. Ten interviews were first read and re-read to identify areas where the *gimnazjum* FL exam was mentioned. Initial open codes were created to identify broad topic areas. Open codes emerged as the data was read. Whenever a new code was created, the researcher returned to the first interview and re-read all the interviews to see if this code was also present there.

Initial open coding created 14 codes referring to different topic areas (number given is the code number) as follows:

1. choice of book relates to exam

   In describing the books used in school the teacher specifically mentions that a motive for their selection was the exam.

2. clash between core curriculum and exam content

   The teacher refers to difficulties caused by perceived differences between the core curriculum, which they are mandated to cover, and the content of, or syllabus relating to, the exam

3. extending syllabus focus under influence of exam

   The teacher mentions deliberate inclusion of a new area/areas of focus in the course content and attributes these changes to the exam

4. aim of teaching is exam related

   In stating their aim for teaching in lower secondary school the teacher specifically mentions preparing learners for the exam. This code was also used for any additional classes organized under article 42 of the Teacher's Charter which the teacher describes as exam preparation, or as providing remedial, exam-related, work. (Article 42 of the Teacher's Charter required teachers of *gimnazjum* to offer 2 hours a week over and above their teaching load, which could be used to offer whatever additional teaching support the school deemed necessary.)

5. exam results are important

   The teacher describes the exam results as being important in their work. This includes descriptions of accountability, both external and self-regulated. This code is also used when the teacher refers to the importance of exam results as justification for decisions relating to planning or choice of materials.

6. assessment changed because of exam

   The teacher explains changes made in the subject assessment, or school assessment system, which they ascribe to the exam.

7. planning of work is guided by the exam

   The teacher describes their rationale for planning work in school, or in lessons, and specifically relates this to the exam.

8. exam affects work throughout all three years of school

   The teacher states that the exam affects all classes in the school.

9. assessment is planned to the exam

   When describing the assessment system, the teacher specifically mentions that it relates to the exam. This code is also used when the teacher describes the use of mock exams, or 'diagnostic' tests, which they state are based on the exam.

10. speaking is assessed

    This code, unlike the others, hypothesizes that a sign of negative washback will be narrowing the focus of classroom assessment. As speaking is not directly assessed in the exam it might be noted that speaking is less frequently reported as being assessed. For this reason the frequency of reporting of assessment of speaking was counted.

11. aim is to meet expectations of others which relate to the exam

    When describing the aims of their teaching, the teacher ascribes the decision to focus on the exam to meeting the expectations of others. This may be, for example, parents, the learners, the school principal, or society.

12. course book focuses on the exam

The teacher describes the course book as containing sections, tasks, a syllabus, etc. which is related to the exam.

13. narrowing syllabus to fit exam

The teacher specifically describes excluding areas from the syllabus, or specifically reducing time spent on an area of the syllabus, for reasons which they ascribe to the exam.

14. role of the exam

The teacher describes how they perceive the role of the exam, its aims, its intentions, etc.

The set of interviews were then coded using the main codes. Memos were added as ideas occurred.

> [A memo is] the theorizing write-up about codes and their relationships as they strike the analyst while coding…it exhausts the analyst's momentary ideation based on data with perhaps a little conceptual elaboration.
> Glaser (1978: 83–84)

Verification of coding

As coding is a subjective process, questions can be raised about the reliability of analysis based on it. In an attempt to increase reliability, the following process was followed. A decision was made to have 20% of the sample double-coded, using a second coder. The second coder, an experienced teacher from upper secondary school, was familiar with the BUNJO project and had worked on coding transcripts of observed lessons, so had some experience of coding. A fee was offered for her services.

The researcher supplied the second coder with the main code descriptions (given above) together with extracts from interviews which illustrated each of the main codes. In this way, the second coder had samples to standardize their decisions against. In addition, one complete coded interview was made available to the second coder. The coder was asked to highlight places in the interview text which they chose to code and to apply a code number to each instance.

To ensure random sampling for the second coding, a two-tier sampling process was used. First 60 numbers of schools were drawn randomly from the set of 120 school numbers. Next for each school selected, identification codes were prepared for each of the interviews. This was necessary as there were differing numbers of interviews in each of the schools, ranging from 1–3. Some of the teachers included for interview were categorized N, which meant that they were teaching in *gimnazjum*, but were not the teachers of the groups of learners who were the subjects of the study. Teachers with this categorization were not included in the draw. For each school a separate draw was then made, with one identification code being selected

from the number of interviews in that school. This gave a randomized pool of 60 interviews, one from each school in the sub-sample, for second coding.

The verification process

When the second coding was completed, the two coders compared the codes used in the interviews. The codes used in each interview were noted for each coder and compared for agreement. The total percentage of agreement between the two coders was then calculated. This was found to be 86% which was felt to be acceptable for the purposes of verification.

Quantitative analysis

Once the verification process was completed quantitative analysis was carried out. First the main codes were counted for frequency to allow an overall, quantitative, picture of the data to emerge. Next a deeper analysis was conducted. To do this, extracts from the interview transcripts which had been tagged with each main code were re-examined separately. Where appropriate, sub-codes (known as axial codes) were created to give a more detailed analysis. The memos which had been written were used to help in this process.

Axial coding

In this process the main codes were sub-divided to allow deeper analysis of the topic. This was adopted in the case of many (9) of the 14 codes. Codes 1 (see above), 8, 10 and 12 were excluded as these were binomial (ascribed to the exam/not ascribed to the exam; affected/not affected; included/not included; mentioned/not mentioned). Code 2 was also treated as binomial, as all mentions were found to give the same reason. This will be described below. The axial codes for the remaining codes will be described in the section where that code is reported in the next chapter. The number of axial codes depend entirely on the data and so some main codes have a small number of sub-codes, while others have a larger number.

Analyzing interviews with school principals for this study

Transcripts of 115 interviews with school principals were analyzed by the researcher. Each interview was first scanned to find references to 'examination' using the document search facility in Word. This highlights the key word in the text. The word 'examination' was chosen after experimenting with different key words, such as 'results'. It was found to identify the greatest number of instances where issues relating to the examination were referred to. However, in addition to the scan, the text was also read through quickly to check no other reference to the exam was made. Sections of the interview referring to the exam were then copied and pasted into a new document and,

using a process of highlighting and commenting, they were described in note form. Next, the collection was read and re-read many times, with a process of noting, tagging, and writing of memos, until trends or patterns emerged. A series of themes were identified and extracts from the interviews selected to illustrate the concepts involved.

The findings are reported in the next chapter.

In this chapter we have described the rationale behind the research, pointed out its significance and explained the research concept. We considered methodology used in studies of washback and issues arising from the use of qualitative data obtained from interviews. We explained the design of the study, showed the three phases and relationships between them. The aims of the project as a whole and then the aims of each phase were presented, together with the research questions. In the next section the method was explained, beginning first with information about the participants and the samples, followed by descriptions of the instruments and an outline of the procedures for each of the three phases. For the third phase we also explained the coding procedures which were used to analyze the qualitative data obtained from interviews.

CHAPTER 5

Presentation of Data

In this chapter we present an analysis of the data gathered in a longitudinal project to study the influence of the national external examination in English on teaching and learning in lower secondary schools in Poland. The project took place in three main phases:

Phase I (2008) Baseline study

Phase II (2009) Initial washback study

Phase III (2012) Impact study

Information about the project design was given in Chapter Four.

5.1. Phase I: The baseline study

As described in Chapter 4, the research on the impact of the external examination in foreign languages began with a study designed with the intention of obtaining a picture of the situation in lower secondary school prior to the introduction of the external FL examination.

5.1.1. Questionnaire study

A report on this part of the baseline study was published (Ellis, 2009a) but a summary will be provided here, so as to bring the whole project together in one place and provide a context for Phase III the Impact Study.

In the questionnaire (see appendix 1) teachers were asked to rank what they thought was of importance in their teaching in lower secondary on a scale of 1–6. Their responses were totalled and ranking calculated. The following was found:

1. Vocabulary
2. Speaking
3. Listening
4. Grammar
5. Reading
6. Writing

In a later part of the questionnaire teachers were asked what aspects of language they felt was important for their learners in learning English. The question was differently formulated, included other aspects of learning English, and unfortunately, in error, failed to include listening. However, despite these flaws, some comparison is still of potential interest.

The ranking of areas of importance for learners as perceived by teachers was as follows:

1. Speaking
2. Vocabulary
3. Reading
4. Pronunciation
5. Grammar
6. Culture
7. Writing

When we compare what teachers felt was important and what they thought was important for their learners, we note the reversal of the first two places, with teachers feeling that learners place most importance on speaking, while they themselves placed vocabulary first. Teachers also appeared to consider that learners felt reading is of greater importance than they did. Both parties were similar in their view of the unimportance of writing, but there appears to be, despite the design problems with the item, a disparity between teachers' own values of the relatively low importance of reading and their views about how they thought their learners perceived it, with reading ranked higher for the learners.

The next area concerns assessment. A list of the same areas was given for teachers to mark all for which they give grades. Teachers claimed that they give grades for all the six aspects of learning listed, with marginally more selecting grammar (17%) and vocabulary (18%) than writing (16%) or speaking (16%). The skills of listening (14%) and reading (14%) were a little less frequently checked on the list. It is interesting to note the dissonance between the lack of importance teachers ascribe to writing in their ranking and its relatively high position in the list of what is graded. As the differences in the frequency count of what is graded are small, however, we cannot draw any definite conclusions from this.

Next, the respondents were asked about where they get the tests they use in class from. Teachers showed that they draw on a variety of sources for tests, including their own ideas (26%), but this was quite heavily outweighed by use of existing published materials, with tests from the publisher (21%), from the course Teacher's Book (14%), from the course book (11%), other course books (7%), grammar books (10%) and from the internet (10%).

The next questions asked about the percentage of time spent on the course book in lessons and in the course. The course book was found to play a strong role in teaching, with 78% of the sample stating that 80 per cent of the time in

lessons is spent on it, and the same number claiming that 80% of course time during the school year is course book based.

The strong role of the course book was also found to apply in the next set of questions, which were on the frequency of listening and reading practice in lessons. When teachers were asked how often they practice the skill the choice "whenever there is one in the course book" was selected most frequently from all the options given in the case of both listening and reading, suggesting that the course book also plays an important role in planning.

Questions were also asked about tasks done connected with reading. Closed tasks were most frequently selected, ("True/False questions"– 21%, "Multiple choice" – 16%, and "Matching" – 13%), which may suggest that these are perhaps tasks which appear in the course book (Ellis, 2010), found a predominance of these types of tasks in an analysis of eight course books approved by the Ministry of Education for use in lower secondary school. In addition, tasks focused mainly on understanding specific, detailed information in the text). Other, more traditional, options were somewhat less frequent, with "translation" at 10% and "reading aloud" at 11%. "Students write answers to questions" was also relatively common at 14%.

In an attempt to find out whether teachers support the development of reading skills, rather than simply test reading comprehension, two questions were given, with a series of actions for the teacher to select. An open option was also included. There was no restriction on the number of actions that could be selected. The most commonly chosen actions were as follows:

Teach strategies for reading (selected 35 times)
Explain step-by-step how to find the answer to a reading comprehension question (28)
Give lots of reading practice (28)

Less popular were "Regularly use translation" (20) and "Teach how to use an English-English dictionary" (21).

When asked to select actions they take to help students do a reading comprehension exercise, teachers most frequently chose actions relating to preparing learners before tackling the text, with "Pre-teach key vocabulary" first (36 selections), followed by "Introduce the topic of the text before reading" next (35). "Ask learners to tell you words they don't understand after reading" and "Translate the questions for the learners if there are problems" followed with 33 selections each. "Encourage learners to use dictionaries in class", which could be interpreted as promoting learner autonomy, was the least commonly selected action, with 18 counts.

Asked to select actions they use to help learners find the answers to a listening comprehension task teachers favoured two actions: "write problem words or phrases on the board" (39 counts) and "play the recording again" (38). Least commonly chosen was "show them the transcript" which was selected only 3 times. To be fair, we should remember that this was 2008,

that few classrooms had projectors at that time and that photocopying was severely restricted in most schools. Tape scripts at this time most commonly appeared only in the Teacher's Book.

The majority of teachers (78%) claimed they were familiar with the new FL examination and 68% of these stated that they had done practice tests with their learners. Opinion on how difficult the exam would be for their learners was divided, with 55% claiming it would not be hard, and 45% that it would be a challenge. Among the reasons for this teachers gave responses mainly about the low level and motivation of their learners. Concerns were also expressed about mixed levels of ability and individual learner differences.

To conclude this report on the baseline questionnaire study, we can sum up by saying that the course book was found to be of importance to teachers, appearing to strongly influence choice of materials, and frequency of occurrence of reading and listening tasks. Published tests also featured strongly, although some 25% of the teachers claimed they also used their own ideas. There were discrepancies in the ranking of what teachers viewed as important in their teaching and what teachers considered was important for their learners, particularly in the case of reading, which teachers thought learners considered more important than they did themselves. However, vocabulary and speaking were found to occupy the first two places in both lists, although in a different order. Teachers placed writing very low in importance, both for themselves and for their learners.

5.1.2. Observation study

In order to verify information obtained from the questionnaires, a series of observations were carried out. Information about the choice of schools was given in Chapter Four. For reasons of space, detailed information about the lessons will not be given, but the findings will be summarized in description of a series of areas considered worthy of note. First, attention should be drawn to a design flaw in the observation study, namely the timing of the school visits. Observations took place in the last week of May and the first week of June 2008 and included all three years of the stage. Although the school year officially ended in the last week of June, it was found that in some of the groups learners appeared to be ready for the vacation and not very engaged in lessons. This varied considerably from school to school, with School Two working "normally". In Schools One and Three the variation appeared to depend on the teacher, their plan for the lesson and their management skills, with some lessons observed where learners were fully and actively engaged, while in others teachers were struggling to mobilize learners to take part. In short, it was perhaps not the best time to be gathering observation data, although it did give a very clear picture of the demands groups of teenagers can place on their teachers,

and the importance of choice of materials, planning, rapport and effective classroom management.

Teaching Materials

Of the 14 lessons observed, twelve were centred on the course book. Three of these lessons used course book based tasks for only about half of the lesson time, while the remaining nine lessons followed the book exercise by exercise. One of the remaining two lessons was spent mainly going over a practice test, and in the other learners wrote an essay. The three lessons where the book was only used for part of the time were taught by two teachers with considerable teaching experience (17 and 25 years respectively).

A noticeable feature of the lessons taught by the more experienced teachers was their ability to personalize information, or tasks, in the published material, to make them more engaging for their learners. This was done, for example, by adding a pre-task, or by asking learners to give their opinions on or experience of a topic. Learners in these lessons were visibly engaged, and asked questions of the teacher about the material.

Treatment of reading and vocabulary

Tasks involving reading comprehension featured in six of the lessons, and formed the main part of four of these. Learners, regardless of the year group they represented, appeared to have difficulty with comprehension, which seemed to be resulting from having limited lexical knowledge. Some demonstrated signs of extreme frustration. The texts, which varied in length and difficulty, came from course books, with one exception, where the teacher had prepared a questionnaire for the learners, which they first did as a comprehension task and then used to interview a partner. As a follow up they had to make a short presentation about what they had learnt about their friend. This was done in stages, first making notes and then rehearsing together in pairs, before finally speaking out in front of the group. Where the reading texts were in the course books they were accompanied by comprehension exercises with multiple choice, true-false or open questions. The pattern followed was mainly for silent reading, followed by learners asking the teacher about unknown words (usually in Polish, with explanation given by the teacher in Polish). Next, the comprehension task was done individually and then checked, usually by the teacher nominating a learner to answer. Where learners had difficulty, some of the teachers supported them with a series of guided questions aimed at leading them to the answer. Others asked another student for the answer, and in one instance the teacher encouraged learners to use bilingual dictionaries to find the Polish equivalent of words. New vocabulary was generally noted on the board by the teacher, as single words with either a Polish translation, or a synonym or explanation in English. Words were not written in context.

Grammar

Four of the observed lessons focused entirely on aspects of grammar, while grammar featured as part of a further three lessons. Where it was the focus of attention, learners were expected to do exercises from the course book, either individually and then checked with the whole class, or in front of the class at once. In one case, the teacher first gave learners the rules for the item he was targeting, writing these on the board, and then followed this with a series of exercises. In another case, the teacher had prepared a chart on the board before the lesson, which was partially completed, and learners worked together in pairs, on the basis of practice exercises in the book, to complete the verb paradigm in the chart. This was then checked as a class and the learners then copied the final version of the chart into their notebooks.

Where grammar was only a part of the lesson, it was course book based and came as a follow-up to either a reading or a listening task, where the text was used to elicit and focus learner attention on target items, before doing practice exercises from the book.

It was noticeable that many learners had difficulty, not so much with the grammar items, as with the vocabulary of the examples in the tasks, which in many cases interfered with their understanding of the exercise. Some learners, however, were being asked to do tasks focusing on grammatical items which appeared to be beyond their level, for example, an exercise on the passive voice, using a variety of tenses, where learners were unable to give the required part of any of the verbs in the task.

In the course of other activities some teachers would point to individual verbs in texts and ask learners for the base form, or the past form. Learners, in speaking, or responding to the teacher, were seen to have a tendency to use the present tense in many instances when past was needed, and to have difficulty with subject/verb agreement. Most of the teachers used immediate, overt correction of mistakes, supplying the correct form at once, with no attempt to elicit the correction from the learners.

Listening

Listening to a recording featured in only three of the lessons observed and only in one of these was it a task focusing on developing the skill of listening. In the other two the listening was used to introduce new language, which in both cases was grammar. In these two lessons the teachers, both very experienced, extended the course book tasks, by first exploiting the listening text for comprehension and only after that introducing the language focus. One of the texts was a dialogue and this was used by one of the teachers to get learners producing their own dialogue in a restaurant. In both these cases, the teachers did substantial preparatory work before listening, setting the context, pre-teaching key words, setting an initial

listening task and checking learners understood what was expected of them. Where the listening was for understanding of a text, it featured as the continuation of a text which had been read by the learners. They had already had considerable difficulty with the vocabulary of the written text, which the teacher attempted to deal with through eliciting translation, or if this failed, supplying the meaning in Polish for the class. No vocabulary preparation was done prior to the listening exercise and once the recording was started it was very quickly visible that learners gave up trying to answer the comprehension questions and started talking to each other in Polish. The teacher had considerable difficulty getting the group back on task, and in the end resorted to getting them to read the transcript of the listening aloud round the class and asking them to translate it sentence by sentence.

Speaking

The amount of opportunity learners were given to speak depended, of course, on the focus of the lesson. Where learners were doing exercises from the course book, they often responded to the teacher's request for the answer with a single word, or short phrase in English, but if asked to explain why they thought this was correct, often switched into Polish. Some teachers, however, asked learners to read the whole sentence aloud and also corrected pronunciation mistakes. Two of the lessons included more extended oral tasks, one with learners creating dialogues and the other with learners interviewing each other in pairs. Some of the teachers insisted that if learners spoke in Polish, when asking the teacher a question, they should repeat, but in English. In other situations, teachers switched into Polish and responded. Students were also asked to read aloud instructions, or fragments of text from reading exercises, and attention was paid to pronunciation by some teachers. In two lessons no such correction was made, despite the fact that learner responses were not very intelligible. Opportunities for longer production in English, or the chance to freely develop responses were relatively few in the observed sample of lessons.

Interaction

In most of the lessons, the classroom was dominated by the teacher at the board, with the class working together in lockstep, and the talk directed by the teacher, who asked questions and nominated learners to answer. Pair work featured only in the three lessons taught by very experienced teachers, and no group work was observed. In some lessons, learners worked individually, doing exercises based on the book, which were then checked together in class. In one lesson, the teacher set the students an essay topic which they then worked on individually, asking him for help when they needed it. The teacher monitored closely, correcting and advising the learners. Learners were seen to have great difficulty with vocabulary

in this task, and appeared to be first composing the text in Polish, and then attempting to translate it.

Classroom language

The language used in class varied from teacher to teacher, and the same teacher sometimes changed the language they used, depending on the level of the group. So, while one teacher taught a year three class speaking English most of the time, the same teacher in a class two lesson used English only about half of the time in the lesson. Where teachers used predominantly English, Polish was used mainly to explain vocabulary items, or to clarify problems with content information in a text. Teachers working with less able groups tended to use mainly Polish as the classroom language, using translation as a scaffolding device to support understanding of texts, to explain vocabulary, to explain grammatical rules, or to compare the differences between items which learners had confused. In these instances teachers often asked the learners to translate from English into Polish, particularly instructions or comprehension questions.

Tests

In several cases the teachers commented in the post-lesson discussion that they had tried practice tests for the FL external exam with their learners, in order to diagnose what kind of difficulties they had. In one observed lesson the teacher went through such a practice test, which had been done by learners in class 2, doing the test round the class with the students, eliciting their answers and asking them to explain what they had given. In this way it was possible to see what was challenging for the learners. In general, vocabulary was the main area of concern for the teachers. They predicted this would hold the learners back in both reading and listening tasks. Listening was also cited as being something that was worrying, particularly tasks which did not only test understanding of specific information. In the observed lesson, and also from comments of teachers, it was clear that in reading comprehension learners were adopting a strategy of picking out a single word and then guessing wildly, assuming that the question contained the exact same words as in the text, so "matching" the answer, on the basis of it containing that same word. As such direct referencing is rarely used in the FL exam test tasks, except in distractors, this is not a successful strategy. Teachers commented that learners had become used to direct referencing in course book exercises. The learners were also used to true/false tasks, and so had become accustomed to guessing. Once again the course book effect was noticeable, this time in how learners anticipated tasks and in the strategies they had developed to deal with them.

To sum up, what was observed in the 14 lessons appeared to confirm information obtained from teachers in the questionnaires, with course book based lessons predominating. The difficulty learners were observed to have

with vocabulary helps to explain why teachers placed it at the top of their list of what is important in their teaching at this stage.

5.2. Phase II: Initial Washback study

The first administration of the new external FL exam in lower secondary schools took place in April 2009. In order to begin to assess whether the examination was having an effect on teaching and learning in *gimnazjum*, questionnaires were designed for students in class three who had taken the new exam and a second for teachers teaching those classes. A report on this research was published (Ellis, 2009b), but a summary will be given here, to bring the whole project together.

5.2.1. Learner questionnaire

Only selected questions will be described and discussed, for reasons of space. The questions which directly relate to other parts of the project have been chosen.

The first question of the learner questionnaire asked learners to rank aspects of language in order of importance in the learning of English in lower secondary school. It mirrored the question asked of teachers in 2008. Similar questions were asked both of teachers and learners during the main study.

The next block of questions asked learners how frequently each of the four different skills had been practised in English lessons during the past school year. Respondents had four options to choose from: once a week; in every lesson; rarely; according to the course book. In 2008 teachers had been asked similar questions, but only referring to reading and listening. Learners were also asked to select from a list and check all aspects which they had received grades for in English during the school year. In the questionnaire in 2008, teachers were asked what they give grades for. However, the list of options was shorter, but an open option was included. Questions about what is assessed were asked in Phase III the Impact Study of teachers in interview, and of learners in one of the questionnaires.

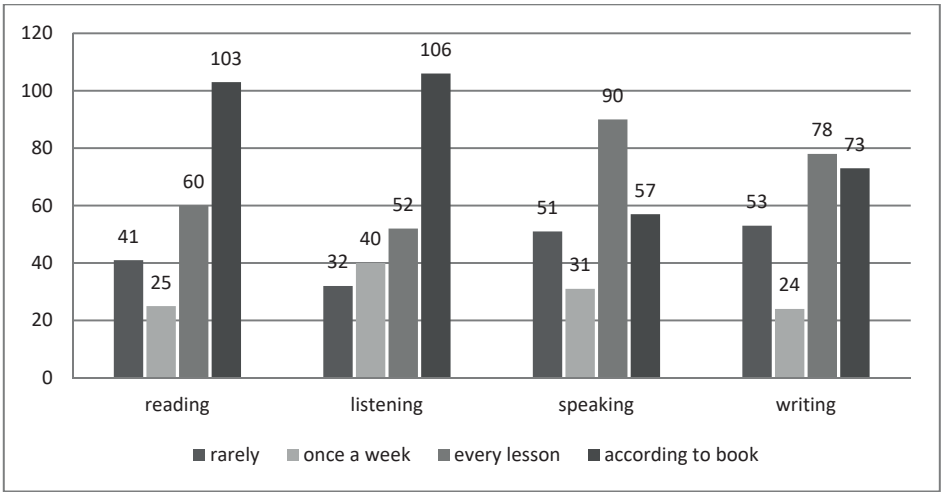
The next set of questions asked for the learners' opinions on difficulty of the exam which they had taken. These were closed, with a series of answers to select from. The final question in this set asked learners to rate the importance of the exam for them on a 5-point Likert scale. This question was asked during interviews with learners in the impact study.

Findings

Learner responses to the first question, "In your opinion what is the most important in learning English in lower secondary school?" were analyzed and rankings calculated. These were found to be as follows:

1. speaking
2. vocabulary
3. reading
4. listening
5. grammar
6. writing

The next set of questions refer to the frequency different skills are practised in English lessons according to the learners. As can be seen from the graph below (see graph 5.1) the largest number of learners thought that reading and listening was most often practised if such an exercise was found in the course book. The numbers who felt these skills were rarely practised were the lowest of the four skills with 17.9% saying reading was rarely practised and 13.9% listening. The question remains, however, how frequently the skills appear in the course book, so assuming that reading and listening are practised with similar frequency on the basis of these responses would be incorrect.

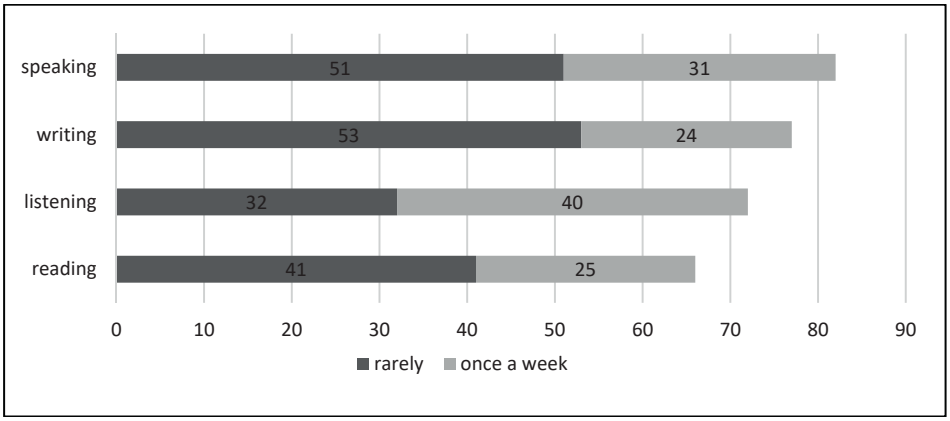


Graph 5.1. Frequency of different activities in English lessons according to learners

For speaking, the largest number of learners considered that it was practised every lesson. It should also be noted, however, that 22.3% (51) of the learners felt that speaking was rarely practised. Opinions on writing were divided between those who thought it was practised according to the course book and those who felt it happened every lesson, while 23.2% thought it was practised rarely.

If we aggregate the numbers of learners who thought a skill was practised rarely and those who thought it was only practised once a week (remembering that these learners have three 45 minute English lessons each week), we see that the productive skills of speaking and writing appear

to be less frequently practised than the other skills (see graph 5.2). Reading is practised most frequently. What we do not have in this data is information why this is the case.



Graph 5.2. Aggregated frequency of practice of skills: "rarely" plus "once a week"

The next question asked learners to state in which aspects of language learning they had received grades during the school year. They were given a choice of areas and asked to tick all of those which applied.

Table 5.1. Initial study. Aspects of language learning for which learners claim they receive grades

	grammar	vocabulary	reading	speaking	translation	listening	writing	projects
n =229	184	150	146	130	119	105	93	73
%	80.3	65.5	63.7	56.8	52	45.8	40.6	31.9

From Table 5.1 we see that the largest number of learners (80%) claimed they were given grades for grammar. This stands out above all the other categories. Vocabulary and reading follow with 65% and 64% respectively. Speaking is next at 57%. Below this come translation (52%), listening (47%), writing (41%), and project work (32%). It should be pointed out, however, that there appear to be problems with this way of collecting information about what is graded. Questionnaires were arranged for analysis in sets, according to the group in which learners had English. It was observed that there were substantial differences in responses about what they received grades for, between learners who attended the same group, taught by the same teacher, which calls into question the reliability of the information. These are teenagers who were being asked to think back over a school year. It is quite possible that learners remember differently. It is also possible that they remember selectively, perhaps affected by their personal preferences, strengths or even weaknesses.

The next questions to be analyzed are directly connected with the examination. Learners were asked to express how important the *gimnazjum* foreign language examination was for them on a five point scale from "very important" (1) to "not important at all" (5).

Table 5.2. Learner responses to the question: How important is the *gimnazjum* FL exam to you?

Value	1 Very important	2	3	4	5 Not important at all
n = 227	34	65	76	33	19
%	15	28.6	33.5	14.5	8.4

We see that there is a trend towards the exam being considered important by these learners. If we aggregate the positive values (1 and 2 on the scale) we see that 99 (43.6%) of the learners felt the exam was important, as opposed to 52 (22.9%) who gave negative vales (4 and 5 on the scale). Approximately one third (33.5%), of the learners (76), however, were undecided as to their opinion.

Learners were also asked their views on which parts of the examination they had found the easiest and which the most difficult. 43% found *Reading* the easiest, while 17.8% declared it the most difficult part of the test. 49.8% thought *Listening* was the hardest part, while 32.1% claimed it was the easiest. *Communication* was found the hardest part by 32.4% and the easiest part by 24.2%. Interestingly the Central Examination Board (CKE) report on the examination (2009: 37) states "Learners coped best with the tasks assessing listening comprehension", while finding that the average results for reading were the lowest of all parts of the test.

5.2.2.Teacher questionnaires

In the first questions, teachers were asked directly if they had made any changes in their teaching in class 3 as a result of the new examination. The majority (95%) agreed that this was the case. The next question, dependent on teachers having answered the first question in the affirmative, gave a list of statements for teachers to select from and mark all which applied (see Table 5.3). The most frequently selected actions related to two areas, choice of materials and choice of activities. Teachers often (62%) added an exam preparation book to their teaching materials, and many of them used practice tests (81%). Influenced by the content of the exam, teachers increased the number of listening tasks done in class (81%) and to a lesser extent the number of reading tasks (52%). Some also reported changes in the type of tasks done for listening (19%) and for reading (28%). Picture description tasks were done by approximately half of the teachers (57%), despite the fact that this task was removed by the Central Exam Board in November 2008.

Table 5.3. Initial Study. Changes made to their teaching by teachers of class 3 lower secondary as a result of the new examination

I did practice tests	81 per cent
I did more reading tasks	52
I did more listening tasks	81
I did picture description tasks*	57
I changed my course book	28
I added a new book, especially to prepare for the test	62
I changed the type of reading exercises I usually do	28
I changed the type of listening exercises I usually do	19
We practised matching examples of language to the Polish function word (np. *Przeprosić* – 'I'm sorry')	24

*the exam was expected to include a task where learners were to write a description of a picture. This task was, however, dropped by CKE in November 2008

Teachers were asked to assess the attitude of the majority of their learners to the new exam from four options given. 29.1% selected *It really motivated them to learn*, 50% felt *They showed some interest in it*, 12.5% said *They were indifferent to it* and 8.3% chose *They said it was not important to them.* In addition, teachers were given a series of statements and asked to comment on any effects they had noticed on the least and most able learners as a result of the exam. Table 5.4. shows the results.

Table 5.4. Initial Washback Study. Responses to two questions to the teachers: *Did you notice any effect on (a) the less able learners (b) the most able learners because of the new exam?*

	Less able learners (%)	More able learners (%)
They asked for help more often	27.3	6.2
They appeared more engaged in lessons	13.6	37.5
They seemed to be trying harder	27.3	28.1
They asked more questions in class	9.1	18.7
Nothing seemed to change	22.7	9.4

Teachers suggest that the examination had had a positive effect on many of their learners. This is particularly noticeable with the more able learners who were reported as being more engaged (37.5%) or trying harder (28.1%). The less able learners were found to request help more frequently (27.3%), to be trying harder 27.3% and to be more engaged (13.6%). However, 22.7% of the teachers declared that there were no changes. This appears to be

an indication of the examination increasing the motivation of many of the learners, suggesting a positive washback effect.

Asked to choose between three statements about the effects of the examination, 71.4% of teachers chose *The foreign language exam in gimnazjum had a positive effect on my learners and on my teaching.* 19% declared it had had no effect and 9.5% that it had had a negative effect.

75.2% of teacher respondents said that they would make changes to their teaching in the coming year as a result of the examination.

While this is a very small sample, the indications are that the new examination had had an effect on these teachers in terms of their choice of materials, their planning and, in a few cases, the type of exercises done in association with reading and listening. The teacher themselves declared that this effect was positive. Teachers also reported an effect on learners of different levels of ability, and summarized this effect as being positive.

To conclude this description of the Initial Washback Study, we observe that there appears to be evidence of change taking place, following what is reported by teachers. Findings from learners are, however, less clear-cut. There is some slight indication that the productive skills of speaking and writing may be less frequently practised and assessed, but the evidence is far from conclusive. Indeed, what exactly is assessed is unclear, although it would appear that grammar is given priority. If, in fact, productive skills are being assessed less frequently, we have no information from the data from the Initial Study to explain why this is the case.

5.3. Phase III: The Impact Study

The following sections describe data obtained from Phase III the Impact Study, conducted in 2012. This mainly comprises data from interviews with teachers, school principals, and learners, which is supplemented with information extracted from questionnaires given to learners and teachers.

5.3.1. Teacher interview data

The data from 301 interviews was analyzed from transcriptions, using procedures described in Chapter 4. This included initial open coding to identify broad topic areas. Data from each of these topic areas were then re-examined and, where appropriate, more detailed codes were created. These more detailed codes were then counted for frequency to allow a first, overall picture of the data to emerge, based on quantitative analysis. This will be reported first. The coding process also included highlighting and extraction of fragments of the interview texts. Selections were then made from these excerpts, to find examples which could be said to typify each of the detailed codes. These are reported in the second sub-section, to provide a deeper understanding of the phenomena under consideration.

The frequency of occurrence of each of the codes will now be reported, code by code.

Code 1: choice of book relates to exam

214 of the 301 teachers interviewed (71.1%) claimed that they chose the book or books because of the exam.

Code 2: clash between core curriculum and exam content

33 teachers (10.96%) from the data set (n=301) referred to a clash between the core curriculum, which they specify includes targets on speaking, and the exam content, which they regard as not testing speaking. All of these teachers gave the same reason for their concerns.

Code 3: extending syllabus focus under influence of exam

14 teachers (4.65%) mentioned extending the syllabus they use in their teaching. Of these, 8 extended it to include the teaching of writing, as this was a new addition in the 2012 examination (see chapter 3); 4 extended the syllabus to include more grammar, explaining this was a result of the new Use of English section in the exam, or the new extended level. One teacher described extending the syllabus to include both writing and grammar, and one teacher described extending the syllabus to include "all skills", which may relate to the inclusion of writing, but we cannot be sure.

Code 4: aim of teaching is exam related

187 of the 301 teachers (62.12%) mentioned the exam in response to the question: "What are your aims in teaching English in *gimnazjum*? What is most important for you?". Of these 103 (55.1%) claimed that they had dual aims, with developing learners' communicative skills in the first place and preparing them for the exam second. 50 of this 187 (26.73%), by contrast, declared that the first of their dual aims was first to prepare learners for the exam and only secondly to develop other skills, which were most commonly described as communicative abilities. 34 of the 187 (18.18%) gave a single aim, which was to prepare learners for the exam.

Besides the main aims of teaching, the examination was also given as an aim for additional classes organized under article 42 of the Teacher's Charter (39 cases). In some instances (16 of the 39 cases, 41%), these were described as additional classes designated for exam preparation in class three, but in 9 of the 39 cases (23.08%), these were remedial coaching for learners having difficulty, specifically focused on exam preparation. The remainder (14 cases) were specified as exam preparation classes open to all learners.

Code 5: exam results are important

96 of the 301 teachers interviewed (31.89%) mentioned the importance of the exam results. Reasons given varied and can be divided into three

categories, labelled 'social conscience', composed of two codes where (a) the teacher explains the results are important because of their significance for the pupil, as they help to determine which upper secondary school they will enter, and where (b) the teacher expresses a personal sense of involvement in/responsibility for/ learner test outcomes. The second category was 'accountability' which included three codes: (a) the importance of the results for the school's reputation, (b) teachers report they are assessed on the basis of learners' test scores, and (c) teachers report school policy requires them to analyze the results of their learners. The third category was 'other' and was used to code miscellaneous responses.

The 'social conscience' code was applied in 34 instances out of the 96 (35.42%), the 'accountability' code in 46 instances (47.92%) and the third 'other' code in 16 instances (16.66%). The breakdown of frequency of sub-codes is shown in the table below

Table 5.5. Code 5 'Exam results are important': Axial categories and sub-codes

Category	Sub-Codes	Frequency (%) *
Social conscience	Exam results count for school selection	22 (22.9)
	Teacher expresses personal involvement	12 (12.5) **34 (35.42)**
Accountability	Scores important for school reputation	9 (9.37)
	Teachers are assessed on basis of results	14 (14.58)
	School policy requires analysis of scores	23 (23.96) **46 (47.92)**
Other		16 **(16.67)**
		n=96

* In the frequency column counts and percentages in bold indicate the total for the category

Code 6: assessment changed because of exam

Six instances of this code were noted. Of these 3 (50%) related to written work being assessed for the first time in response to changes to the exam, 1 to the intensity of grading changing, with the number of grades increasing in class three, in response to the new extended level of the exam. 1 stated that the revised exam now determines what is assessed in class one and the last 1 that speaking is less frequently assessed than previously.

Code 7: planning of work is guided by the exam

146 of the 301 teachers interviewed (48.5%) mentioned that planning of their work is guided by the exam. The three largest sub-codes within this were:

- Everything is focused on the exam 38 (26.03%)
- We make sure the exam syllabus is covered 32 (21.92%)
- We include test tasks in lessons 31 (21.23%)

Other explanations were much less frequent and were grouped as follows:

Exam focus is introduced gradually	9
Class 3 focuses on the exam	7
Extra hours are organized to ensure syllabus can be covered	6
Corrective programme based on analysis of exam results is implemented	6
Areas of focus are selected from the exam	5
Programme for teaching writing was introduced	3
Programme of mock exams is followed	3
Diagnostic tests based on exam are used, analyzed and form basis for focus of work	2
Exam tasks are used in formative tests	1
Test-taking strategies are introduced	2
EU funded innovation (3 year programme) to prepare learners for the exam	1

Code 8: exam affects work throughout all three years of school

251 of the 301 teachers interviewed (83.39%) claimed that the exam affects work throughout all three years of the *gimnazjum.*

Code 9: assessment is planned to the exam

The difference between this and code 6 should be made clear. Code Six was only applied when the teacher specifically referred to changes being made in assessment. Code Nine was used when the exam was referred to in relation to assessment, but no reference to change was made. It will be seen that some of the aspects mentioned under Code Six re-occur, but this is not an error. Code 9 was applied in 36 instances (11.96% of the 301 teachers). Most frequently it was used to refer to mock exams being used (17 cases, 47.22% of the 36 instances). There were 7 instances when teachers referred to what is in the exam being assessed (19.44%). 5 teachers stated that they use the rating criteria from the exam for assessing writing (13.89%). 5 teachers design test tasks in formative tests to match exam tasks (13.89%). One teacher explained that writing is assessed because of the revised exam and one teacher told how assessment intensifies as the exam nears.

Code 10: speaking is assessed

In the definitions of the codes it was mentioned that this code is different in nature to the others used. It was decided to count instances where the teacher specifically mentions that speaking is assessed to check the hypothesis that speaking might be less frequently assessed, for the reason that it is not directly tested in the exam. However, application of this code proved difficult. The first reason for this was that it was often not clear if what the teacher was referring to was actually speaking. They used terms such as

'responses in class', and 'answers' to individual questions, where the focus of the assessment could equally well have been other than the spoken skill. Such instances were coded as 'not mentioned'. The code was applied when the teacher specifically referred to "spoken production" ,"spoken interaction" or "speaking," although there were cases where the teacher then qualified "speaking" as referring to "reading aloud, when I check pronunciation" which meant that the code was not applied. The second problem was that the teacher was asked to explain what they gave grades for, which in some cases led to shortened responses, such as "all four skills", without information as exactly what was being referred to. Some teachers grumbled that it was a lot to remember, and others were somewhat uncooperative and rattled off seemingly random elements. In this way we see that the code itself was problematic, and the question used to elicit information on this also somewhat dubious in value. As a consequence, frequency derived from this code should be treated circumspectly.

183 of the 301 teachers (60.8%) were found to assess speaking. There were 3 instances when teachers specifically stated they did not test speaking, one as the group were very weak and the teacher did not want to demotivate them, and the remaining two as the teacher did not want to cause students "unnecessary stress". The other teachers were classified as not mentioning testing of speaking, with all the difficulties entailed which have already been described. Where teachers did state that they tested speaking however, there were mentions that the teacher was operating within the school assessment system, which centrally established the weight, or frequency of the assessment of different areas, implying that the decision was not the teacher's to make. Where a weighting system was in place, more weight was most often given to written tests, and speaking was valued as of secondary importance. The reason for this was not related to the exam, but to the assessment system itself, which was most often a school system applied to all subjects. In 6 other instances teachers referred to testing speaking, but added that this took place once or twice a semester, (where a semester is approximately 18 weeks, with classes taking place 3 times weekly), citing logistic difficulties relating to the time it required as reasons for this. To conclude, the testing of speaking in formative assessment is an area requiring further research and for the purposes of this study this code will not be given further consideration.

Code 11: aim is to meet expectations of others which relate to the exam

There were 7 instances of this code. In 3 cases the teacher ascribed the reason for focusing on the exam as being to meet expectations of parents and pupils; in 3 cases the teacher referred to pupil expectations as being the reason, and in one case the teacher cited it being the school principal who placed primary importance on the exam.

Code 12: course book focuses on the exam

85 of the 301 teachers (28.24%) referred to the role of the course book in relation to the exam. This was either in response to the question "What do you think of these books?" which came after the teacher had described the books they use throughout the three years of *gimnazjum*, or in reply to the question "What role does the *gimnazjum* exam play in planning and choice of materials for teaching English in *gimnazjum?*" Of the 85 responses, 75 (88.23%) described the course books as introducing the types of tasks which appear in the exam, containing sections dedicated to the exam, or as introducing the exam gradually. In 10 instances (11.76%) teachers appeared to feel coerced by the book into introducing exam-related material. By contrast, many of the other teachers positively asserted their satisfaction that the exam is introduced systematically throughout the material.

Code 13: narrowing syllabus to fit exam

This code was applied 12 times (4% of teachers interviewed). In 10 instances (83.33%) teachers referred to a loss of speaking in lessons because of the fact that it is not directly tested in the exam. This was used to refer to a reduction in time spent on speaking activities, or a diminished focus on speaking in the course. In 2 instances the teachers referred to restricting the syllabus focus in the case of weak groups.

Code 14: role of the exam

Teachers were not asked about the role of the exam itself. Responses given this code appeared spontaneously when teachers were talking about their aims in teaching in *gimnazjum*, or in response to the question about the role of the exam in planning and choice of materials in their school. The code was applied 34 times (11.3% of teachers). Responses were sub-divided into 7 sub-codes as described below.

The exam is:

- Closely related to the core curriculum 8 (23.53%)
- A threshold to cross 5 (14.7%)
- A passport to the future 4 (11.76%)
- A showpiece of the learner's achievement 4 (11.76%)

The exam:

- Shows the effects of the work of the teacher 5 (14.7%)
- Checks the work of the teacher 5 (14.7%)

We could also group these into categories referring to the exam from the learner perspective (a threshold to cross, a passport to the future and a showpiece of their achievement) and those referring to the exam from the perspective of the teacher (shows the effects of the teacher's work, checks the teacher's work, and is related to the core curriculum). In this case, note that the teacher perspective outweighs the learner perspective by 18 instances to 13.

Qualitative analysis

The aim of qualitative analysis is to enable us to come to a deeper understanding of the nature of the influence of the external exam in English on the work of teachers in lower secondary school and to obtain insights into the mechanisms of how washback works.

Let us begin by giving examples of what teachers said, to illustrate the aspects investigated in the coding. Here, rather than going through the codes chronologically as they were given in the previous section, the codes will be grouped into themes, with the aim of obtaining further insight into the questions posed in the previous paragraph. The first theme is Beliefs and Attitudes of Teachers relating to the exam. Gathered under this theme are codes relating to the aims of teaching (code 4), the role of the exam (14), and clashes between the core curriculum and the exam (2). Once teachers have determined their attitude to the exam they may, in response, decide to narrow the syllabus (13), or extend the syllabus (3), so these codes will be linked as a sub-section of this theme. The next theme is Planning of Work and Choice of Materials and their relation to the exam. Here we will group the codes relating to planning (7), choice of book (1), "book features exam-related material" (12), and "exam affects all three years of school" (8). The next theme is Assessment in School and the Exam, which will include "assessment is planned to the exam" (9), "assessment changed because of the exam", and "speaking is assessed" (10). The final theme is Levers, and this aims to investigate pressures which teachers indicate they feel on them to focus on the exam. This includes "aim of the exam is to meet expectations of others" (11) and "exam results are important" (5).

Theme One: Beliefs and Attitudes of Teachers relating to the exam

Before we start to illustrate the codes associated with this theme with examples, we need to remind ourselves of the context in which the research is taking place. As we saw in Chapter 3, teachers in lower secondary school are legally required to cover the core curriculum, which gives as its main aim for the end of this stage of education that learners achieve "effective communication in the foreign language in speaking and writing" (MEN 2009: 42). Schools in the year 2011–2012 were required by local educational authorities to report on coverage of the core curriculum. From data gathered from school principals, we know that all the schools in the sample were taking measures to monitor and report on coverage of the core, which involved teachers being required to complete forms, registers, or schemes of work, with detailed reference to the core standards and items contained in it (See Paczuska, Kutyłowska, Gajewska-Dyszkiewicz, Ellis & Szpotowicz, 2014). In this way we can be confident that all teachers interviewed would be fully aware of this document and its importance.

The FL external examination, as we have seen in Chapter 3, purports to test all the skills and knowledge in the core curriculum, although speaking is tested indirectly. The exam was revised in 2012 and now includes two levels, basic and extended, which are designed for all learners following the continuation programme III.1, for those who started English in primary and are continuing it in lower secondary. Of importance is the fact that up until 2011, points from the exam did not count towards selection for upper secondary school. From 2012, for the first time, points will count, but only those obtained in the basic level test. The interview data in the project was collected in spring 2012.

As we saw in the report on the quantitative data, more than half of the teachers interviewed (62.12%) included the exam in describing their aims in teaching in *gimnazjum*.

Let us now look at an example of what one teacher said in response to the interview question: "What are your aims in teaching English in *gimnazjum*? What is most important for you?".

Extracts from all interviews are given in translation (done by the author) but follow the structure of the interview, as far as possible verbatim. These are transcripts of spoken utterances and so include false starts and reformulations and are often ungrammatical.

School 8 Teacher 1 is very aware of the examination. She describes it being what they have to focus on, as it represents "the final effects [of our work] which are evaluated." She states "as a teacher most important for me is good preparation for the test and giving a strong basis for continuing learning after this." However, she also stresses the importance of communication, "because let's say it's the most important skill for the kids, as wherever they go on holiday, where we have Comenius, we have all sorts of other things, so somewhere there has to be communication in English" (this school was involved in a European Comenius project and school exchange with the Netherlands.) She states that the exam plays a "very large" role in planning, "as it is an external test of skills, which show the extent to which the school prepared [the learners]" and adds "so, if it comes to English, I think that it's like this in most schools, unfortunately it's [the exam is] the most important thing."

She considers that the exam underlies much of their planning in school:

> Because really all the books, all the tasks, and also our in-school diagnostic tests, all of it prepares them mostly for the exam. Because if we didn't do that we would be shutting the kids off from the road ahead, to the best schools, so it is the most important thing for us too.

If it comes to her attitudes towards the exam, she has given it priority in her aims but we see that she appears to be placing importance on it for two reasons. First from the learner perspective, as she appears to associate solid preparation for the exam with helping learners get a firm foundation for

learning in the future, and giving them the best possible opportunities for getting to good upper secondary schools, and secondly, from an institutional perspective, concern for external evaluation of the school on the basis of the test results and, by implication, of her own work, She seems, however, equally aware of the importance for the learners of being able to use the language in real life situations. Here use of the word 'unfortunately' when describing the exam as the most important thing, may suggest that there is some dissonance between her personal beliefs and the ethos of the school. We can see something of the dilemma in which this teacher, like many others, finds herself. On one hand, she has what could be described as sound pedagogic and practical reasons to focus on communication, while on the other hand, her social conscience tells her that she has an obligation to her learners, whose futures depend to some extent on the test results, and also that she has an obligation to her employer, the school, to get good results.

Dual aims for teaching

We saw from the quantitative analysis that many teachers had dual aims. Here, in response to the question about her aims, School 55 Teacher 3 explains:

> Sort of two-track you could say. For sure, what is a very important issue for everyone learning English whether it's in primary school, in *liceum*, or if it's an adult, that's communication in the language, the skill of using the language, and, particularly for lower secondary learners, practicing those skills which are tested in *gimnazjum* exam… I want to teach the pupils so that if they went to England they would be able to use the language and be independent. That's the first priority. So that they understand what someone is saying to them, that they could respond, and, of course, that they are also able to use the language properly, and write the *gimnazjum* exam well, that's why it's sort of two-track.

Like the previous teacher we see that she talks about wanting her learners to be able to use the language in everyday situations in a real context, but we see that this teacher places communication first, as her primary aim, followed by being able to do well in the external examination.

Other teachers find themselves challenged by this sense of there being dual aims.

> Most important for me is to prepare the learners well for the *gimnazjum* exam. And there's a bit of discrepancy, because the *gimnazjum* exam is only written. So we focus on teaching grammar, vocabulary, on doing typical exam tasks. But there's not enough time for communication, because there's no oral exam. So there's always a little communication, but it's sort of pushed into the background.
>
> School 39 Teacher 2

The "discrepancy" mentioned by this teacher (39/2) appears to refer to discrepancy between the aims of the core curriculum, which are for learners

to be able to communicate, and the examination, which has not direct test of speaking, although this is not overtly stated. What is clear, however, is that the teacher responds to what she perceives as the aims of the exam, by narrowing the syllabus, and reducing the focus on "communication".

> I'm in a serious dilemma, because on one hand most important for me is that any student, regardless of whether it's *gimnazjum* or older, or younger, is able to communicate in the language. Yes, but communication isn't really the most important thing for the learners. So communicating, communicating is key for me, but I am evaluated, I can't escape the fact that the grades the learners get and their scores on the exam, the measure of my work is the exam. So I constantly have this dilemma between choosing what I think is the most useful and important exercise, for example. I would like to be able to practise speaking more often, and I can't do that because I am mainly evaluated on writing.

School 68 Teacher 1

This teacher (68/1) expresses the frustration of being torn between what she feels she should be doing, which is teaching learners to use the language in communication, and knowing that her work will be assessed on the basis of exam results, which she sees as primarily a test of written skills.

The problem of the exam focusing more on the written form is referred to by another teacher, but approached rather differently:

> In the exam there's no speaking, and in life it's speaking that comes first. So we have to really plan our work very well, so that they'll be prepared for real-life and at the same time prepared for the exam, to join the two together.

School 65 Teacher 2

The teacher (65/2) sees the "discrepancy" as a challenge to be overcome by effective planning of her course and looks for ways to combine both.

The next teacher shows quite a different attitude:

> I think that the exam is really constructed quite well, yes from the perspective that it gives learners such practical, life-skills. So I think that if I do exercises with them on speaking it is not a problem for the *gimnazjum* exam, rather the opposite. Because quite a large part of it is on communication, but it's tested in another way. So we also do lots of listening exercises, because of the *gimnazjum* exam, and speaking. I think that is preparing them for the tasks that test communication.

School 104 Teacher 3

Teacher 104/3 appears to have taken time to consider the exam content carefully and concludes that it is testing practical skills. There are tasks in the test which focus on communication "tested in another way" (i.e. indirectly) and she has interpreted that her learners will learn how to do these tasks through practising speaking, so for her, unlike teachers 39/2 or 68/1, there is no conflict.

The next teacher (92/1) also sees no problem, as she has understood the relationship between the core curriculum and the exam.

> The gimnazjum exam is based on the core curriculum, so they are *connected*. Planning materials, planning work I base on the core curriculum and, I'm telling you, the two things connect, so it's not that I'm teaching to the exam, I'm simply teaching on the basis of the core curriculum, on the basis of the programme which is adapted to teach that and also to prepare for the exam.
>
> School 92 Teacher 1

> I have to admit that I don't organize my teaching to the exam….I just teach what I think is important. Obviously I have to cover the core curriculum, the core has to be covered before the exam. Because there might be a grammar question which the children could come up against in the exam and they would resent me. But besides that, as I said and I'll say it again, I don't teach to the exam and still the children pass the exam. They pass it quite well. I was very pleased with how the mock exams went this year, both the basic and the extended levels. How the real exam went, we'll see.
>
> School 77 Teacher 2

Teacher 77/2 also sees the relationship between the core and the exam, although she does not overtly state this. She sees her task as to cover the contents of the core curriculum and reports that this is effective, with her learners obtaining satisfactory results in the exam.

We can see that teachers' attitudes to the examination appear to be influenced by how they perceive the relationship between it and the core curriculum and how they respond to this. We shall return to this important issue in the discussion in the next chapter.

Learner aims and motivation according to their teachers

Teachers provided interesting insights into what they consider is important for their learners in lower secondary school. One of the trends to emerge concerns the relationship between learner aims and their motivation.

> Interviewer: What in your opinion is most important for the learner in learning English in *gimnazjum*?
> I think that the aim which the learner sets themselves. Or which the teacher sets for them. If a learner has an aim that, who knows, they can achieve, it is in some way important for them and it certainly motivates them to learn.
>
> School 63 Teacher 3

This teacher (63/3) indicates a factor which we will see has bearing on attitudes to the examination. She points out that having a clear and achievable goal in learning motivates students.

> They think about upper secondary school here, whether where they will go that language will be needed or not. I see very little that it's [needed] in life, or if they think about it for future work. That seems still very far away. I think it's very far away. More from the point of view if it will be useful in upper secondary than what further with that language. Sometimes I try, so that they'll think, look how it is in work, it's difficult, the job market is as it is. But for them it seems that, when I say that maybe you'll go away somewhere and that language will be really necessary for you, they just look at me. It seems to them that it's the very distant future. At the moment they are thinking, what will happen if I take that exam, and later, upper secondary. I think that that is what concerns them most for now. I have that feeling. That's what I think.

School 65 Teacher 2

Teacher 65/2 explains her interpretation of the learner's perspective. Although she tries to interest them in long-term aims, such as needing the language for future employment, or travel, her learners seem to be fixed on the near-future and have more immediate concerns, such as the potential usefulness of the foreign language in the next school and taking the FL external exam.

The next teacher appears to have similar views, but takes a different approach

> Of course I would like them to know the language as well as possible, so they would be able to make themselves understood, that they are simply aware that they can cope on a trip abroad, on holiday, maybe at work. I also help them understand that if they, for example, hear about a job advertisement in the Employment Centre, even here in Poland, trying for a position, knowledge of a foreign language is required, and even more than one. So my aim is to prepare them for adult life, from the point of view of foreign language. And here in the short-term I look at how [to get] the best grades, the best results in the *gimnazjum* exam.

School 94 Teacher 2

Like the previous teacher, 94/2 is concerned with helping learners understand the importance of the foreign language for their futures. However, she appears to view getting good results in the exam as a step in that direction, and sets that as an immediate goal.

> Theoretically they understand that they have [English] in the *gimnazjum* exam and that the points at this moment are counted towards recruitment [for upper secondary], so I think for them it's important to be well-prepared for the exam, right? Some of them consist, they also understand that it gives them the possibility to travel abroad, because, for example, they have parents who have gone, right? Or brothers, or sisters, and they see that it is really useful, right? But there's also a little group who simply treat it as another subject and that's it. They do the minimum or less.

School 116 Teacher 2

This teacher (116/2) reminds us that lower secondary is a school for everyone, with no selection. As a result the youngsters are very different from one another and also come from different home backgrounds. She suggests that this affects their attitude to English and also their motivation. For some learners, English is simply "another subject" to be learnt and is unrelated to life in the future. Even the external exam is not enough to motivate them.

This teacher (116/2) adapts to her learners by adjusting the aims of her teaching according to the level of the group

> It depends on the group. If it's a weaker group then, for sure, the main aim is to prepare them for the exam, so that they write it as well as possible, even though it's a less advanced group, so to make up for what is missing from primary school, and to prepare them well for the exam. If it's a stronger group, then apart from the exam, there's also communicating. Here I place the emphasis so that's, it's not so that they write the exam nicely for me, but that also later in the upper secondary school they really have a foundation, right? So that they're not afraid if someone stops them and asks them to say something, right?

School 116 Teacher 2

This teacher (116/2) has made a conscious decision to narrow the focus of aims in the lower ability group to the exam alone. With the more able group she teaches communicative skills, in addition to preparing the learners for the exam, with the longer-term aim of preparing them for the challenges of upper secondary school as far as speaking are concerned. This is purposive, selective narrowing of the syllabus, according to the language ability of the learners as perceived by the teacher.

This takes us into the next thematic area, relating to planning of work and choice of materials.

Theme Two: Planning of work and choice of materials

The attitude the teacher has towards the exam, its role and its importance, seems to interact with other factors to determine the approach they take in their teaching. Some factors which were identified in the data, and which appear to function as variables, fall within the theme of planning of work.

The level of the learners and planning decisions with regard to the exam

One strong factor which emerges is the question of how the teacher perceives the level of the learners and their potential with regard to the exam. The level and ability the learners are perceived to represent appear to have a strong influence on teachers' planning decisions. The first teacher describes the situation of young people entering class one lower secondary:

> In the first class, my main aim is to even out the level, because kids come to us with very different levels, and most often they are packed into one group. Despite the fact that they say, whether it's a class or a group, that's, it's a continuation of learning, there are kids who, despite the fact that it's a continuation, know practically nothing at all. There are also kids who thanks to that continuation, could have also had some private lessons, I don't know, maybe the school did everything, and they're at a very high level, and now joining those two is a challenge in the first class, right? The next thing is that the difference in level often is not the result of some difference in the pupils' talents, is it? It isn't that some are worse because they're less clever, but only because they've, they have come from some sort of worse preparation. Generally, my aim is to prepare learners for the *gimnazjum* exam and in the long-term also for *matura* [school-leaving exam], so to ensure they get the basic, necessary grammatical and lexical material and the communicative skills which are required at the end of lower secondary school and which later will be really useful in *liceum.*

School 58 Teacher 2

This teacher (58/2), unlike the majority of the teachers in the data sample, works in a school which does not have streaming into groups for language ability, and is faced with the challenge of learners of very different levels of ability in one group. Nearly all the teachers in the sample refer to the question of learners of differing levels of ability, but this is most frequently in terms of groups of different ability levels. The reasons this teacher surmises for the differences appear to apply across the data set, with the additional factors of home background and parental attitude, parental aspirations and engagement in their offspring's learning, which are also mentioned by teachers as contributing to the pupils' differing levels of language attainment. The challenges of catering for these differences lead teachers to various planning decisions.

> Unfortunately lower secondary ends with the *gimnazjum* exam, and everyone certainly somehow works so that the young people pass the exam, and certainly the teaching is aiming in that direction. I'm pleased that I have quite good children, so they will certainly pass the exam. So here, I am trying to work on the development of their different skills. It's not only strictly directed towards the exam, it's not that we only do that and nothing more….if it comes to a more advanced group, then I don't really take the exam into consideration at all. Because if I wanted to take an exam preparation book [pol. *repetytorium*] then the kids would simply stay in the same place, in terms of language development. But, in the current second class I have a group which is a bit weaker, and there I take that into consideration, so I will have to take the exam preparation book and do some deeper preparation for the exam in class three. So really it all depends on the group and I'm sure that there is no sense adding an exam preparation book by force, for groups that are quite strongly advanced, if it comes to the level of the exam.

School 112 Teacher 1

In describing their choice of books for lower secondary school, many teachers refer to the selection of an exam preparation book for class three. These books tend to follow a similar format, regardless of the publisher. They are arranged thematically, according to the topic areas outlined in the core curriculum, and include large numbers of exam type tasks in all the skills, clearly labelled basic or extended, according to the level of the exam. Some also offer strategy training and offer tips for how to get good marks, explaining, for example, how the open writing task is marked and what the rater is looking for. Teacher 112/1 explains that the decision whether or not to choose such a book for class 3 depends on the level she perceives a group has reached. Already she has decided that the group she is teaching in year 2 will need such a book in class 3, as she sees them as being weak and needing specific exam preparation. By contrast, she rejected the choice of such a book for her current class 3, on the grounds that it would hold them back. Her belief appears to be that if learners are already at an appropriate level, then they will cope with the exam anyway, without specific preparation, while weaker groups need specific exam support. We will return to this factor in the next chapter, when considering the mechanisms of how washback comes into play.

The next teacher also refers to the level of the group, but this time in relation to how she treats the course book.

> Let's put it like this, if you have a course book, yes? And the core curriculum for that, then it's easier to plan the whole cycle for three years, isn't it? To prepare them for the exam. Anyway, I won't hide the fact that it depends whether I've got a stronger, or a weaker group. More advanced, or less. So either I stick to the course book more faithfully, if it's a less advanced group, because then I am sure, right, that I'm doing with them what I have to do to prepare them well for the exam, which is my priority at the moment. But with a stronger group, I won't hide the fact that I also photocopy a lot of things for them, I look on the internet so that they can also develop their speaking more.

School 116 Teacher 2

The attitude of Teacher 116/2 is typical of a large number of teachers, for whom the course book is the mainstay of their course. She is confident that following the book (or series of books) will mean that the contents of the core curriculum will be covered and so learners will be ready for the exam. The book becomes the primary (or sole) source of material for a less advanced group and appears to drive planning decisions, while for a more advanced group, the book is a starting point to which additional materials may be added.

Dependence on the book, or even belief in its power, feature frequently in the data set. The next teacher, like the previous ones, speaks of the differences between learners, but refers to another aspect of this:

> Many pupils work systematically and really put a lot, a lot of effort into it, but in general they don't have that level of language knowledge yet, because that needs a lot of work on your own, generally a language needs a lot of work by yourself, and if they don't revise, then it escapes them. Apart from that, it's difficult to compare learners who have never had extra lessons, and many of them have. They are always in a different situation, they have more hours, so it's easier for them. But generally the thinking is always that, what happened in other schools, what happened here between classes, how that affects grades, so that there aren't such big differences. So, because of that, we have the exam preparation book. It has tests very similar to the exam tests, and so I think that the differences will even out.
>
> School 68 Teacher 2

Teacher 68/2 points out that students approach foreign language learning with differing levels of engagement. She believes that foreign language learning does not happen only in class, but requires independent work from the learners. Learners who have access to extra lessons outside school, she believes, have an advantage, as this exposes them to more contact with the language, apparently implying that they do not need to be so independent in their learning. It would seem that she believes that learners who only have English in school need to make the effort themselves to revise the material at home. She appears to imply that this is not happening in many cases. Despite the differing levels of access to English, the school policy is to even out the level for all learners, regardless of their situation and she believes that using an exam preparation book will do this. The thinking appears to be along the lines that solid revision in class, on the basis of the exam preparation book, will compensate for the fact that learners make differing amounts of effort and have differing exposure to the language outside school. She expands:

> …The most difficult is with those creative tasks, because there is little [time]. You have to know [something] already and then you have to transform it and it's on the whole knowledge [i.e. grammatical syllabus], you can come across different things. It's also difficult for them to get hold of all this knowledge and remember it. All the more because, as I said, our young people are not used to working independently at home, it's very hard. As I said already, I force them with short tests, there's something else, to explain to them that they have to work for themselves, and you'd be surprised, but in the end it works for them.
>
> School 68 Teacher 2

In this excerpt the teacher (68/2) is referring to the semi-open transformation tasks which appear in the extended level exam. They have a grammar focus, and can be based on any grammatical item in the syllabus for the whole educational stage. The teacher indicates that first the learners have to 'know' the grammar points being tested and then apply a transformation, suggesting that she means learners must understand and recognize what is being tested and know how the same meaning can

be expressed differently. This requires higher order analytic thinking skills and the ability to apply knowledge productively in a new context. Her concern is, however, not for the cognitive aspect involved, but for the learners to have internalized the material and be able to use it. In order to push her learners to do this, she uses frequent short tests and explains that the learners have to revise by themselves at home. She claims that she has found that this strategy, of using tests and explaining to the learners, brings good results.

The overt testing of grammar was introduced in the *gimnazjum* exam for the first time in 2012 and so constitutes a new challenge for the teachers who are the subjects of this research. Teacher 68/2, in her description and explanation, is grappling with how to encourage and support her learners, in order to enable them to be able to deal with these "new" exam tasks. While it is the new exam which is driving the change in her teaching, we can see that her approach has the potential for positive change in the learners, by encouraging them to be more independent, to be more systematic in their learning and to apply knowledge in practice.

The fact that different learners are at different levels is, however, approached quite differently by some teachers.

> In some groups we have to practically throw grammar out, because they will never really learn it. It would only demotivate them. So it depends on the group. We select the contents, I select the contents which they'll be able to get, so, for example, with a weaker group I pay attention to reading and listening, more than to language functions, or grammar. But in the strongest groups, of course, on everything, so the exam makes itself felt in the teaching process.
>
> School 23 Teacher 1

Teacher 23/1 describes narrowing her teaching programme with less advanced learners, consciously choosing not to work on the areas which require language production (tasks tested in the extended level) in favour of a focus on reading and listening. Although not stated, this appears to be a pragmatic decision to focus on the skills which are allocated the largest number of points in the exam. The teacher, however, rationalizes her choice as concern for her learners, not wishing to demotivate them by facing them with tasks she thinks they would not be able to do. In contrast to Teacher 68/2, who encourages independent learning, this teacher (23/1) is autocratically making all the decisions for her learners.

> I would prefer that the *gimnazjum* exam included an oral part so that I could develop speaking more. But here I don't have time, to be honest. If it comes to class 3 I have to, to put it simply, speed up and finish the book, that's the exam preparation book [pol. *repetytorium*], in which there are 15 sections. So I focus on that. I focus on listening, on reading and on communication [pol. *reagowanie językowe*], plus, this year, happily also on writing. On writing short letters. And unfortunately I can't, I don't have a chance to practise speaking at all. In

class 2 they spoke quite a lot in English, but in class 3, practically nothing. Not until after the exam will I have an opportunity to practise that skill.

School 103 Teacher 1

This next teacher, 103/1, is an extreme example of the teacher for whom the course book is the syllabus. Here she explains that, because of pressure of time to finish the exam preparation book, speaking has been cut out of her course in year 3. She does, however, intend to practise speaking "after the exam", that is, in the remaining two months of the school year. It is interesting that she does not see any relation between the exam tasks in "communication" and speaking. This is not uncommon in the data set. We shall return to this important issue in the next chapter when we consider the mechanisms by which washback comes into effect.

The extent of the influence of the exam on planning

The next theme to emerge is that of how teachers respond to the exam in terms of their planning. As we saw in the quantitative analysis, the choice of teaching material is strongly influenced by the exam. We have already seen in the previous section that many teachers choose to use an exam preparation book in class 3. Teachers also frequently cite the exam as a reason determining their choice of course book for other classes, as we see in the following extracts.

The *gimnazjum* exam is the key aim for us, so the course books are chosen so that they have the largest number of exercises preparing for the exam. Of course the books that we've chosen contain all the thematic material which is to be revised before the exam. The types of exercises, the tests prepared, especially in class 3, we work hard so that they will be compatible with the ones in the exam, that's very important.

School 79 Teacher 1

The teacher here (79/1) checks the syllabus needed for the exam is covered and is concerned that the types of tasks included in the book reflect the tasks which will be in the exam.

The course books are chosen for the new *gimnazjum* exam. The books were changed… because the books we followed were still for the old exam, without division into basic and extended. So the exam caused the change in the books. These changes, the books were already changed once, because in the meantime the new exam came in, with a division for basic and extended, but even before that the new core curriculum came in, didn't it? So here the exam was the key reason, the motive, to change the books. Because of the exercises which are divided into basic level, extended level, exam exercises, whole separate sections in the book, you see? Plus the whole construction of the book, the way in which it's organized, the vocabulary, the way the grammar items are chosen. They are divided into things that are obligatory in the exam, into the fifteen topics for vocabulary and it's always marked which topic area a word comes

from, if it's People, or School, Family Life, Entertainment. Yes, the exam plays a key role in the choice of book here.

School 80 Teacher 3

Teacher 80/3 clearly illustrates the thinking processes behind deciding to change the course book and shows unambiguously how the exam is a major factor in the process.

We choose course books, I obviously don't make that decision alone, but together with the other teachers of English and also with the principal. We take note of what the exam is like and choose the appropriate book with that in mind, so that it contains all the information, everything which the pupils have to learn in order to pass the exam well.

School 103 Teacher 3

Teacher 103/3 explains how the book is chosen after analyzing the exam and deciding what pupils need to pass the exam well. The exam is a major deciding factor in choice of course books in lower secondary school.

Let us now examine how teachers refer to the exam when describing their course planning and the roles it plays.

There's a growing tendency that to prepare for doing well in the exam you need to do tasks, tasks, tasks, so that they [the learners] practise them, so that they won't be stressed when they go to the exam…

We try to choose those materials which are connected with the exam, the types of tasks….and to show the learners the exam, what it really looks like. Because it's not the same in the course book, and later, that at least once in a while, we practise the exam format with them, in the first and second class. The form of the exam as it really is, so that they know what the answer sheet is, that they have to transfer their answers themselves, to practise that with them. And we try in fact to make sure that everyone knows what the exam is like, to show them, so they understand and so that it's useful.

School 93 Teacher 1

Teacher 93/1 describes a process of attuning learners to the exam tasks and the test format, with the aim of reducing stress when they come to the exam itself. This teacher draws attention to what she perceives as being potentially problematic for the learners, the fact that they have to first write their answers in the question booklet and then transfer those answers to a separate answer sheet. This attuning process starts early in the stage, in the first class. She also, like many teachers, believes that doing lots of exam tasks is the way to develop learner skills.

The next teacher also introduces the exam gradually.

The whole of our teaching is adapted to the *gimnazjum* exam, everything we do in lessons is typically applied later, so reading comprehension, listening comprehension, that's everything. I choose the tasks so that they are very

similar to those that are later in the *gimnazjum* exam. And in the Teacher's Charter hours we also do some tests.

Interviewer: Does this apply only to class 3 or to other classes too?

Teacher: We start with it from class 1, so that, it's not that, in class 3 we already have more material. In class 1 we start from basic things, fewer of them, fewer of those tasks. In class 3 and at the end of class 2 there are already more, to prepare them already and sort of get them used to the idea that the *gimnazjum* exam will take place soon.

School 58 Teacher 1

Teacher 58/1 reports that the entire planning process is geared to the exam. This includes not only the course content, but also the types of tasks selected to use in class. The exam focus also extends to the additional classes offered under article 42 of the Teacher's Charter, which she says are dedicated to doing practice tests. The intensity of exam focus increases gradually, starting from fewer and more basic tasks in class 1 and increasing as the exam draws nearer, illustrating what Watanabe (1996) described as the 'seasonality' of washback.

The next teacher illustrates a similar tendency, but in a somewhat milder form. He first claims that the exam influences planning mainly in class three, but continues:

However, in class 1 already we prepare for the exam indirectly. If we do an exercise which is similar in form to what is on the exam, I always inform the students, whether it's in class one or class 2, that they will meet similar exercises in the exam, so as to draw their attention to how to do that task. But it's in class three that we do the largest number of those exercises. Because they already have a certain level of knowledge and we can polish their skills.

School 73 Teacher 2

This teacher (73/2) also 'attunes' learners to the exam tasks, but apparently not by deliberately choosing to do lots of practice test tasks. Instead, he follows a general programme in the first two years of school, but takes time to point learners towards what is coming in the exam, by drawing their attention to tasks which they come across, which are similar in format to the exam tasks. He does, however, explain that class 3 includes a lot more test tasks, which he appears to justify by implying that at this stage learners have already covered the curriculum and so time can be spent polishing their skills and preparing for the exam.

It would be irresponsible to suggest that all the teachers in the sample are heavily exam-oriented in their teaching, as this is definitely not the case. Many take a more pragmatic approach, as illustrated in the next excerpt.

I'd like to say that during two and a half years we do material which covers the requirements of the core curriculum, following the course in the course

book, which is later reflected in the exam tasks. But in class three, from the new half-year, so somewhere in February, I try to go through material with the students, according to typical exam tasks, so that they can see, even technically, what it's like, that the knowledge which they have learnt so far will be required in the exam in this, or that form, because it's very often the case that, following the course, we do certain things in pieces, and in the exam they are needed as a whole. So in the last weeks before the exam I try so that the learners get the whole. I think that it's easier for them to see and get used to on specific test materials.

…In the first class simply talking to the kids about the exam, it's such a distant future that it simply doesn't get through to them. But somewhere in class 2, in the second half of the year, when they are already nearly feeling like third graders, we begin to talk, we begin to write with the exam in mind, because that is always the most difficult for them. Whereas, in the earlier stage of lower secondary they are still very immature teenagers and the perspective of two years for them is cosmic.

School 96 Teacher 2

Many of the teachers interviewed explain that it is the core curriculum which drives their planning. Some teachers believe that the relationship between the core and the exam is so close that there are no difficulties in preparing learners for the exam if the core is covered. Others, as we saw in the earlier section, feel conflict with the fact that speaking is tested indirectly in the exam, or even feel that speaking is not tested at all, which affects their planning decisions. Teacher 96/2 represents the type of teacher who believes the core and the exam are closely related, and for whom the core drives the course plan. She also places her decisions firmly in the context of the age group she is teaching, reminding us that at the start of lower secondary these are young teenagers aged 12–13. She explains that for them something which is to take place two years in the future is very distant, as from the perspective of a 13 year old two years is really a very long time. So, aside from the fact that she decides to plan around the core, she considers it not worthwhile even mentioning the exam in class one, delaying this until late in class 2, when the learners have begun to look ahead. Like several other teachers, she plans a course to introduce writing which starts in year two for logistic reasons, as learning to write letters takes time and she feels it cannot be left until class 3 to do this.

Other teachers also refer to logistics in their course planning, particularly with the introduction of the extended level test for the first time in 2012. Many see this as placing additional requirements on them, needing them to cover more material, to focus more specifically on grammar (tested overtly for the first time), to extend the scope of the vocabulary they teach, and to introduce the teaching of writing. It appears that analysis of the exam requirements, particularly those of the extended level test, is common practice and is impacting on teachers' planning decisions. A sense that there

is more to cover is used as justification for working towards the exam from classes earlier than year 3. The next excerpt shows how writing is dealt with in the three-year plan.

> We think about the exam, that, for example, different types of tasks appear that we have to get the first class used to, so they are worked on and already in the first class my colleagues and I introduce writing tasks, sort of in stages, because in class one, let's say there's an introduction, in the second class there's already pressure, because we are also practising tests a bit, and in the third class there are typical exam tasks, also developed, plus, of course, mainly preparation for the exam.

School 22 Teacher 2

Teacher 22/2 phases writing in gradually, starting from year 1, beginning with a general approach and becoming specifically exam-focused by year 3.

As the external exam in foreign language is relatively new for these teachers, with only the fourth administration in 2012 (and the first in the new two-tier format), there is also a sense that teachers are still adjusting and learning how to deal with it in their work.

> Interviewer: What role does the *gimnazjum* exam play in planning and the choice of materials for teaching English in your school?

> Teacher: The *gimnazjum* exam? It plays a big role, but I wouldn't say that it's deciding, because I already made several of those mistakes, focusing on the contents of the *gimnazjum* exam from the previous years, I crammed the kids, I prepared them for the test which had been the previous year, or two years earlier, the contents that were in them, and it didn't work at all. I only now understand that after X years of teaching. The tests always surprise you, at least foreign languages do, there's nothing certain in the contents of the *gimnazjum* exam, but the form, yes. The form of the tasks, we practise those and I always sensitize the kids to the form of the tasks. But the contents are different, and so different from year to year that it's hard to speak about teaching to the *gimnazjum* exam, to select teaching material for the exam, no. I think you have to do everything that is in the core curriculum, and only then can we be sure that we're preparing the kids for the *gimnazjum* exam.

School 101 Teacher 2

Here is a teacher (101/2) who has learnt from experience how *not* to use the exam in planning. Having attempted to teach strictly to the exam, on the basis of analysis of past papers, the teacher concludes that this is not the best approach. She attributes this decision to the fact that the exam contents are unpredictable and has settled on covering the core curriculum as effective and adequate preparation for the exam. What she does do, however, is train learners how to do the type of tasks that appear in the exam, acknowledging that, as these do not change, they can be included in her planning. We see that her attitude to the exam has changed in the four years, and changed quite

markedly, indicating that attitudes are dynamic rather than fixed. Here we are led to consider whether, in general, attitudes to the exam change as a teacher gains more teaching experience, or whether this depends on the individual. Or is the factor not general teaching experience, but experience of working with learners who are taking the exam? Is it perhaps coming to a deeper understanding of the philosophy of the exam, or becoming more aware of certain aspects of the exam that were not present earlier? We will return to these issues in the next chapter.

The final aspect to raise in this section is the question of response to the exam results. It is a requirement of the school, if, following analysis, the results of the exam in the school fall below the national average, that a "corrective programme" is prepared. This will be discussed in more detail in the interviews with school principals. A teacher here gives a simple explanation of the process:

> In our school the foreign language team analyze the results of the exams, these are the national exams mainly, but twice a year we also have mock exams. And on the basis of those results we decide what went well, what went badly and we try to correct what needs to improve. For example, we try and do more writing exercises in the lesson, if writing went badly. More listening exercises, if the questions, the tasks were not satisfactory. So we simply respond to the results of the analysis and we try to do more exercises on that topic during lessons.
>
> School 73 Teacher 2

The school is required to show how they intend to improve the results of the following year group. The strategy described by this teacher is to allocate more class time for and to do more exercises in the parts of the test that was found to be done poorly. She cites the example of listening, an area often identified as problematic in national reports from the Central Exam Board and theoretically predicts that writing, which featured in the exam for the first time, will need attention. This interview took place in April, even before the *gimnazjum* exam, and the detailed results for the school were available until August.

Let us move on now to the next theme found in the data, which is that of assessment and the exam

Theme Three: Assessment in School and the exam

We saw, in the quantitative analysis, that teachers mentioned the influence of the exam on assessment in school much less frequently than when referring to planning, or the choice of materials, and we should bear this in mind when illustrating this theme with examples.

> To be honest I assess everything. I do tests of grammar and vocabulary, sometimes separate little tests for grammar and separate for vocabulary. I also try and assess the skills. That means we always do it at the end of a chapter, apart

> from the tests of grammar and vocabulary, a test of listening comprehension, and reading comprehension is also assessed. It's similar with writing. To be honest, I would like to introduce that writing is also tested in the lesson too, because they write in the exam, but in most cases writing is assessed as homework. And the worst is with speaking, to be honest. I mostly grade them on the headphones, because they are very embarrassed about speaking out. I try and test everything.
>
> School 67 Teacher 1

This teacher (67/1) is fairly representative of many teachers in the sample, with the exception of her solution to the logistics of assessing speaking, which is novel. Following the core curriculum, she attempts to assess all its aspects. We see the influence again of the course book, with assessment coming at the end of a chapter, but it is not clear if the tests being referred to are those prepared ready for the book, or the teacher's own work. The difficulty with assessing speaking is related to the learners, who are reported as not wanting to speak in front of their peers, rather than to a decision to reduce the amount of assessment of speaking in response to the exam. This teacher has found an innovative solution, assessing learners in the language laboratory through headphones. She is aware that a change is needed in how she assesses writing, wanting to move it into class, rather than as homework, because writing is now included in the exam. Presumably there is concern about who is in fact doing the written task, or how much support the pupil getting in doing it, over which the teacher wants more control. This teacher reminds us of the problematic logistics of the assessment of productive skills in the classroom, which may also contribute to their being less frequent.

One aspect of the new exam appears to be that the assessment of writing may be on the increase.

> The pupils now regularly do writing tasks, because one of the obligatory parts of the exam is [productive] writing, so now they also write several pieces of work in the semester. At the moment mainly at home.
>
> School 22 Teacher 2

Where writing tasks are assessed there is also a slight trend to use the rating criteria from the exam for marking, as we see described below:

> And the assessment of writing, I try to assess as it is assessed in the *gimnazjum* exam, so separately for accuracy, spelling, the coherence of the answer and logic.
>
> School 12 Teacher 1

There are indications that teachers are also planning their assessment programme with the exam in mind.

> In the course of a semester the learners get between 10 and 15 grades for English. They are tested practically every month on a section of the course

book. The skills which are required in the exam are assessed, so we have separate grades for reading comprehension, listening comprehension, [spoken] production and conducting dialogues. Generally, all the skills that are required in future in the exam are assessed.

School 67 Teacher 3

We see that for Teacher 67/3 the exam seems to be the main motive in deciding what is assessed. She does, however, also assess speaking, presumably recognizing its relevance in the communication tasks in the exam.

The final aspect of assessment which is related to the exam and more widely in evidence than the areas described above, is the use of practice tests and mock exams. The next teacher offers a detailed description of how such tests are used in her school.

In the first class we do tests, of course, adapted to the level of class one. But they know that there are exam papers, they know what the types of tasks look like in the exam. They know that they have basic and extended parts. They know what they have to write, what types of tasks and how to do them. So we don't only focus on teaching the material but also on techniques for doing the tasks. For example, how to go about doing listening comprehension so that they hear the most and can cope with it, right? So it's not only, listen and do ABC, because that's not what it's about. Also, already from class one, those sorts of tasks appear in the course book, and of course we have copied materials, that we do. In class two we do a mock exam already. An official school level mock exam, where they come dressed up, and it's all organized. And they have their first exam, so that they can see what it's like. And, generally, I try once in a while, every two or three months, to do a test paper with them, so that they see more or less what it's like. And usually they get used to it, and in class three, it's clear for them what they have to do, when they have to do it, what types of tasks there are, so it's no stress for them.

School 70 Teacher 2

The programme of testing used in School 70 is fairly representative of quite a large number of schools in the sample. What is clear from this explanation, however, is that the motivation for following this plan is not simply teaching to the test in order to raise scores. What this teacher appears to be suggesting is that test papers and mock exams are used to familiarize learners with task types and procedures, in order to reduce exam-related stress. She points out that the course book includes test-like tasks from the first class, a fact also indicated by many other teachers in the sample, and that she uses such tasks not to test learners, but also to teach them strategies for coping with such an exercise. The example she selects is listening, a part of the exam in which many learners perform less well, possibly for the reason that it is stressful. The learner has no control over the speed of delivery of the text and knows that they will only be able to hear it twice. Encouraging learners how to deal with such a task seems legitimate teaching of a skill

associated with coping in authentic communicative situations, where we may not hear everything said to us, and may understand only parts of it. The teacher seems to be using the test task to develop important life skills, rather than simply doing practice tests, which Hamp-Lyons (1998) would classify as acceptable test preparation. Other teachers in their interviews limit their information to the fact that mock exams are done to get learners used to the form of the exam and the associated procedures, but do not describe whether they work with learners to develop coping skills.

The final theme, in this qualitative analysis of the teacher interviews deals with the forces at work on teachers which push them to make decisions related to the exam. This has been named Levers.

Theme Four: Levers at work on teachers

There were a limited number of instances in the data set where teachers referred to pressures being exerted on them. These fall into different categories. Some teachers themselves feel obligated to do the best for their learners in preparing them for the exam, for the reasons that the points for the exam will play an important part in selection for upper secondary school, which may impact on the learner's future. Teachers, in this sense, hold themselves to account. Other teachers report feeling accountable to others, primarily to the school principal, who they understand is also being held to account by the supervisory body. Other teachers report an awareness that exam preparation is expected of them by their pupils, who want to do as well as possible, or by the pupils' parents, who see the exam as important for their offspring. Some teachers see it as important that their school does well, as they feel this is key for the reputation of the institution, which in turn may affect its future. Good results, or results which are better than other schools in the area, will help attract more candidates, while poor results may have a detrimental effect. In uncertain times, with financial pressures on local councils, poor results could even impact on decisions about the school's future, particularly if accompanied by a falling roll.

> I would prefer that the exam had a lesser role, but unfortunately it plays a very big role. It's because I think that teaching to the test alone is not like conscious teaching of the language. But, anyway, the tests also encourage the learners to learn, to get some new knowledge, and learn different things, but in this situation today, unfortunately this exam, we as teachers are "evaluated", and we know that the principals and other institutions are looking at our exam results, and that's why I think that, unfortunately, the importance of the *gimnazjum* exam is growing these days.

School 13 Teacher 2

The teacher's sense of discomfort with the exam can be felt in the fact that she uses the word "unfortunately" three times in this one extract. She makes an interesting differentiation between teaching the language and teaching

to the test, which she appears not to view as teaching at all. Her discomfort seems to stem from her sense that she sees her work as being assessed on the basis of the exam results, not only at the school level, but also outside in the wider community. The next teacher shares some of the same concerns.

> Interviewer: What role does the *gimnazjum* exam play in planning and in the choice of materials for teaching English in your school?
>
> Teacher: A key role, because the *gimnazjum* exam is taken not only by my pupils, but above all it's taken by me, and the way they pass it is a witness to my work, and my involvement, and about what materials I chose for them. So every *gimnazjum* exam, please believe me, is an enormous stress not only for the pupils, but above all for the teacher.
>
> Interviewer: And who holds who to account?
>
> Teacher: For the *gimnazjum* exam results? I think it's a pyramid principle: we have to answer to our principals, our principals most likely to the education department [of the local council], the education department most likely to the Local Education Authority (LEA) [pol. *kuratorium*], and here I don't know, someone above, someone there, someone checks. Anyway, we have the summary, after every exam it's summarized, we discuss our results in our team. Then it's presented and discussed at a staff meeting, the principal speaks as the leader of the school, I'm sorry, the School Exam Commission, and here several times there have been quite, how to say, critical remarks. Up to now English has been rated high, in comparison with other exams, so it's been good so far. We'll have to see what will happen this year, when for the first time the pupils will also be assessed at the extended level.

School 68 Teacher 3

This teacher (68/3) identifies so strongly with her learners that she describes herself as taking the exam. She sees the exam as providing evidence of the quality of her work, her planning and the effort she has made. It is not clear if this strong sense of responsibility for her learners' achievements comes entirely from her, or whether it has been caused by what she describes as the "pyramid" of accountability. She is aware of the pressure on her to get good results, as these are required by the management, who in turn are answerable to their superiors. She is also aware that failure to get good results could lead to an uncomfortable situation, where there could be criticism in front of the whole school staff, as she has seen this experienced by her colleagues who teach other subjects. This is description illustrates the many and complex levers at work in relation to the exam: the teacher's conscience, her solidarity with her learners, her loyalty to her school, and her desire to avoid public censure.

Sometimes the teacher has direct pressure placed on them overtly:

> I don't have much choice. The principal has already made it quite clear, the most important thing are the exam results.

School 31 Teacher 3

As we see with Teacher 31/3 her boss has set the priorities and she has no choice but to conform. The next teacher is able to put the pressure into perspective, as she understands where it comes from.

Interviewer: And who holds who to account?

Teacher: We are held to account by our management. We always test, discuss and analyze the results of the *gimnazjum* exam and come to appropriate conclusions. Then we apply a corrective system if something didn't work, what to do in the following year so that the results will be better. Later there's pressure on the school, on the directors, from their supervisors. If the school does badly, then they are the ones they are after, and so it's a chain. We have to make demands on the pupils, they make demands on us and demands are also made on them.

School 79 Teacher 2

Teacher 79/2 clearly describes how a "chain" of accountability leads to a focus on the exam. We will return to this in the next chapter in discussion of the washback mechanism.

The next extract shows another way that a teacher may choose to rationalize the pressures they experience on them.

Interviewer: And who holds who to account?

Teacher: We are held to account by the directors, aren't we? We present our conclusions, we see what the results are like. The principal shows us bar graphs and you can see, can't you? If we're above or below the *powiat* [regional district council] the voivodeship, the national average and so on. Maybe it's not on the basis that it's personal and something threatens us, or something like that, right? But, anyway, if those results are lower, that motivates us to look for a solution, so as to change something, doesn't it? Both in our targeted work plans [pol. *planach wynikowych*] and in the way we teach, how we present the contents, doesn't it? So it's accountability in the sense of motivation to work, to make certain changes.

School 116 Teacher 2

Teacher 116/2 does not seem to find the need to account for results threatening. Rather, she sees the analysis of exam results and the search for reasons for them as stimulating, as a motive for change and improvement. Like Teacher 79/2 she has rationalized the situation, seeing the larger picture, being aware of the mechanisms at work and takes a positive stance. The question of why some teachers respond positively to these pressures, while others appear to find them frustrating is one which will we return to in the discussion in the next chapter.

The next trend in the data is where teachers feel pressure not so much from "above", as from the learners themselves, as illustrated in the following extract:

> I think, I personally think that the exam isn't the most important. But I am forced to it by the attitude of the pupils, and so I have to prepare them for it as well as possible

School 35 Teacher 3

The next teacher also feels pressured by some of her pupils, and draws attention to the problems this creates:

> I don't really like it very much, but there's a strong pressure to teach to the exam. The learners draw attention to it, that they want to learnt to do it well, they want to be able to read, listen, unfortunately speaking is pushed into the background, to put it mildly. That's because they often learn "just for the exam", in quotes. But of course there are those pupils for whom language acquisition in general is important, that is speaking and writing… But there are others who learn for the exam to please their parents. I think that's what it's like in our school. I don't think that's so bad really, but really it should also be important for them not only to do well in the exam, but also to learn how to be able to use the language and later make use of it. Their knowledge of the language, abroad, for example.

School 103 Teacher 1

Teacher 103/1, unlike 35/3, thinks that only some of her pupils are focused exclusively on the exam. Their interpretation of what is in the exam appears to be literal, and they consider that as there is no oral part, then speaking as part of their course is unnecessary. By contrast, there are other students who are interested in developing their language skills, not just in working for the exam, and for them speaking and writing are both important. Here the teacher seems to imply that the exam-focused learners have realized that points from the extended part of the exam do not yet count for school selection and so writing, which is only tested in that part of the test, is not "necessary" for them. In addition, the teacher indicates a lever at work on some of the pupils, which is to do well in the exam to please their parents. She points out the difficulty caused by the fact that for the exam-focused learners developing productive skills, the ability to use the language, do not equate with what is needed for the exam.

The final extract illustrates some of the frustrations a teacher can experience as a result of the levers at work and the pressures exerted on them.

> Well, unfortunately, the exam plays a big role. I say unfortunately, because it seems it's the next exam which was introduced to limit us in a way. Because a large part of lessons are spent preparing learners for the exams, that's teaching them certain exam strategies, selecting material to meet the requirements, sort of adapting them to the exam requirements. And really those lessons are already squeezed onto some sort of treadmill. We have fewer possibilities to do any extra things, which would be more interesting for the learners, but unfortunately it's what is expected of us, I mean the parents and the pupils

expect that the school will prepare them well for the exam, that the exams will be passed at a high level. And here, unfortunately, I have to have, in this way I also have to be supervised.

School 114 Teacher 1

Teacher 114/1 exhibits the traits of a teacher who feels trapped by the system. The language she uses is indicative of the pressure she is experiencing. The exam is viewed as a conspiracy to "limit us in some way", forcing class time to be spent on exam-focused work, using exam-focused materials, leading her to describe lessons as a "sort of treadmill". The forces at work in this factory-like image of the classroom are the parents and pupils and, perhaps less immediately, the knowledge that her work will be evaluated on the basis of the learners' results. We will return to what may cause some teachers to respond negatively to pressures in the discussion in the next chapter.

In these sections we have investigated four themes which emerged from the Teacher Interview data: Beliefs and Attitudes of teachers to the exam, Planning of work and choice of materials, Assessment in school and the exam, and Levers which teachers perceive are exerted on them in their work. The themes were illustrated with examples taken from the data in an attempt to offer insight into the processes and mechanisms at work in whether the exam exerts influence in school and if it does, how this might come about. Theories on the working of these mechanisms will be discussed in the next chapter.

5.3.2. Additional Teacher data

In two parts of a longer questionnaire, administered electronically by a dedicated online platform, teachers were asked questions which pertain to the present study. In the first set of questions teachers were asked to rate on a four point scale from *not at all important* to *very important*, the importance of different aspects of language, when learning foreign languages. The items were then ranked and the following outcome obtained:

1. speaking
2. vocabulary
3. listening
4. reading
5. grammar
6. writing

Learners were asked a similar question.

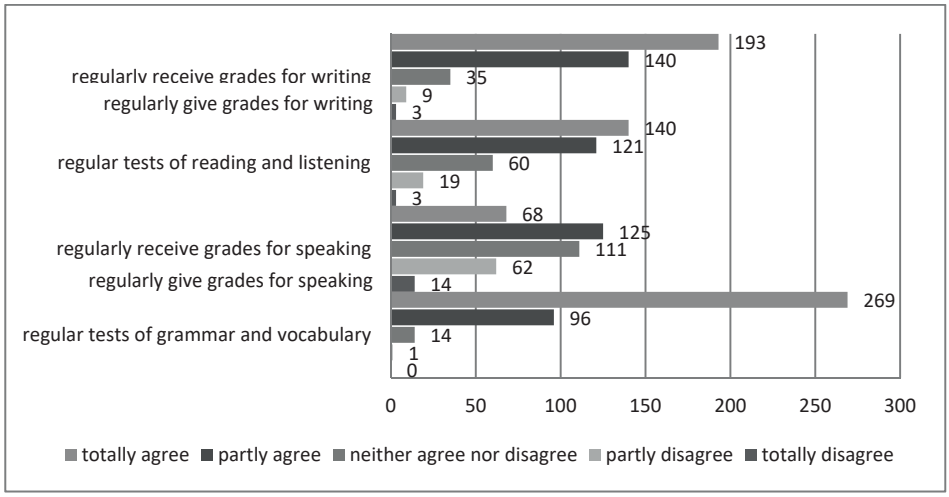
The next set of questions related to assessment. Teachers were asked to state how often they gave tests, or grades for different aspects of language.

In Table 5.6 and on graph 5.3. below we see that tests of grammar and vocabulary are most frequently used by teachers. Grades for writing follow, with tests of reading and listening in third place. The distribution for

speaking differs from that of the other skills with fewer teachers choosing "totally agree" (68) and a large number giving neutral responses (111), indicating that there is some ambivalence towards the grading of this skill.

Table 5.6. Phase III Impact Study: Teachers on how frequently different aspects are assessed in English lessons

	regular tests of grammar and vocabulary	regularly give grades for speaking	regular tests of reading and listening	regularly give grades for writing
totally disagree	0	14	3	3
partly disagree	1	62	19	9
neither agree nor disagree	14	111	60	35
partly agree	96	125	121	140
totally agree	269	68	177	193

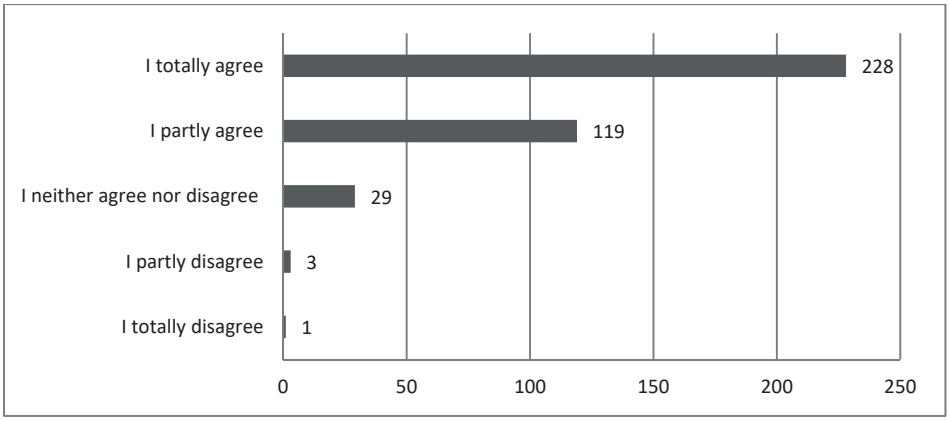


Graph 5.3. Teachers on how frequently different aspects are assessed in English lessons

The final question in the questionnaire which is of relevance here concerned how teachers opinions on the statement "We do lots of exercises during class preparing for the *gimnazjum* exam in English." They were asked to respond on a five-point Likert scale, from 1, *I totally disagree* to 5, *I totally agree*. As we see from graph 5.4. below, the majority of teachers (91.3%) strongly or partly agreed with the statement. Learners were asked the same question.

This concludes additional data obtained from teachers, which is of relevance to the topic of this study. For information about other aspects in

both the teacher questionnaires and interviews, please see the interim report on the first part of the BUNJO project (Paczuska et al., 2014).



Graph 5.4. Teachers on "We do lots of exercises during class preparing for the *gimnazjum* exam in English"

5.4. Learner interview data

Analysis was made on the basis of transcripts of interviews which had been coded following a procedure described in Chapter 4, by a team of coders. To obtain the quantitative analysis, the frequency of the axial codes for the response to each question was calculated by the researcher. As learners described different aspects of the exam, and as each of these were individually coded, the total frequency count is sometimes greater than the number of respondents. In preparing the analysis reference was also made to the relevant section of the internal report prepared by the sub-contractor on the interviews.

Learner responses to each of the five questions asked will be described in turn.

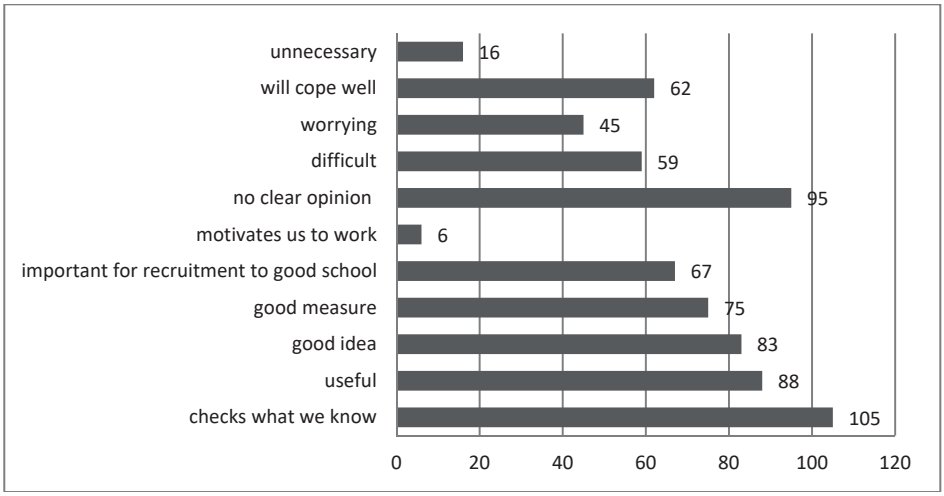
Question One: What do you think about the *gimnazjum* exam in foreign language?

Learners gave a variety of responses. Nearly 20% (19.8%) of them either said they did not know, or had no clear opinion, yet the remainder gave answers of varying lengths and complexity. The majority of these were positive.

As we can see in graph 5.5 learners commented that the exam will "check what we know" (105 counts) and judged it as a good and objective measure (75). It was felt to be "useful" (88) and "important for recruitment to a good school" (67). The introduction of the examination in FLs to the lower secondary school external examination was seen as a "good idea" (83). A small

number (6) felt the exam "motivates us to work." There were 62 mentions that the learners "will cope well" with the examination.

Fewer of the responses were critical, with 59 counts of mentions that the exam was "difficult", or that it was "worrying" and stressful (45). A small number (16) said the exam was "unnecessary".



Graph 5.5. Responses to Question One: What do you think about the *gimnazjum* exam in foreign language?

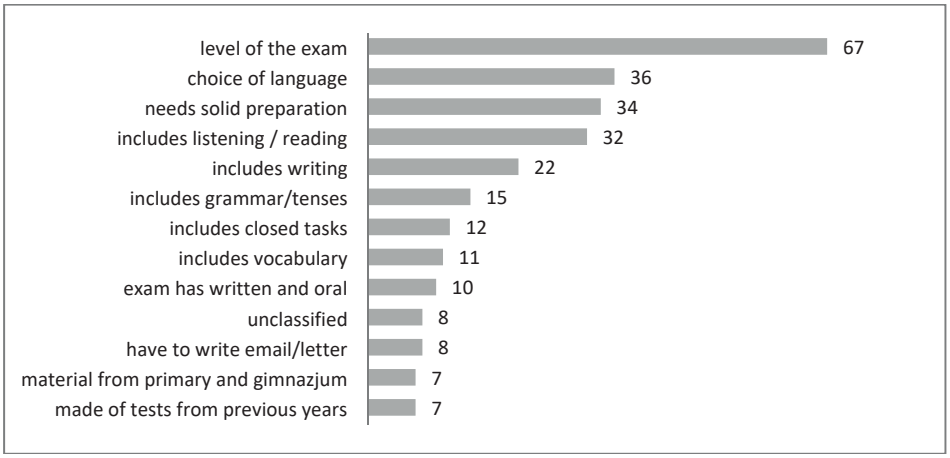
To sum up, generally we could describe learner responses to the FL examination as positive, viewing it as a useful objective measure of what they have learnt. However, learners' views of the level of the exam are divided (among those who referred to this), with 62 positive expressions about the expected outcome and 59 mentions that the exam was difficult. A number also mention the stress associated with the examination.

Question Two: Do you know something about the *gimnazjum* exam in foreign language (FL)? What?

Slightly above half of the 480 students interviewed, 277, that is 57.7%, claimed to know nothing about the FL examination. Of these, boys gave a negative response a little more frequently than girls, 62% as opposed to 54%.

Among the affirmative responses (203) the largest number (67) of mentions referred to the fact that the exam had two levels (see graph 5.6). Next were three groups of responses at similar levels, around 30 mentions each. The first of these was that learners could "choose the language" they took in the extended level of the examination (36 responses), followed by "the exam needs solid preparation" with 34 mentions, and "it will include listening and reading" with 32 mentions.

The trend for description of what the exam contains continues, but with fewer responses in the following categories. There were 22 mentions of "there will be writing", 15 of "it will include grammar (or tenses)" and 11 that "it will include vocabulary". 12 mentions referred to the type of tasks, labelled "there will be closed tasks". 10 mentions (erroneously) stated that the exam consists of a written and oral part. 8 mentions said that "you have to write a letter/an email". 7 mentions thought the exam "covered all the material from primary and *gimnazjum*" while 7 thought it consisted of "tests from previous years". The remainder were single responses and a group of 8 vague and unclassifiable answers.



Graph 5.6. Responses to Question Two: What do you know about the *gimnazjum* exam in foreign language? Number of mentions.

To sum up, we can see that a large number of learners know nothing about the exam and the remainder know a little, at a very general level. The highest number of mentions seem to relate to what might be considered information that is key for the learners: they have a choice of language, there are two levels and they need to prepare conscientiously for the exam. A very small number give more details, but this is limited to less than 3% of the information given. We also see that a few of the learners have mistaken ideas about the exam, such as that it is based on past tests, or that it includes an oral component. These are first year students interviewed near the end of year one. For them the exam is still quite far in the future.

Question Three: Is the *gimnazjum* exam in foreign language important for you?

In response to this question there were a total of 491 coded segments. Of these the majority, 433 (88%) were positive. There were only 24 (4.9%)

negative mentions. The remaining 21 mentions (4.2%) were unclear and could not be classified.

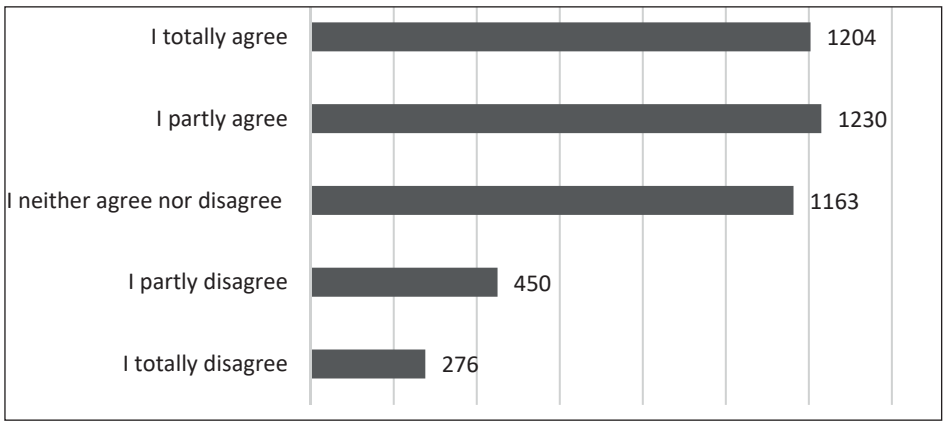
Question Four: Is it as important as the other parts of the exam?

In response to this question there were 482 coded segments. The largest number, 376 (78%), stated that the FL exam was as important as the other papers in the examination (Mathematics, Natural Sciences, Polish, History and Civics). 41 (8.5%) felt that the FL exam was more important than the other parts, while 65 (13.5%) felt it was less important. As we saw in the previous question these responses differ from what we were led to expect by some of the teachers who were interviewed.

Question Five: Has the *gimnazjum* exam in foreign language been mentioned in school?

There were a total of 549 coded segments (i.e. exceeding the number of learners in the sample) in response to this question, usually caused by the fact that some learners repeated themselves in their answers. Of these the majority, 71.6% (393), were that information had been given about the examination in school. 25.2% (144) were that no information had been given, and the remaining 2.2% claimed not to know whether information had been given, or not. This information is slightly contrary to what was found in question two, where more than half of the answers were that the speaker was unable to give any information about the exam, but logical if we bear in mind that being told something about the exam does not mean that learners retained that information, or were able to explain it.

5.4.1. Additional learner data

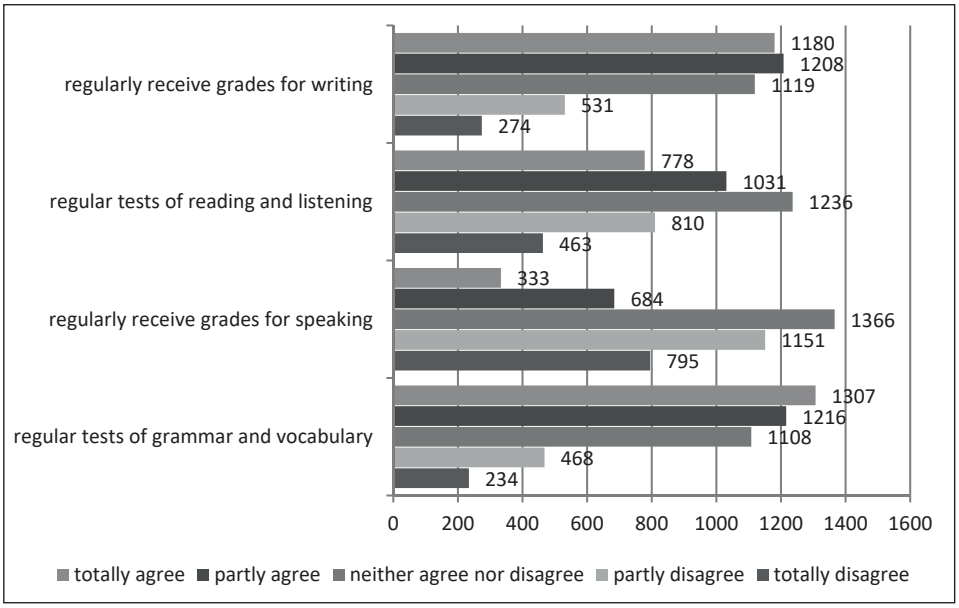


Graph 5.7. Learners on "We do lots of exercises during class preparing for the *gimnazjum* exam in English"

Apart from the questions about the *gimnazjum* examination in the learner interviews there was one question directly about the exam in one of the two written questionnaires completed by all learners in the sample where learners were asked to respond on a Likert scale to "We do lots of exercises during class preparing for the *gimnazjum* exam in English." The findings were as follows (see graph 5.8 and Table 5.7.).

Table 5.7. Impact Study. Learners on how frequently different aspects are assessed in English lessons

	regular tests of grammar and vocabulary	regularly receive grades for speaking	regular tests of reading and listening	regularly receive grades for writing
totally disagree	234	795	463	274
partly disagree	468	1151	810	531
neither agree nor disagree	1108	1366	1236	1119
partly agree	1216	684	1031	1208
totally agree	1307	333	778	1180



Graph 5.8. Learners on how frequently different aspects are assessed in English lessons

From the graph we can see that there is a strong trend for learners to respond positively to the statement, with 56.3% (2434) stating either *I totally*

agree, or *I partly agree*, as opposed to 16.8% (726) who expressed either total or partial disagreement. 26.9% (1163) were undecided. This seems to indicate that slightly more than half of the learners in the study consider exam preparation exercises are already being done in class one of lower secondary school.

There were four statements in the same questionnaire on the frequency of different aspects assessment in English lessons, also to be responded to on the same Likert scale described above.

The results of the analysis show that it is tests grammar and vocabulary that occur the most frequently, followed by regular grades for writing. The distribution of "regularly have tests of reading and listening" is more varied, but is still positively skewed, although we see a large number of ambivalent responses. These may have been caused by the formulation of the question, which grouped these two skills together. Speaking is clearly much less frequently graded, although again there are a large number of undecided responses.

Learners were also asked, in a second questionnaire, to rate on a four point scale from *not at all important* to *very important*, the importance of different aspects of language when learning foreign languages. These were then ranked and the outcome was as follows:

1. speaking
2. vocabulary
3. writing
4. reading
5. listening
6. grammar

Teachers were asked the same question in their questionnaire.

These items have been selected in order to allow comparison with other groups of respondents, and also with responses from learners in the Initial Washback Study in 2009. We will return to this in the next chapter.

5.5. Data from interviews with school principals

Interviews with school principals were conducted in the BUNJO project to collect data on a number of topics, one of which was the role of the external examination in foreign language.

Eight themes were identified and these will be described below, using illustrations taken from the data. The main themes were:

1. Analysis of exam results
2. Comparison of results with others
3. Corrective programmes
4. Systemic responses to exam results

5. External exam results as one part of a complex school assessment programme
6. Changes in learners noted in response to the new examination
7. Possible reasons for different responses to the exams from schools
8. School leadership styles

Theme One: Analysis of exam results

Principals describe how teachers do detailed analysis of the results of the *gimnazjum* exam, making use of the full report published each year by the Central Exam Board (CKE) or the regional exam boards (OKE). These reports show the linking between the test item and the core curriculum, explaining which standard in the core, and which aspect of that standard is being assessed by a particular test item. Some schools then do their own item analysis for each of the groups who took the exam, using the detailed breakdown of results they receive from their OKE. This is often done by a team of subject teachers, who are specially tasked to analyze the results, discuss them and come to conclusions. They are able to gauge the achievement of a given group. In some cases they also calculate the facility value of each test item, to ascertain which test items were easy and which were more demanding for the different groups in their school (see School 1 below). They also analyze the different parts of the core curriculum which were assessed, to find what went well, and what was problematic for their learners. A presentation and/or report is usually prepared and formal conclusions drawn up.

(the school principal is referring to the report prepared by her teachers)

> The teachers get together in a team. Foreign language teachers, because that's the team we have, they…come up with the successes and the failings. For example, the skills that are best achieved, the ones according to the OKE…Identifying the main idea in a written text. Finding specific information in a text, they were the ones that were done best… And the ones that couldn't be done, or were done poorly: asking for permission, giving or refusing permission. Those were the questions, that's how it was…
>
> And what didn't come out well, I can read it for you, seeing as we're explaining it in detail. For example, the skill which caused our students the most difficulties was identifying the main idea in a text… Most likely causes: reading the text too quickly and without sufficient attention. Weakness in the ability to generalize or make conclusions. They have to work on that, simply…
>
> School 1

The principal explains in some detail how results are analyzed using facility values to compare the performance of different groups. Here the results being discussed come from two sets of mock exams, one prepared by CKE and the other by a commercial company who offer mock tests to schools each year.

> We do analyses. These are the ones [mock exams] from CKE and Operon, because there were two. We do it like this, in each class, the number of points the class got on average, what result was the highest, what was the lowest. We work out the mean. And, for example we do the facility value of the test… and here if we take the exam from CKE some classes, the basic level part went, I'd say, not too badly, because the class came out with that level of facility and even had 0.74. So for them the basic part was quite simply, easy….But we're a bit worried, because some classes, for example, were moderate. We have three class threes. It was, the difficulty of the test was average, it was 0.64. It was moderately difficult. Hmm. For the basic level. Moderately difficult. Whereas for the extended level, for example, from Operon, it was difficult, 0.42 for the extended. And from CKE too, the extended level was difficult.
>
> School 1

Most of the principals describe the analysis more generally, but show the approach that is applied:

> My colleagues in their team do this [analyze the results] in detail. Not even so much from the perspective of the raw scores, what got what points, or how many, rather from the perspective of which questions came out worst and which questions they were, which went the worst, and we think about why they went the most badly, and what the students might not know…
>
> School 43

In other schools the principals have a less rigorous approach, leaving each teacher to do their own analysis and make their own conclusions.

Theme Two: Comparison of results with others

Analysis of the results often involves comparison of the school's outcomes with that of other schools:

> We do a very careful analysis, looking particularly at the results. We look at the types of task that were done poorly, what were the reasons for those weaknesses, why, for example, a certain task didn't work out at all. We look at the contextual factors too, that means what human material, learners, we have to deal with. Could we have tried [to get] more from those pupils? Was it a one-off incident that something in the exam didn't work out, or did other factors come into play? We look at that just for these reasons, which areas caused the biggest problem. And then, after that detailed analysis, we make a presentation at a meeting of all the school staff. We prepare a multimedia presentation, we calculate the facility values for each of the exercises at the level of classes, and at the level of the [whole] school. We compare the facility value at the level of voivodeship, and at the national level. It's a really very long and detailed analysis. Next the teachers prepare a programme which aims at improvement, and simply they show for themselves the number of exercises, and which exercises they have to increase, and what action they need to take in the following year, so that the result will be better.
>
> School 111

The principal of School 111 compares how difficult learners in her school found the exam tasks, with how difficult they appeared to be in the wider context of the voivodeship and at the national level. We can imply that if a task was found difficult by her learners and not by the wider population, then this is interpreted as a cause for concern, and serves a focus for the action plan for the following year.

Other principals are more concerned about the appearance of their results to those outside. School 90, below, is concerned about their position in an unofficial local "league table":

> We try to compare ourselves with the *powiat* [area covered by the district council], with the voivodeship and we look at which place we are.
>
> School 90

The next school (97) focuses on the national stanine scales, which show the mean scores of the school in relation to the whole population of test takers in the country. The position on the stanine scale, which show the school's mean score in terms of the number of standard deviations above or below the national mean score, was used in 2009–2011 as a measure by the LEA to determine whether corrective action should be taken by the school to improve results. The stanine scale was replaced in 2012 by a centile scale.

> Well, that's to say, the result of the *gimnazjum* exam, as the result of an external exam, is of course important, which is why, well, every school would like those results to be quite good. But we have different youngsters. We have different year groups, and we have five classes in each cohort. There are classes that are better, better let's say intellectually, there are others a little weaker. Well, all of that makes a difference, of course, to the exam results, the learning outcomes. Well, up to now we didn't, in any subject, have to write a so-called "corrective" programme, at the moment we're holding at the same level on that stanine scale, … this year it will be different, in centiles. I'd like to say something more about the stanine scale. Also at the moment we're holding a satisfactory level. But English, which came in not long ago, as an exam subject, that came out quite well.
>
> School 97

The next school (104) uses the comparison with others as a point of wider reference, to put their own scores into perspective. Yet, the approach taken seems to be to look for deficits which need to be identified and corrected, rather than nervousness about position, or concern for appearances.

> After analyzing the real exam the teachers, in the subject team, prepare a report analyzing the results of the exam both qualitatively and quantitatively, and they prepare another corrective programme, together with recommendations which are implemented in the following school year, and so the results play a really enormous role, all the more because all the mock exams which we do, and the *gimnazjum* exam itself, we try and make it have an external reference, that there is that element of reference to other schools, other pupils, as far as

possible even on the national scale. Then it's simply easier to determine our place in the ranks, to describe our weaknesses and to get on and try to make up for them

School 104

This final school explains why comparison of results with other schools is of importance:

We try so that our results are the highest and that word about our good results gets around, because it's a good signal to, let's say, the locality, to the surrounding area.

School 117

We need to bear in mind the fact that schools compete for pupils. Good results are seen as a good advertisement for the school.

Theme Three: Corrective programmes

If the school's results are considered unsatisfactory, which is interpreted as falling below the national average, indicated on the stanine scale prepared by CKE, then according to the regulations set out by the Regulation on pedagogical supervision in schools (see description in Chapter 2), the Local Education Authority (LEA) requires that a "corrective programme" is prepared. This is approached differently by different school principals. Some go "by the book" without question, while others are more circumspect, seeing difficulties in blind implementation of the regulations. These principals find various ways to adapt the requirement to their schools.

In the first description we see a principal who rigorously adheres to the regulations.

After the exam the recommendations are implemented the following school year. Yes, the teachers are required to do more exercises, and to plan their schemes of work so that there are more exercises in those areas where the pupils had problems in the exam. And that is checked at once in a diagnostic observation.

Interviewer: A propos, who is holding who to account? How is that checked? Who is accountable?

Principal: It's the principal who is checking by doing that diagnostic observation, checking to see if the recommendations in the report are being implemented. Mainly the principal, the principal has to account for the results in the annual report to the LEA.

School 69

Not only are teachers in School 69 required to plan to compensate for the shortcomings of pupils in the previous cohort, but they are also checked by the principal, who observes their lessons to make sure the "corrections" are being put into practice.

> Above all, on the basis of the results, the teachers have some information which skills the children are weak in [lit. lame, limping]. So they have to work more on those skills. And so, because of that, they have to organize lessons with that in mind. If what's needed is work with listening texts, then the teachers do that in their lessons. If there are some problems with grammar….then they have to work on that.

School 93

Interestingly, the principal of School 93, in other parts of the interview, makes no differentiation between how the teachers are expected to respond to mock exams, which take place at the end of year 2 and early in year 3, and how they are required to respond to the external exam itself, seeing the same approach as equally appropriate.

> Our results are always very good and that's very pleasing, but no-one wants things to stay the same, to stay as they are. They want, so to say, for them [i.e. the results] to keep getting higher. So, every time the teachers do an analysis of the external exam they catch, so to speak, the weaknesses of our pupils and then, of course, we know it's a different group of learners we'll be working with the next year, but they try to pay more attention to what went a little less well, to spend more time, so as not to let it happen that in the next exam the same, let's say mistakes, are made, or that the pupils have the same problems.

School 9

This principal, although not required to prepare a corrective programme, as the results are well above average, still insists that the teachers follow the procedure and apply what could be described as preventative measures in the following year. The motivation for this seems to be the desire that the results, although already high, become even higher. We shall return to this practice in the discussion in the next chapter.

Many of the principals question the requirement to use the results to adjust the programme for the next cohort.

> I mean it's a point of reflection on what worked, what didn't work. We got those results, but it's a question of conscience, why they are like that. But the results of the final exam, the results of the final exam themselves, they don't really give very much. At that moment you have to refer to…It's always possible to say that the learners were very stressed and that's why the results are so poor. On others stress has a mobilizing effect, so, at that moment, you have to refer that exam to internal exams, and compare if at that moment students who regularly got very good results suddenly had a crisis, because those exam papers for all our classes are standardized at the same level. They're tried out one year in one class and then the following year, for example, they are improved. Of course, we know that there's a market that functions in school, a market for information. So a test from one year can't be used in the second year. But if it turns out that a task was badly formulated, badly edited, then that has to be taken into consideration in the subsequent years and conclusions drawn. So the results of the *gimnazjum*

exam are a kind of information. But that information by itself, which doesn't tell us much, it has to be referred to the characteristics of the learners themselves.

School 39

The principal of School 39 reminds us of the differences for the learners between low-stakes formative and high-stakes summative tests. Mock exams done during the school year, or progress tests done in class, he argues, give more reliable information about the pupils than a one-off summative test, which is in addition high-stakes. Further, formative assessment provides a picture of the learners which is built up over time, providing a more reliable body of evidence for evaluating the true ability of a pupil, than a snapshot taken on one day in the summative test. He argues that the results must be contextualized, as they refer to individual learners, each of whom is different, and responds to the stress of an external exam in a different way.

The next principal questions the usefulness of preparing a corrective programme for the subsequent year groups, as each year group may be quite different:

> Put it like this, over the last years, since that exam started, it's hard to say that the results of that exam have helped to plan the work of the following year. Because each of the next year groups is different. It's specific for a particular group, one is better at one thing, the next is better at something else. This one is more active than that one. And even if we had fantastic results, we could say to ourselves, well what we're doing is great, let's just keep doing the same and we'll get those results again, but that's not true at all, because in a year's time the results might be tragic, because it turns out that the group which is in the second class, or the first class, by the time they get to the *gimnazjum* exam, if we work in the same way with them then they might not get the same good result. So those results of the *gimnazjum* exam are simply a sort of snapshot of what we've done, if we've managed to prepare those children for the next stage of education at a reasonably good level, that information is for somewhere there in upper secondary schools only, it's for them is that information.

School 51

As we see, the principal of School 51 believes that the information from the exam results should feed *forward* to the next stage of education, to the schools those pupils will go to, rather than backward to the next cohort of the *gimnazjum.* We will return to this idea when we consider how the washback mechanism comes about.

> Well it varies. There are better year groups, and worse. But our education authorities don't like it if we say it's a worse cohort. But I know full well that that's how it is. One cohort is good and the exam goes well, and some years are weaker. I don't know where it comes from. And then the scores are lower. But it's hard to say that it's the fault of the teacher. It's hard on the basis of the exam to say if a teacher is good, or if a teacher is bad. There have to be other things that make up the assessment of their work. The same if the exam [results] are

> poorer, then it's not the teacher, well, they should not suffer any consequences. Because there's also the stress of the learners. Well, we don't know what factors came into play, it's hard to say that it's the fault of the teacher. Well, you could try to do that sort of assessment, if, as different students take the exam, right? That's to say, you would have to analyze whose pupils did worse, right? Try that kind of analysis. And then maybe you could come to some conclusions, or [decide on] consequences. But generally we analyze our results at the level of the school, the whole school, don't we?
>
> School 85

The head of School 85 appears to be wrestling with a number of issues relating to the exam results. His main argument against the requirement to apply the corrective programme to the next cohort is, like that of School 51, that each year group is very different and that what went badly in one year does not necessarily apply to the next cohort. Like the principal of School 39, he points out that there are many other factors, such as stress, which may have had a bearing on the results. So he initially declares that for these reasons the teacher should not be held accountable in any way for the results of their pupils. However, he then begins to consider whether, if more complex analysis of results were done, which checked outcomes at the level of individual learners, or presumably also at the level of groups, whether it would not be possible to call teachers to account, if it turned out that their classes were performing less well than others. Fortunately, he dismisses this questionable idea, on the grounds that currently the statistical analysis done in his school is very general and stops at the level of the school as a whole. He has, however, demonstrated a thought process that, if pursued and put into action, could have serious consequences, and we will return to this in the next chapter. He also indicates that the LEA does not share the view that each year group has its own character, and that a cohort can be weaker than the one preceding. There is a slight suggestion here that the LEA applies a bureaucratic approach, applying the letter of the law, rather than being understanding about the real situation. This is another issue to which we will return in the discussion in the next chapter.

Theme Four: Systemic responses to exam results

Some principals respond to the analysis of results with solutions at a systemic level:

> The kids have a problem [with listening], so I say, we'll do it in small rooms, so the sound will be better, because in the sports hall, you can't hear. There's an echo there. And we have seven commissions, for example. We do our best so that the students can hear.
>
> School 1

Here the principal, responding to information that students had poor results in listening, re-organizes the examination, so that instead of having

all the students together in the hall for the listening test, they are divided into seven smaller rooms. This creates logistic problems, as all the tests have to take place simultaneously and each room has to have a separate 3 person commission to supervise the test, but the principal is adamant that the school needs to support the learners.

Results of exams are analyzed, as we have seen, and then used in various ways, with the aim of increasing the effectiveness of the work of the school.

> All the conclusions are then written down. We write them in the registers of the extra-curricular lessons, and they're also placed in the schemes of work, so that nothing is overlooked. For example, if some weaknesses were noted in the minutes of the school staff meeting, then every teacher concerned, at that moment the foreign language teachers get sort of instructions what to work on, what to include, and we as principals check too, to see if it has been included in the schemes of work, or if it's in the extra-curricular lessons, that it's in the remedial classes [pol. *zajęciach wyrównawczych*], or in the interest groups.

School 24

Other responses are more long-term, involving improvement of resources. The next two principals have invested in dedicated language classrooms with new equipment, in the belief that this will improve test results:

> What you see here, this equipment, is one of those conclusions, but more mine than the teachers', because this room is from September this school year. And I think that thanks to this, that we have this room, the results will be better. So if you go into the teaching resources too, as support, then the results will be better.

School 19

(interview takes place in newly equipped language classroom)

> The external exam in foreign languages in our school is low, I mean the results of that exam are low and that's why we took this step, to buy this language workroom, to buy it.

School 29

Often the principal changes the allocation of hours in response to exam results:

> All the time we're analyzing the results, and the very fact that the results aren't the best is why English here has such emphasis placed on it, in the form of three hours a week for all three years

School 31

The principal has allocated one of the "principal's hours" to increase the intensity of English lessons in one of the three years of the school plan, as statutorily foreign languages have a total of 14 hours over the three years, to be divided between the lead FL and the second FL. This is most often divided

for the lead FL as three hours in two years and two hours in remaining year, with different decisions as to which year gets fewer hours.

> Up to now the results of the *gimnazjum* exam, that's 2 years for foreign languages, were not taken into account for getting into upper secondary school. But the teachers still checked the level of learning. And it wasn't adequate. Because the youngsters, knowing that the exam wouldn't count towards being accepted to school, didn't apply themselves 100% to the exam. From this year, like other subjects, if languages come out poorly then extra hours will be allocated for foreign languages.
>
> School 94

The school principal here makes clear that if the results this year are not satisfactory then more class time will be given to FLs.

> For sure, the results of all the exams are sort of, hmm, for us some information about how knowledge is growing, and what more we need to do. That means, then we think about whether we should have more remedial lessons, or maybe more lessons for those talented learners.
>
> School 20

Another option are extra-curricular lessons (two hours per week offered by each teacher in lower secondary school, under article 42 of the Teacher's Charter) which the principal can decide how to use. Here the principal ponders the question whether to add more time for learners having difficulty, by increasing the number of remedial lessons, or whether to focus on the gifted and talented learners, and offer them lessons instead.

> **Theme Five: External exam results as one part of a complex school assessment programme**

For some schools the external exam and its results are simply one element of a more complex programme of assessment in school. This takes three main forms:

1) students are given a test at entry to the school, variously referred to as a "test of competence", a "diagnostic test", or a "placement test", the main aim of which is to divide learners into groups according to language ability for the purposes of streaming. At the end of class 3 the same test is repeated and calculations are made of language gain. All learners do the same test and in some cases comparisons are made of gain between groups. The external exam results in this approach are generally seen as an objective measure which sums up the achievement of the learners.

> Every year after the end of the teaching cycle there are, the foreign language teachers organize tests to check, that, I mean they organize a test on entry in first class and when it's the end of the school year, in class three they check how the level of learning has increased.
>
> School 94

2) Students take a test on entry, as above. At the end of each school year there is a test of achievement given to all the learners, based on the material covered during the year. The team of subject teachers generally work together to produce one test for all the groups. Teachers analyze the results carefully, using them to diagnose learners' strengths and weaknesses and take these into consideration when planning the following year's work. As above, the external examination is seen as summing up of the work of the teacher and/or school.

> We also have an internal system of assessment, when the pupils come to us in September, initial tests, which allow us to find out what those pupils have brought with them when it comes to language skills. It's a test of English. After three semesters, in class 2, there's assessment at the so-called half-way point, which checks what the effects of the work are from that period. And the third stage is a mock exam in class 3. And we look at those three tests to see what is happening with a given student and their language skills. This is one of the elements. And apart from that, of course, the teachers do progress tests, on the basis of which they give information if there are problems, if there is progress and what progress.
>
> School 26

A variation on this is where the results of the achievement tests are compared with the baseline entry test (or with each other) and diagnosis is made of learners' progress, or language gain (difficulties inherent in this will be discussed in the final chapter).

3) Students take a test on entry as above. At the end of the first year there is an achievement test given to all learners. In some cases, either instead of this test, or in addition to it, an exam based on the external examination is prepared. This is usually described as being adapted to suit the level of first year learners, but following the same format as the final exam. In Year 2, usually at the end of the first or beginning of the second semester (around February), a mock exam is held. Results are diagnosed with a view to diagnosing how prepared learners are for the final exam, and what weaknesses still need to be worked on. Early in Year 3 a further mock exam is held with the purpose of gauging what still needs to be polished, or what areas need to be strengthened. Some schools organize more than one mock exam in Year 3. In most cases the schools make use either of the mock exam offered by CKE, or use other tests offered by commercial providers, or publishers. The results of the final external exam are carefully analyzed with the aim of making a plan which will lead to improvement of scores in the final exam in the following year (problems inherent in which will be discussed in the next chapter).

> At first we have a sort of test, you could say on entry, where we decide what level the children we have are at. And then, after each year, the teachers who have taught each of the classes, do a test and decide, assess, try to calculate the gain in knowledge, if there really was growth, if those kids have developed, have

> greater knowledge and skills, can use the language better. And in subject teams it's analyzed. We also do mock *gimnazjum* exams, both in the third class and in the second. In fact, today is the day when we have the mock exam in class two. And also those, the results of those exams we analyze, we check what we need to correct, what we have to work on more. We see different things, because let's say one year group has one problem, and another has a different problem. We try to work it so that it comes out well, and so our kids will be able to continue their learning at a, at the level of advancement that we are satisfied with, because it is also an opinion about our school… In the third class we try to do three or four mock exams. And then we analyze those exams, come to conclusions and prepare a corrective programme. It's not really possible to apply the exam, for example, from one year group and work on those problems with the year below, but maybe that's another problem completely, that's why we try, we do those mock exams in class 2 and really those results, from that exam, are very important for us, because we want there to be the greatest, the highest, if we're talking about school, standard of teaching.
>
> School 16

We see in this description that the principal attributes the implementation of mock exams in class 2 to the fact that the external exam itself is considered not helpful for diagnosis of what to remedy. As an alternative, this school, like many others, uses the mock exam to diagnose weaknesses *prior* to the main exam and adjusts the programme accordingly in response. This appears to be a more workable solution in answer to the LEA requirement for corrective measures than applying corrections to the next cohort, but it is not without consequences, as we will discuss in the next chapter.

We could place these three approaches to a school-based assessment programme on a continuum from the first, where the exam is seen as an external summative measure, while the main thrust of the assessment programme is on measuring the progress and language gain accrued by learners over the three years of school, using a simple entry-exit repeated measure. The second approach shows a more complex whole-school approach to assessment, which monitors progress more closely, using annual measures, with the external examination providing a summative measure at the end of the educational stage. The final approach could be described as exam-driven and uses the external examination as the motivator for the whole assessment programme, throughout the three years of school. The thrust in this approach appears to be to maximize the school's results.

Categorization into these groups is not clear cut. To illustrate this, here is a description of a complex approach applied in one school. It combines strong focus on the exams with a variety of other measures, firmly situated in the context of the school:

> We start of course with a test on entry. Our teachers prepare a test at the level of class six primary. It checks them [i.e. the learners] and to that we add also their grades, an analysis of the grades on their primary school certificates.

They check, because if a child later shouts "but I had a 5!", yes, well a 5 and we compare it with the test, if a 5 is a 5. If it's a 5, maybe the child is nice and quiet and always, for example, did their homework with the help of Mummy at home, but their knowledge is not at the level of a 5. So we do this analysis, and every October the teachers get the results… and everyone has to know that we have a group of talented children. Because if we do, we've noticed an analogy, with physics, mathematics and geography. These are new subjects in lower secondary and they make us think. And the information, that I've got a group of talented language learners, then I know that in geography I can do something more with those children, and the physics teacher will be able to, it means that those children have good memories. Just that. I think that's where it comes from. So, in October we have the first information, and at the end of the first semester we have an analysis of the general level of the class. And also in November, between times, we do a background and behavioural analysis of class 1. Done by the pastoral care team, because if a child is weak, if the child's results are poor, it doesn't mean that the child isn't clever. Only that a child has a difficult situation at home. So, we add to the results what we now know about the home situation, and then we work.

At the end of the first semester we see, unfortunately from the grades, what the learning is like… Then we have the mock exam before the final exam in class 3. Sometimes in December, sometimes in January, it depends on when the winter break is. And that's the next stage of our evaluation, because we see how well our third classes are prepared for the exam, and what their level is. What we have to polish. Do we have to place more emphasis on preparation lessons, or on development lessons? And we come to conclusions and put more stress on that type of lesson. The third element is May. When they come after the exams, they come to a test, sometimes the principal's test, sometimes the teachers', and again we talk about what the level looks like for which class. At the end of May we have exams in class 2 and there's also a language exam. And again that tells us what level the second class is, 50%, 60% prepared for the exam, so in September we have to work on those weaknesses with a group… they are conclusions for our work in the future. At the end of August, at the last school staff meeting of the year, the conclusions are presented and here are the conclusions of the teams and the languages team, who are quite hard working, and they give their conclusions about their work, and so in September we start according to those recommendations. So, the entry test in class 1, so we work from the beginning and correct. Sometimes we have to stop and think, why are there these gaps, where do these gaps come from? Are these gaps brought from primary school? Or do we have to change something in our methods? This year we changed the course books. After the evaluation we came to the conclusion that the level of the books wasn't right for our children. We had to adapt the programme and the books.

School 70

Here we see that although ostensibly the school belongs in category three, with a strong reliance on the exams in terms of the programme of assessment and evaluation, the principal is also applying a more holistic

approach, verifying the results from primary school, marking out gifted or talented children on entry, and taking the children's home background into consideration. The evaluation programme includes both entry-exit measures (the test done after the exams in May in class 3 is the exit measure) which appear to be used for evaluation of the work of the school, and measures to check preparedness for the exams in mock exams at the end of Year 2 and again in the first semester of Year 3.

The principal goes on to explain why foreign languages play an important role in the school:

> I place a lot of emphasis on languages because we have quite a large group of children from homes with a difficult financial situation. And I know that in a year, or two or three, when they have left lower secondary school, those youngsters may emigrate to look for work. These are youngsters with no support from home, and language will be a key for them. A key to the gates of Europe, to a life. Many families, we have many Euro-orphans, so I know that language will be a passport for them, so that they can go to mum or dad, and not get lost there. Find themselves there. So I can let them off physics a bit, because they don't have to go to university. Not all my children go to university. Most of them go to vocational school, but they have to have language. For future hairdressers, future drivers, future car mechanics. They have to know the language, even if just in speaking.
>
> School 70

Thus we see that, although at first sight the school appears to be exam--focused, the principal's main concern is that the learners gain a sound knowledge of the language. Her concern is not for the reputation of the school, or for good results on the exam, but to prepare her students for life after school. This underlines the importance of placing the research in context, of maintaining a rigorous ethno-methodological approach and not being tempted to over-simplify in order to come to neat conclusions.

Theme Six: Changes in learners noted in response to the new examination

As we noted earlier, when first introduced, scores from the foreign language exam were not taken into consideration for school selection for an introductory period. In 2012, the year in which the BUNJO interviews took place, points from the examination in foreign languages counted for the first time. In effect, the examination became officially, rather than nominally, high-stakes. Of the 115 principals interviewed, 9 specifically mentioned that during the introductory period their students had not been motivated with regard to the exam.

> Up to now it's been very difficult because of the fact that the results had no effect on the fate of the students. So the students wrote, they didn't have really any motivation to write that exam as well as they could. So they didn't make

any effort for the results to be the highest and that's why it was very hard as a school to gauge at what stage, or at what level the students were ending the stage, that is *gimnazjum.*

School 49

This attitude has begun to change, as noted by the following principals.

Up to now the situation has been such that the language wasn't considered as, no, another way. The exam wasn't taken into consideration in any way for selection of school, and upper secondary schools did not really pay much attention to the results of that exam. And it simply varied. But it's not without significance that this year languages start already as one of the main elements of the exam. That, for sure, has meant that the youngsters feel a little more pressure and the need to study. So, I think that that's the direction it's going to go.

School 68

Well, I'll put it like this, up to now the exam was treated a bit like a fifth leg, to put it baldly, because it wasn't taken into consideration anywhere, so the learners also treated it like that. We had situations when we did a mock before the first language exam, when the youngsters almost took no notice of that mock, or the grade for the mock, because it was hard to get through to them. Well it helped a bit, they became a little more certain, I mean more certain, and they've stopped ignoring the exam. They treat it as seriously as the other two parts, that's the arts and the maths and natural science parts. Because up to now you could say that it seemed to them a bit as if language was unnecessary and redundant. But it turns out that it wasn't quite like that, and I'm telling you that now, at once, we have better results, because we can see that the youngsters are, I can tell you I was at the mock exam in English, as a member of the commission supervising it, and I watched to see how they worked, because it interested me. They listened to a text from one of the recordings, it wasn't a problem, they listened in concentration and they wrote.

School 52

Well up to now the results of the *gimnazjum* exam, let's say for 2 years there were *gimnazjum* exams [in FL], which didn't count towards recruitment anywhere. And that had an effect on the way learners worked, because they were well aware that it wouldn't have any effect on later, on their later study. But now you can see that the learners are approaching foreign language learning more seriously.

School 101

From these three accounts we see that these learners approached the exam pragmatically, not expending unnecessary energy when they were aware the results would not have an impact on their futures. The first principal (School 68) noted that the youngsters now appear to feel more pressure and need to study. The second principal (School 52) notes a clear change

in the students' attitude to the mock exam, while the third principal (School 101) thinks that learners have changed their attitude to FL learning and are now treating it more seriously. These examples illustrate one aspect of the washback mechanism at work – the exam has become of personal importance to the learners, because the points from it now count for selection to upper secondary, and so they approach not only the exam, but also the subject itself, differently.

Other principals indicated that it was not necessarily only the learners who took a less serious approach to the exam:

> Up to now the results from English, from FLs in general, were not taken into consideration for recruitment to upper secondary school so, to be honest, we did not look at the results in the context of organizing our work for the following year. It's just this year that, for the first time, the results will be considered in selection for upper secondary, and here of course we'll look at those results and analyze them and come to conclusions for our work for the coming year.
>
> School 35

It should be noted that schools like 35 were in a minority (7) and that the majority analyzed and responded to the results in the ways we have seen above. In general, the decision to include FLs as a paper in the lower secondary school external examination is seen as a positive step.

> To be honest, I am very glad that this part has been added to the external exam, because for years all the burden was on the arts subjects and maths and natural sciences, while foreign languages sat by the stove and were comfortable…
>
> School 66

Theme Seven: Possible reasons for different responses to the exams from schools

We have seen that different schools respond to the exam in different ways and now we will attempt to find evidence which may explain why this is the case.

> Interviewer: What role do the results of the FL *gimnazjum* exam play in the planning or monitoring of foreign language teaching in your school?
>
> Principal: Well they play a very big role, because unfortunately it works like this, that after those results schools are assessed. I mean they are information which the local authority looks at, which the LEA inspectorate looks at, and of course which principals look at. The first thing which happens is that a principal looks at their results. When the results of other schools appear, they compare more or less how good those results are, or how bad they are, and so we really look carefully at those results…But last year we had the third or fourth highest results in English in [name of large city], for public schools, and in fact it was a result about 10 or more higher than the average in [name

of large city]. It was very high and, of course, there was huge delight. Exam results are, on the whole, very important for us.

School 74

The principal of School 74 explains that the reason the results of the exam are important is because on the basis of them the school will be held to account, both by the local council, who employ and pay the teachers' salaries and maintain the school buildings, and by the local educational authority who are responsible for quality in education. Concern for the school's ranking appears to be caused by the need to check that the school is performing well when compared with other, potentially rival, schools.

The principal of School 88 below explains the process in a little more detail:

> It's like this, the teacher presents [the analysis of the exam results] to the principal and then later the directorship has to account for all those results once again externally. So it's not that it stays in our school. But it is all presented externally. The analysis takes place in school, above all in the whole school staff meeting, conclusions are made, considerations about how to improve the results in the future and then it's all presented outside. That's how it looks.

School 88

Apprehensiveness about the LEA is expressed by the next principal, who bases this on her experience of being held accountable.

> The school is held to account by the inspectorate, by the head of the LEA [pol. *kurator oświaty*]. Accountability at this moment hasn't applied to foreign languages, but to the other two parts of the exam. So if the school gets results below the right level, the right mean, then unfortunately the principal is summoned and a corrective programme is written. Up to now we've had very good exam results, above the average level. So we were glad. We'll see what happens next.

School 14

The principal of School 14 suggests that the principal of a school which performs below average is "summoned" to the presence of head of the LEA, a powerful person who has the power to brand the school as performing inadequately, with all the connotations this has for the principal themselves, as it will feature on their own performance record. This triggers a process where the inspectorate could commence a hands-on supervisory role in the school.

The principal of School 50 describes the process of analyzing the results of exams, not only of the external exam, but also of all the mock exams done in school. She continues:

> [The results] are analyzed for quality, first quantitatively, what they look like, what the mean score is for the classes, and then that is compared to the mean

for the school and we also compare, particularly in the case of the external results, the results of the external exams, with the results in the local district [pol. *gmina*]. Actually, in our situation we are the only lower secondary school in the local district, so we can't compare at that level, but at the level of the district council [pol. *powiat*], the voivodeship and on the national level. And from that we come to conclusions, and then we implement those findings like this, for example, we apply some, we produce a new set of tasks, tests, yes, that simply take the skill that was found to be the weakest, [to see] could it be better demonstrated in this way of teaching. It's a pity that up till now there's no value-added [pol. *EWD*] for foreign languages. There is for the maths and the arts parts, but I hope that with time the regional exam board (OKE) and the Central Exam board (CKE) will go more in that direction for analysis of foreign languages. How does it work in practice? In practice the teachers hold themselves accountable for the outcomes. I mean, there are several teachers and they compare, they must compare, how it went in their classes and how they did, compared with the others even. And of course, at a meeting of the whole school staff this information is presented, publically here in school, and later it's in the local newspapers, for example.

School 50

This school, like many others, places the responsibility for analysis of the results and the making of decisions about what needs to be changed in the hands of the FL teachers. Here the principal of School 50 seems to be suggesting that the question of accountability for those results has also been devolved to the teachers. She believes that the language teachers will self-regulate, by comparing the results of their classes with external results, but also in a way regulate each other, by comparing the results attained by learners in the classes they taught with those taught by their colleagues. The hidden barb is that this information is then made public, not only within the school to the whole staff, but beyond the school, to the local community, through the local media. There are subtle and powerful motivators at work in this process to which we will return in the next chapter. The principal comments on the use of value-added in other subjects for calculating the contribution made by the school in the learner's progress. Value-added calculators are made available by CKE, or the Educational Research Institute in Warsaw, for the other two papers of the exam, but are not available for FLs. Several principals commented positively on the use of value-added, claiming it gives much more information about the quality of education in the school than relying only on the results of the final exam.

The next principal suggests that the desire to improve the results of the external exam comes from the teachers themselves

Above all [the exam results] give us a basis on which to work further, with the aim of improving the whole educational process. Our exams last year were not analyzed, but they are always at a high level, not just English, but all the other

subjects. But that doesn't mean we rest on our laurels. The teachers are raising the targets higher and higher, so that the youngsters get even better results.

School 75

We do not have sufficient data from the principal to know whether the teachers' desire to set ever higher targets is the result of a competitive policy, like the previous school (50). What we do know, however, is that at the end of each semester, throughout the three years of school, there is a whole staff meeting where coverage of the core curriculum is discussed in detail. Here teachers who have fallen behind with what was planned in their scheme of work are asked to explain in public the reasons for the backlog, and are expected to find alternative times when they can make up what has been missed. These are then noted in the minutes and subsequently checked. We may surmise, therefore, that this is a school where teachers are expected to hold themselves to account and where there is a high level of administrative invigilation to make sure that this is the case. It is possible that in such an atmosphere teachers may feel the need to prove themselves by being able to show that their learners are making visible improvement. One difficulty in this is, however, the fact that the examination results for the foreign language exam are not comparable year on year, an issue we will return to in the next chapter.

This approach may be contrasted with schools where the exam results are only one small part of a wider system of evaluation which takes place. In School 76, for example, there is an extensive program of internal tests, all of which are carefully analyzed with a view to identifying areas which need to improve, and these are then compared with the formative assessments carried out by each individual teacher.

After each awarding of term grades [pol. *klasyfikacja*] there's a team meeting where the teachers discuss what, if any, changes are needed, for example in the scheme of work, or in the approach they use when teaching the lessons. Because sometimes those sort of things are necessary to, I don't know, make the lessons more attractive, to motivate the learners more. It all depends, because there are lots of those sort of meetings. Anyway at each of those meetings they come to some conclusions and make some recommendations, which they then put into practice. I think that's the most frequent form of evaluation. Of course, the teachers use some questionnaires related to specific issues which they want to find out about, but that's just from time to time on, as I said, key issues. But the whole analysis process, that's a cycle.

School 76

It would seem here, that rather than getting the teachers to compete with each other (as in School 50), or worry about external exam results, the focus is on cooperation and discussion among teachers, in order to come up with joint decisions which have the aim of improving the quality of learning. Note that the principal appears to view the practice of using questionnaires

to obtain more information on important issues as normal ("Of course, the teachers use…."). This would seem to suggest that the learners, or perhaps their parents, are being consulted, which was, in fact, found to be extremely uncommon in the BUNJO schools. We will return to the issue of cooperation vs competition among teachers in raising the quality of foreign language teaching and learning in the next chapter.

The next principal hints at another motivator in the process of why the role of examinations become increasingly important.

> I think they [the results] play a very important role, because it's a very objective assessment, no-one questions it, not the parents, or the pupils. And that assessment and those points have a very great influence on the learners, on the family, on the career of that pupil and on their future approach to that language. I think that the idea of this exam in English is spot on. The school is additionally motivated by the parents and I think that the results will keep on improving.
>
> School 72

The principal of School 72 points to the power of parental influence in wanting their offspring to do well. The reasoning behind this is natural, the results play a deciding factor in the future of the young people and therefore the family wants the best possible outcome. Implicit in this response is also the question of the social responsibility of the school to ensure that all learners get the best possible opportunities to get the best possible results, which will give them the best possible start in the next stage of education.

Theme Eight: School leadership styles

The final pattern to emerge from analysis of the interviews with principals was that principals appear to fall into two categories: the school manager and the pedagogue. The first case, the managerial principal, is concerned with efficiency, with getting results and with doing their best to run a successful institution. By contrast, the principal-pedagogue places the learner at the heart of the process. Their focus is on providing an environment which supports learning and supports all learners. They want to see that all learners are achieving the best they are able and that the school does everything to make this possible. In the first case, exam results seem to play an important role, as they are viewed as a symbol of the school's achievement, or as an indicator of how the school performs in comparison with others. In the second case, the results are viewed more as one element in a complex process which is used to obtain information about the learner and their progress. Results here are often seen as important, in that they provide objective, externally validated information, but not as dominant, and not as the main driving force in the educational process.

> Each year we have files with analysis of the learning outcomes, where the teachers write their comments. From the tests they do themselves, or those

that are done from outside, or by me, and those decisions are taken into consideration in the teaching programmes, and in the so-called targeted plans of work[1] [pol. *plan wynikowy*] for the next year. What still needs to be corrected, to be completed and what is happening…

Let's say a teacher comes to conclusions from those results for their work with others, because we can suppose that if something didn't work with that group, then it won't work with the next one either, right? So that's the way we look at accountability, the teacher evaluates their work. Obviously it's clear that when the principal is also informed of the learning outcomes he decides what was better than in the previous year, or what was worse than in the previous year, or what didn't work. But the principal bases that on the analysis that every teacher does, every exam subject teacher does themselves. They have the results, they do it themselves. The principal, of course, looks at those results, but they're dry figures. Whereas, the teacher looks at the figures on the basis of "that task didn't work, I need to find out why." "Did they really not have enough time for it, was it much harder, do they need to concentrate on it longer?" Qualitative analysis, right? While the principal concentrates on quantitative things. He compares the mean obtained, it was 30.5 points and this year it was 30. Obviously, the exam this year might have been harder, and easier in that year, and so on. But I try not to do it on the basis of I appreciate a teacher less because they got that result.

School 116

What is striking from this long extract is that there is not one mention of learners. Learning outcomes have been reduced to a bureaucratic process. The responsibility for "qualitative analysis" is entirely that of the teacher and the basic assumption on the part of the principal is that something is wrong and that it needs fixing.

Let us contrast this with the approach of the director-pedagogue.

The next account comes from a school which is part of a complex where the primary school and *gimnazjum* are linked together.

When the exam results come we get to know those results. Here we have documentation of a learner's progress, not just the results of the *gimnazjum* exam. Now it happens that in class 3 [primary] there's a test and in class six the external exam. We've been doing this for a few years. There were tests, internal diagnostic tests, and we tracked the progress of learners from class 3, what were they capable of. Each year we make a file, like this [shows the interviewer], for ourselves. And on the basis of diagrams, on that basis we compare what was the increase in learning from class 3, class 6 and then in the classes of *gimnazjum*… Of course, you realize that this is a rural district and there is no private tuition. As far as can, we [help] those children, simply, when the teachers, on a voluntary basis, ran an interest group, the kids very willingly joined it. And that's more or less what it looks like with us. I can show you, please look. In context, we do it,

1 These are teaching plans which give assessment requirements for each teaching goal at different levels of attainment.

it's called a contextual analysis of the results obtained by the learners in the exam in the school year. This was last year, for example. The results, results of the class three mock exam, results of the exam, mean scores, how those relate to the mean scores in each of the different subjects. Then there are the results of the exams in maths, the arts and foreign languages. Total number of points. And then we take factors like the learner taking part in a competition, taking part in extra lessons, ones organized not only by our school but, for example, by the local council, learner engagement, cooperation, how cooperation with the parents worked out, and of course the last factor which has an impact on the exam results, the learner's attendance. And then comments. And on that basis for every learner in every class we monitor and then we can see their progress.

School 84

Clearly, this is also a form of bureaucracy, but the focus of interest here is the learner, the progress they make and all the possible factors which might have a bearing on that. From this explanation we can deduce that the school implements a variety of measures to support learning, working with parents, encouraging participation in extra-curricular activities, tracking to see what progress an individual learner is making. One of the characteristics of the "principal-pedagogue" is their strong identification with the school, exemplified by the use of "we". In contrast to the previous director, who appeared to be at a distance from the teachers, and could perhaps be said to maintain a hierarchy, the principal–pedagogue is hands-on and fully involved in the analysis of results.

This has been an overview of the qualitative analysis of 115 interviews with heads of school. The decision to use a grounded theory approach and search for trends and patterns was made on the grounds that what is of primary interest here is to try to see what mechanisms are at work in these schools in response to the external examination in foreign languages, and what factors seem to have a bearing on why some schools respond in one way and not another. We shall explore these mechanisms and factors more fully in the next chapter.

5.6. Limitations of this study

This study attempted to be longitudinal but suffers from the fact that samples at the three different stages are different in size and conformation. This seriously limits the comparison that can be made between and across the phases. In response, care has been taken in drawing conclusions. A further weakness is the comparative lack of information from learners who are major stakeholders in the process of the introduction of a new national exam. In Phase III the Impact Study this was an artefact of adopting data from another study not designed specifically for the purpose of investigating test impact. In addition, interviews with learners, aged 13–14, yielded

less information than was anticipated. It appears to have been an effect of the type of questions used and interviewer technique, difficulties which did not emerge in piloting and were not anticipated.

What could be considered a drawback to this research is the predominant role of qualitative data and dearth of statistical analysis of empirical evidence. While this was deliberate, in that the aims were descriptive, it could lead to charges of subjectivity. Attempts were made to control for this through verification measures, but the fact remains that interpretation of the trends observed is individual and therefore open to claims of lack of objectivity. While every attempt was made to adhere to grounded theory principles, this could be seen as a weakness.

It was originally planned to compare information obtained from teachers and learners in the Impact Study (Phase III) with data from observation of lessons, but the scope of the study had to be reduced to make it feasible, consequently potentially reducing the reliability of the study. While an attempt has been made to present rich contextual data, the fact that the researcher was not able to physically visit the schools in the Impact Study is a disadvantage. Schools visits during the two earlier phases were found to be invaluable, but the scale of Phases Three prohibited this.

For reasons of space, detailed case studies of schools have not been included, other than brief references in the next chapter where data from teachers and school principal, in the context of their school, are contrasted. Investigating rich data from single schools would seem to be a particularly fruitful source of information about washback and impact mechanisms.

At the end of a longitudinal study it is easy to be wise in hindsight. Without doubt my own skills as a researcher have grown considerably in the process. Were I to begin again, I would be less likely to make the mistakes which, inevitably, have been made. It appears that researching washback and impact is, like the phenomena themselves, a dynamic process.

In the first part of the chapter we summarized the findings from Phase I the Baseline Study. This used teacher questionnaires, lesson observation and post-lesson discussion. Next we briefly described findings from the Initial Washback Study conducted twelve months later in 2009, which used learner questionnaires and a small number of teacher questionnaires. The largest part of the chapter presented data from Phase III the Impact Study, collected three years after Phase II in 2012. We began with data from teacher interviews, presenting first a quantitative analysis divided into four themes: Beliefs and attitudes towards the exam; Planning of work and choice of materials; Assessment in school and the exam; and Levers at work on teachers. Additional data on teachers was taken from a small number of questions on a questionnaire.

The next parts of the chapter presented quantitative data which gave an overview of information from learner interviews and questionnaires. School principals were the focus of the final section. Eight themes were

identified in the interview data: analysis of exam results; comparison of results with others; corrective programmes; systemic responses to exam results; external assessment as one part of a complex school assessment programme; changes learners noted in response to the new exam; and possible reasons for different responses to the exam from schools. This ends the presentation of analysis of the data. In the next chapter we move to discussion and consideration of the role of the external examination in English on the work of lower secondary schools in the light of what has been found in the analysis.

CHAPTER 6

Interpretations and discussion

In this chapter first the research questions posed at the three different phases of the study are discussed in the light of the findings described in the previous chapter. Next, on the basis of what has been learnt from this research, we present what has been discovered about the mechanisms of washback and impact and the factors which contribute to their coming into effect or not. We then suggest the implications of this research for policy makers, teacher education institutions, school principals, teachers, learners and their parents/carers. Finally we propose directions for further research and conclude with closing remarks.

6.1. Discussion of findings of preliminary studies

We will return briefly to the preliminary studies, the Baseline Study, conducted in 2008, which aimed to capture a picture of teaching and learning in lower secondary schools prior to the introduction of the new FL examination, and the Initial Washback study from May 2009, approximately three weeks after the examination, which investigated whether the first administration of the exam had had an effect on teaching and learning in class 3. We summarize the findings according to the research questions which were posed. This section serves as background to the Impact Study, which is the focus of this work.

6.1.1. Phase I The Baseline Study

In this section we will discuss each of the research questions in turn, with illustrations from the findings presented in the previous chapter.

1. What is the role of the course book in teaching?

Teachers were found to be strongly influenced by the course book they used. For a majority of teachers the course book was followed for a dominant percentage of lesson (80%) and course time (80%), according to findings from the questionnaire, and this was corroborated by what was seen in observation of lessons. Variation was noted in individual teachers

observed, which appeared affected by the aim of the lesson, the material in the book and by the experience of the teacher. The course book appeared to determine the frequency at which reading and listening tasks were done, according to options selected by teachers in the questionnaire, and also the type of comprehension tasks used.

2. Which aspects of language are the focus of teaching and learning?

When asked to ranks aspects of language according to what they considered important in teaching English in lower secondary school, teachers placed vocabulary, speaking and listening at the top of their lists. This suggests that oral communication skills were the focus of attention. However, in the lessons which were observed, the situation varied considerably, with only a few lessons giving affordance for learners to practise oral skills in pairs, or offer longer responses in class. Where lessons were closely focused on exercises from the course book, learners tended to have little opportunity to respond in more than a few words, or a short phrase. Some teachers, however, asked learners to read the whole sentence of an exercise aloud when they gave their response, although this was not generally the case. The language of the classroom also varied, with some lessons containing strong use of Polish. It was noted that when asking questions for information, which could be considered genuine communication, learners tended to use their first language, rather than English. Some teachers, did, however, ask learners to repeat their question in English, although this was not done in the majority of cases.

Teachers placed grammar, reading and writing at the lower end of the ranking, yet in the lessons which were observed, reading tasks and grammar exercises were found to feature strongly. When asked to say what aspects of learning English they considered most important for a learner, teachers ranked speaking, vocabulary and reading in the first places. Perhaps the fact that reading ranked higher in this list, than in what teachers themselves considered important, could be attributed to the fact that teachers are aware of the amount of reading present in the course book, which we have seen is the deciding factor for the majority of teachers in whether reading is done. There is no evidence to support this premise, but what does appear, is that there is a dissonance between what teachers profess is important and what was seen in the lessons observed. Of course, this was a small selected sample and it is inadvisable to set too much store by it. The discrepancy may, however, be a further indication of the powerful influence of the course book, which may decide what is done in class and how, even if the teacher believes that the focus of learning should be different.

3. How are reading and listening taught?

In the questionnaire teachers indicated that they support the development of reading through teaching strategies for reading, explaining step-by-step how to find the answer to a reading comprehension question, and through

substantial practice in reading. Translation and teaching how to use English-English dictionaries also featured, but less frequently. In the six lessons observed which featured reading, teachers were seen to use a variety of ways of supporting learners who were experiencing difficulty. Most commonly this was through asking a series of guided questions aimed at helping learners find the answer, but which at the same time provided a model of strategies the learners could use. This modelling was not, however, made overt. There were instances where teachers asked learners to explain why they chose a particular answer and how they arrived at it, which required learners to make the strategies they used clear. Where the answer was incorrect, teachers pinpointed the moment at which a wrong assumption had been made and asked focal questions to help the learner see where the mistake had arisen, and correct themselves. This was not always the case, as if an incorrect answer was given some teachers simply asked another student to provide the correct version. Generally speaking, however, the observation seemed to confirm the first claims made by teachers. However, only one teacher was seen to encourage learners to use dictionaries, suggesting that the questionnaire response may be misleading. Translation was commonly seen to be used, almost exclusively in many lessons, to give meanings of unknown words, both orally and on the board, where bilingual word lists were common, suggesting its use is more frequent than some teachers appear aware.

Tasks for reading were almost always determined by what was offered in the course book. It was rare to see pre-teaching of any words which appeared in the text. Most commonly teachers first asked learners to read and then invited them to ask for the meaning of unknown words. In the light of recent findings from research on reading, which suggest that guessing unknown words from context is possible if approximately 98% of the text is understood (Schmitt, Jiang & Grabe, 2011), it is doubtful whether this is an approach which benefits learners who do not have well-developed lexical resources. Learners were observed to have great difficulty with comprehension caused by poor understanding of vocabulary. Wild guessing, based on taking a single word from a comprehension question and searching for the same word in the text, was seen to be commonly applied. In post-lesson discussions, teachers confirmed that learner problems with comprehension seemed to stem from limited knowledge of vocabulary. Of note is that where teachers took time to introduce the topic of the text, do pre-reading activities and to pre-teach vocabulary, learners were visibly more engaged in the comprehension activities and appeared to have less difficulty.

For supporting listening, teachers most frequently selected the options "write problem words or phrases on the board" and "play the recording again," and these strategies were observed in all of the three lessons which featured listening. In the third lesson, where learners had extreme difficulty, the teacher made use of the tape script, which was an option infrequently

selected in the questionnaire. As with reading, the comprehension tasks for listening in the observed lessons came from the course book. In the previous chapter we described in detail how pre-listening activities were used to support the learners in two of the observed lessons, and how in the third lesson, where no pre-listening tasks were done, the comprehension task had to be abandoned.

To sum up, it would appear that teachers, and particularly less experienced teachers, have some difficulty with the treatment of vocabulary in both reading and listening texts. It may be that there is a generation gap, with older teachers rigorously doing pre-text work, while their younger colleagues ignore, or minimalize this. Whatever the reason, comprehension of both reading and spoken texts was observed to be very challenging for many learners.

4. How frequently are reading and listening practised?

As we have already indicated, the majority of teachers stated that the frequency of practice of reading and listening was determined by what was in the course book. Questions can be raised, however, about the practice of listening. In the lessons which were observed, listening to a recording featured in only three of the 14 lessons, and in two of these the aim of the listening task was not development of the skill of listening, but elicitation of a language point, which was subsequently used for practice of grammar. Two issues arise, the apparent low frequency of listening in comparison with reading, if it is the course book which determines when listening is done (three of the 14 lessons observed featured listening, as opposed to six with reading), and the aim of the listening task. The one lesson which focused on more extensive listening showed that learners had great difficulty with the task. This could, very tentatively (as it was only one example, in a problematic group), suggest that such types of listening are done infrequently. A further issue here relates to the language of the classroom. If Polish predominates in a lesson, learners are not being exposed to the same amount of listening as when English is the main classroom language. Listening is not only developed from tasks based on recordings. It would seem that some teachers may have forgotten this.

5. What is assessed?

Information about this was obtained only from questionnaires, where it was found that teachers claimed to assess all of the six aspects of language given. Vocabulary and grammar were selected marginally more frequently than writing and speaking, with reading and listening slightly behind. Teachers tended to use tests from published sources, rather than produce their own. In the observations, one lesson was seen where the teacher checked through a test done with a second class. This was a practice exam test, taken from a newspaper.

6. What is the teachers' knowledge of and opinion on the new exam?

In the questionnaire, most of the teachers (78%) claimed they were familiar with the new exam and approximately two thirds of these (68%) stated that they had done practice tests with their learners. In post-lesson discussions following observations, several teachers mentioned that they had done practice tests with their learners, with the aim of checking the learners' skills and to help plan what needed to be done to support them. The *Informator* produced by the Central Exam Board (2007) contained sample tasks, but no sample test papers. Teachers were consequently curious what the actual exam might be like and how their learners would fare.

Teachers were of different opinions about the perceived difficulty of the examination, with 55% considering it would not be difficult for their learners and 45% that it would be hard. If we look ahead to the results of the first examination (see Chapter 3) we understand that both groups were accurate, as the first exam produced a bi-modal distribution, which can be interpreted just as these teachers predicted, that the exam was easy for some of the students and difficult for others.

We will not summarize the findings here, as this might inadvertently place emphasis on some aspects rather than others, which is not the intention. The baseline study is to serve as a point of reference for comparison with what happened after the introduction of the examination. In this way we hope to be able to gauge what has changed and how.

6.1.2. Phase II The Initial Study: preliminary effects of the examination

This study took place in 2009, one year after the Baseline Study and approximately three weeks after the first examination. The participants were those who had taken the examination, learners in class three lower secondary, teachers who had taught them and others who had taught the same year group.

Each of the research questions will be discussed in turn. As detailed analysis was presented in the previous chapter we will not repeat this, but focus on the second part of each question, the comparison with the baseline study. The first questions relate to learners, who had not been investigated directly in the baseline study, except through observation of lessons.

1. What aspects of language do learners consider important in learning English in *gimnazjum*? How does this compare with information from the baseline study?

When compared with findings from the baseline study (see Table 6.1) we see that teachers accurately predicted what was important for their learners, with the ranking from 2008 identical to that of 2009. There is no indication of any change having taken place over the year.

Table 6.1. What is considered important in teaching and learning in lower secondary school: Initial Washback and Baseline studies compared

	Initial Washback Study	Baseline study	Baseline study
Rank	Learners 2009	Learners according to teachers 2008*	Teachers 2008
1.	speaking	speaking	vocabulary
2.	vocabulary	vocabulary	speaking
3.	reading	reading	listening
4.	listening	listening	grammar
5.	grammar	grammar	reading
6.	writing	writing	writing

* note that teachers were also asked to rank other aspects. These were removed from the list.

2. According to learners how frequently are reading and listening practised? How does this compare with information from the baseline study?

Detailed information on the frequency of practising these skills was given in Chapter Five. If we compare the learners' responses with those of the teachers from 2008, we note that both groups chose "according to the course book" as the best description of the frequency with which the skills of reading and listening are practised. The strong role of the course book was also noted in lessons on skills observed during the baseline study. The situation appears unchanged.

3. According to learners what is assessed? How does this compare with information from the baseline study?

Information obtained from the questionnaires placed grammar clearly above the other aspects listed, with vocabulary in second place, closely followed by reading. Consequently there is no indication of washback from the examination on what is tested. Neither grammar nor vocabulary were overtly tested in the 2009 examination. Test papers included reading, listening and communicative tasks [pol. *reagowanie językowe*], which indirectly tested speaking. If we were seeing washback we might expect more evidence of grades for listening, which is not the case. However, if we compare what teachers in 2008 claimed they grade with what these learners said, we notice striking differences for writing and speaking between learners and teachers, with much higher claims from teachers than from learners (see Table 6.2). Clearly this should be treated with care, as these are two separate samples. These teachers did not teach these learners.

What a teacher chooses to assess is influenced by what they consider important, and this varies from individual to individual. If we consider that teachers in 2008 placed grammar fifth on the scale of importance, we

might, however, wonder why it is that grammar came out a clear winner in terms of what is graded, according to learners. The reason may lie in the fact that teachers in 2008 declared a strong preference for tests from published sources, which tend to feature grammar. Course books often follow a grammatical syllabus and consequently grammar, together with vocabulary, are the prime aspects tested (see Ellis, 2010 for analysis of course books used in *gimnazjum*).

Table 6.2. Comparison of what learners claim they are graded on and what teachers say they grade in percentages

	Learner responses 2009 (n=229)	Teacher responses 2008 (n=47)
Grammar	80	93
Vocabulary	65	100
Reading	64	79
Speaking	57	85
Translation	52	–
Listening	46	76
Writing	41	89
Project work	32	13

4. Is the exam important for the learners?

Answers to this question are not clear cut. More learners thought the exam was important than did not, with slightly less than half positive when values were aggregated (44%), as opposed to approximately a quarter (23%) negative. The remaining third of the respondents (33%) were ambivalent. In context, we must remember that in 2009 the points from the examination did not count towards selection for upper secondary, unless learners were applying for schools with intensive English programmes, where they could be taken into consideration. These type of programmes were not widely available, so this was unlikely to have affected a large number of respondents. It is therefore understandable that for some learners the exam was not particularly important. This is not yet, in effect, a high-stakes test.

The next questions concern teachers. Detailed findings were presented in the previous chapter and they will be only briefly summarized here.

5. What (if any) changes did teachers make to their teaching in class 3 because of the exam?

Nearly all (95%) of the teachers stated that they had made changes in the teaching they did in the third class of lower secondary school because of the examination. Teachers were found to have made changes to the teaching materials they used in many cases, making use of practice tests (81%), and

perhaps adding an exam preparation book (62%). Changes were also found in classroom activities with most of the teachers (81%) reporting an increase in the number of listening tasks they did in class. This increase may go some way to corroborating what was hypothesized after the baseline study, that listening was then done with lower frequency than reading. Others, although fewer (52%), claimed they increased the amount of reading comprehension done. However, we do need to remember that this is a small sample (n=21). About a quarter of the teachers (28%) reported that they had changed the type of tasks done with reading, and about one fifth (19%) that they had changed the types of listening tasks they did.

6. What according to teachers is the learners' attitude to the exam?

Teachers were asked to give their opinion on how the learners felt about the examination.

Responses varied and covered a spectrum from "It really motivated them to learn", selected by 29%, to "They showed some interest in it" at 50%, (making 79% of the attitudes noted positive), through "They were indifferent to it" (12%) and "They said it was not important to them" (8%). We saw that some 23% of the learners gave negative responses as to the importance of the exam, which appears to confirm this. Looking ahead to the Impact Study, in an article based on analysis of data from one of the questionnaires from BUNJO 1, Ellis (2013) found that a group of 12–15% of learners in the first class of *gimnazjum* had rather negative attitudes towards learning English in school (n=4343). It would appear that in this small sample from the Initial Study, the finding that about 20% of the learners are indifferent or negative towards the exam may also reflect the fact that there would seem to be a group of learners in lower secondary school who have ambivalent or negative attitudes towards school and education in general. Lower secondary school is part of compulsory education and caters for everyone, regardless of predisposition. If we compare teacher perceptions of learner responses to the exam with what learners themselves said, we find some discrepancy. Only 44% of the learners felt the exam was of importance to them, and 33% were indifferent. It is possible, however, that the difference lies in the questions which were asked. It is possible, in theory, to show some interest in an exam (the statement given to teachers) and yet remain indifferent as to its importance. This is a design flaw, which should have been foreseen.

7. What effects did teachers notice in their learners' behaviour which they attribute to the exam?

Teachers were asked to respond separately to two questions about what effects they noticed among less able learners of English and among those who are most able language learners on the basis of five options given. Consistent with the thesis suggested in response to the previous question, around 20% of the teachers reported that less able learners showed no

change (23%). Changes in less able learners were noted, however, by around a quarter (27%) of teachers, who reported they were trying harder, and asking for help more often. A smaller number of teachers (13%) reported these learners were more engaged in lessons. This very tentatively suggests, due to the small number of teachers who took part, that the exam appeared to have a positive effect on some learners who are less good at English.

The most able language learners were also reported as responding to the exam, appearing to be more engaged in lessons (according to 37% of teachers) and trying harder (28%). A smaller number of teachers (19%) reported that these learners asked more questions in class. Thus, there would appear to be tentative indications that the exam was also having a positive effect on the most able learners of English. Without interview data from either teachers or learners to investigate this further, or observations of lessons taking place over time, we can do no more than indicate signs of a possible tendency.

8. What is the teacher's attitude to the exam?

From a choice of three statements, 71% of teachers declared that the exam in English had had a positive effect on their learners and their teaching. 9% felt that the effect had been negative, while the remainder reported that the exam had had no effect.

Summary

Effects felt shortly after the introduction of the examination seem to focus predominantly on changes in teaching materials and an increase in the number of listening and reading tasks done in class, reflecting the importance teachers perceive these skills have in the examination. The course book has been found to play a strong role in other studies of washback, e.g. Andrews (1994b), Lam (1994), Read & Hayes (2003), Spratt (2005). Wall & Horák (2011) not only noted the importance of the course book, but also saw that teachers changed the amount of time spent in lessons on the skills which they thought important in the TOEFL test, a response similar to that observed here in the Initial Study. There is some evidence that changes also took place in the type of tasks done with reading and listening, although this applies to only between one in four and one in five of the respondents. However, deeper changes, which might be indicated by changes in what is perceived as important in language teaching and learning at this level were not found. This is consistent with findings made by Cheng (2005) who reported changes in teacher behaviour were superficial rather than substantial. This may reflect what Chapman and Snyder (2000) referred to, that for change to take place, teachers must have a profound understanding of the examination before they are able to fully respond to it. It may be the case that a single administration of a new examination, to which changes were being made up to five months

prior to its administration, is not enough. Analysis of the results and reports on the examination from the exam boards at regional or national level had not been published at the time the data was collected. Teachers' understanding of the examination at this time was based entirely on their personal experience, on discussion with colleagues and with their learners, or on impressions from the media. There may also be a time factor at play, as the data was collected very shortly after the examination took place. Deeper reflection takes time to take effect. Teachers at this stage of the school year were engaged in carrying out final assessment of learners, settling final grades for the semester and determining which of their learners could continue to the next class. We did, however, see that three quarters of the sample reported that they would make further changes to their teaching in the coming year as a result of the exam. This suggests that teachers are aware that this is a process in which they are engaged, which requires further, and perhaps more profound response.

6.2. Discussion of findings of Phase III The Impact Study

We move on now to the key part of this work and discuss what was found in the Impact Study. This took place between March and May 2012, the year in which there was a revision to the examination, which now has two levels, basic and extended. The second important event was that, for the first time, points from the examination, but only those from the basic level, are counted towards selection for upper secondary school. All young people in the third class were required to take the exam at both levels, but in this year they were allowed to choose which language they took at the extended level.

As with the previous sections, we will first discuss the research questions, taking each of the groups of stakeholders separately, starting with teachers.

6.2.1. The effects of the examination on teachers

Data from teachers includes analysis of transcripts of interviews (n= 301) and extracts from questionnaires (n=380). Detailed descriptions of instrumentation and the coding procedure were given in Chapter 4. In Chapter 5 we presented the findings. Let us now move to discussion.

1. Is choice of teaching materials affected by the exam? If yes, in what ways? What reasons do teachers give for this?

On the basis of other studies of washback (e.g. Andrews (1994b), Lam (1994), Read & Hayes (2003), Spratt (2005), Wall & Horák (2011), we expected that if washback was taking place, then one of the first places it would be discernible would be in the teaching materials. We noted in the baseline study that the course book was found to play a central role in

teaching in lower secondary school, as the focal point around which lessons are planned and carried out. We also found in the baseline that teachers tend to draw on published sources for tests, with only a small number of teachers preparing their own. In the Initial Study many teachers reported that they added an exam preparation book to their materials in class three, and made use of published practice tests. In sum, even before the Impact Study began there were some indications of washback on teaching materials.

At this point it is worth adding a little information about the role the publishers play. The largest educational publishers of foreign language course books in Poland are, in effect, the main providers of in-service teacher education. They offer free conferences and workshops for teachers in all regions of the country, several times a year. As soon as information about the exam reform was available, the publishers began to hold conferences to inform teachers of the changes, offer suggestions how to deal with them and advertise their materials. For many teachers this was the prime source of information about the changes. To be fair, the sessions were very well researched and prepared, and speakers were often associated with the examination boards. We should not, however, forget that the driving function of such events for the organizer is to sell books.

In the interview data from the main study, we saw that nearly three quarters of teachers (71%) claimed that the exam had influenced their choice of book, or books. This was found to refer to materials for all classes of lower secondary and not to be restricted to the final year. Around a quarter (28%) of the sample mentioned the role of the exam in the course book, of which the largest number referred to the book containing sections focused on the exam, as introducing the types of tasks the exam would contain, or, more generally, as gradually introducing the exam (see also extracts from Teachers 79/1, 80/3, 103/3 in Chapter 5). The majority of these teachers reported this as positive information. However, a small number, around 10% of this sub-group, appeared to see this as negative, apparently feeling forced by the course book into focusing on the exam against their will.

It was found that the level of the group plays a decisive role in determining the teacher's choice of material with regard to the exam. Where the group was perceived to be at a level which the teacher considered would mean they would have little difficulty with the contents of the exam, then the course book was reported as being chosen to be in keeping with the core curriculum. By contrast, with groups which were perceived as weak, and consequently seen as being far from the target level of the exam, teachers reported choosing exam preparation materials and following them closely (see extracts from Teachers 112/1, 116/2 in chapter 5). This factor of tailoring choice of materials strictly to the exam, according to the level the teachers perceives learners to be with regard to the exam, appears to be new and has not yet, to my knowledge, been reported elsewhere in the literature on washback. Green's (2007) model hypothesized perceptions of

the difficulty of the test as a factor in washback intensity. Here, however, it appears rather to be a factor affecting whether washback takes place, or not.

Table 6.3. Choice of teaching material relates to perceived level of group with regard to the exam

	Factor perceived to be at work	**Action taken by teacher**
	Teacher perception of level of group	
Choice of teaching materials	Teacher thinks group is at appropriate level for exam	Decision of choice of materials based on curriculum
	Teacher thinks group is below appropriate level for exam	Teacher selects examination preparation book

In order to discover why teachers feel the need for course books which contain exam-related material, we need to consider their beliefs and attitudes about teaching and about the exam. We will first discuss these and then return to the question of how these impact on choice of materials.

2. What is the attitude of teachers as regards the exam in their teaching? What reasons do they give for this?

A number of factors appear to come into play, which seem to influence how the teacher regards the exam in their teaching. These are illustrated in Table 6.4. It should be noted that the order given in the table does not represent the magnitude of the factor. Each of these factors will be discussed in turn. First is how the teacher perceives the examination.

Table 6.4. Factors perceived to affect teachers' attitudes to the exam

Teacher attitudes to the exam	**Factors perceived to be at work** Perception of exam Attitude to exam • Sense of being accountable • Social conscience • School location, situation Attitude to course book • tool or driver

Perception of the exam

By "perception of the exam" we mean what teachers appear to believe the exam is testing. There are numerous examples in the data of teachers attempting to describe the exam and how they understand it. It is also clear that there are very differing interpretations. Teachers seem to have no

difficulty seeing that reading and listening are being tested and appear to view these as being of central importance. A problem, however, seems to appear when it comes to the other tasks in the basic level exam, described as 'communication' [pol. *reagowanie językowe*]. Teachers are divided as to what they appear to think these tasks test, with a number not appearing to recognize them as an indirect test of speaking, which is the intention of the Central Examination Board. These items test recognition of the different language functions which are given in the core curriculum, and require learners to select appropriate questions, or responses, in answer to a prompt given in the context of an everyday situation. Some teachers (although not a large number) were found to report that they do not teach speaking, or do not have time for speaking, because they believe this is not tested in the exam. Others regretted the lack of an oral component to the exam and used this to justify the fact that they focus on written skills in class. These responses seem to indicate that, as these 'communication' items are for the most part read (although one task usually includes listening to a prompt and selecting an appropriate response from a written selection), they are perceived as testing something which is not speaking. This raises the question of the validity of this part of the exam. If the items are not perceived as testing speaking, then at the very least there is an issue with face validity, at worst an issue with content validity. The question of content validity could be raised if the teacher's interpretation of what these tasks are testing leads them to narrow their syllabus and reduce the amount of speaking in the course (reported in 4% of teachers interviewed). If the exam purports to be testing the contents of the core curriculum and the teacher believes that one large section of the core is not included in the exam, to the extent that they limit their teaching of it, then the teacher is misunderstanding the content of the examination and so there is an issue with content validity. It is not enough that the test specifications indicate the relationship between the core curriculum and the exam content, or exhort teachers to cover the whole core, if the intent of these 'communication' tasks is not clear to all teachers and their learners, for we saw in the data that teachers reported that some learners also consider speaking is not tested in the exam (see extract from Teacher 103/1 in chapter 5). It should be stressed, however, that this difficulty was found in only about 11% of the sample.

Where teachers perceive the exam as being closely related to the core curriculum and understand that the 'communication' tasks are communicative skills 'tested in another way' as Teacher 104/3 put it, then there are no issues. These teachers understand that the learner becomes able to select the appropriate response or question, not by doing lots of practice exercises from an exam preparation book, but by engaging in dialogues in everyday situations, or by listening to similar conversations and then holding their own exchange with a partner. It is not clear from the data how the teacher comes to this understanding, or where the misconception

arises. One possible factor may be the teaching material itself. We saw in the baseline study that teachers depend heavily on the course book for their planning. In part of the learner interviews not included here, work with the course book was the characteristic learners most frequently reported when asked to describe a typical English lesson in their school (Ellis, 2014: 29). Data from observation of lessons that accompanied BUNJO 1 (the source of data for this part of the work), which involved approximately one third of the teachers who were interviewed, not reported here for reasons of space, corroborated this impression, with 160 of the 228 lessons observed (70%) categorized as based wholly on the course book, or accompanying practice book. If the teacher simply follows the book, then it is possible that they do not think about what will help the learners be able to do the communication tasks, and treat them in the same way as a grammatical or lexical exercise. As we have already indicated, it is unlikely this approach is going to help learners develop the skills needed to do this type of task.

Attitude to the exam

We saw that the exam featured strongly in the aims teachers set for themselves in their teaching in *gimnazjum,* with some two out of three (62%) referring to it. Most of these teachers (81%) see preparing their learners for the exam as one of two main aims in their work and approximately one teacher in five (18%) as the sole aim. Teachers varied in whether they saw the exam as the primary, or secondary aim coming after the ability to communicate. The reasons given by those who placed communication first included reference to the aims of the core curriculum, which place the ability to communicate in everyday situations as the main aim, or more general references to learners' real life needs, both currently and in the future (see extract from Teacher 55/3 for example).

The reason behind placing the exam as one of the aims of teaching appears to lie in one of three factors which seem to form a sub-set of 'attitude to the exam': a sense of being accountable, having a social conscience, and issues relating to the location or situation of the school. Where teachers felt under some obligation for their learners to get good results in the exam they placed it as one of their aims, and where this obligation was experienced or perceived as pressure, then the exam became the primary aim, or, in more extreme cases, the sole aim. This seems to be a continuum, along which the importance of the exam moves in relation to the strength of the pressure the teacher feels upon them, and the source of that pressure. The greater the sense of pressure and the more important the source of that pressure is perceived to be, then the more important the exam becomes in the teacher's aims. Thus, if the teacher rationalizes to her/himself that the exam is important for the learners, as it helps determine which upper secondary school they will attend, then the perceived importance of the exam is coming from the teacher themselves, and is intrinsically motivated.

In this situation it seems to be compatible with having another aim, which is to develop communicative skills. If the teacher continues their rationalization along the lines that the learner needs good communicative skills in English in their life, for their further study, or for future work, then it seems that the importance of the exam as an aim will be subordinated to communication. The teacher's sense is one of obligation to the learners, but without pressure.

By contrast, if there are forces at work on the teacher, such as signals from the principal that the exam results are important, if the teacher feels that their work is being evaluated on the basis of the exam results, or if the learners themselves indicate that it is the exam which is the most important motive for them, then the teacher feels pressure external to themselves. Where they are able to understand the reasons behind this pressure, and why the parties concerned are exerting it, then they appear to be able to accept the situation and combine it with their own aim, which is usually for their learners to achieve communicative skills. The exam, however, in this case will take precedence. Where the external pressure is perceived as considerable, or where the teacher cannot understand the reason for it, or where they are unable to reach a workable compromise, which seems to be accompanied by a sense that they have little control over the situation, then the teacher expresses negative attitudes towards the exam, or frustration with the situation. It would seem that in these cases that the exam becomes the teacher's sole aim.

The pressures at work on the teacher as regards the exam do not all appear to be external. Some of the pressures seem to be internal to the teacher and to be related to their beliefs about their role as an educator. These have been labelled 'social conscience'. It was noticeable in the data set that some teachers refer to feeling a moral obligation to prepare their learners well for the exam. This is expressed in terms of giving their learners the best possible opportunities in life, helping them get a foundation on which they can build language skills for the future, or, often with reference to socially disadvantaged pupils, trying to be equitable, by making up for deficits from primary school, pushing the learners towards good results and so giving them more chance of a better secondary education.

A variation on this 'social conscience' seems to apply particularly to teachers in rural settings, where there seems to be a clear awareness that the learner has no access to additional support outside the school, either because this is not available, or not economically possible, or because the learner has a difficult home situation. This has been labelled 'school location'. Here teachers, apparently fully conscious of the difficulties such youngsters may experience in the future, when they will have to compete against more advantaged candidates from towns, express a sense of strong obligation to help their learners get good exam results, because they want to do everything possible to help and support them. In this way the strength

of the pressure felt can be exerted by the teacher themselves, in response to their perception of the local situation. One final variation of this, labelled 'school situation', seems to be where the teacher perceives that the school needs to get good results in order to have good standing in the community. Some teachers seem to identify themselves with the school, in the sense that if the school gets good results, then this reflects on their own work, as someone associated with a school that is perceived to be successful, is thought to be a good teacher. Or, conversely, it is thanks to the work of good teachers that a school gets good results and gains a good reputation. In either of these cases the exam results matter, and so preparation of learners to do well in the exam becomes a prime aim.

Once the teacher has placed preparing their learners for the exam as an aim of their teaching, then the logical consequence is to make the examination the criteria for selection of the course book. In this way we return to the first research question and close the argument.

A final factor appears to be the approach of the teacher to the course book. We noted in the interview data that there were two main groups of teachers, those who regard the course book as a tool which they use in their teaching, and those who view the course book as the driving force behind their teaching. The difference in these two approaches appears to lie in the degree of autonomy the teacher exhibits. Where the course book is a tool, the teacher describes an active process of decision-making as to how they make use of it. This includes selecting from it, choosing what aspects to highlight, omitting certain parts, or reducing their salience. Where the course book is the driver, however, the discourse of the teacher is quite different. They express feelings of resignation that they are forced to teach in a certain way, because this is "in the course book". They even express frustration that the course book approaches certain features in ways that are at odds with how they would teach if they were able to choose, suggesting that they are completely passive and allow the course book (or the accompanying support materials) to make their decisions for them. This approach was even noted among teachers who stated that they had had a voice in the choice of the course books used in their *gimnazjum*, a question asked specifically in the interview, which would suggest that it does not result from lack of freedom of choice of the book, but from the teacher's attitude, beliefs, or character in general. A further possibility is that this results from inadequate analytic skills, or lack of training, which may mean the teacher has difficulty determining the aim of a piece of material, or in critically evaluating it. The outcome is that they "do the book", blindly following the instructions, exercise by exercise (see for example, extract from teacher 103/1 in chapter 5). What could be described as the course-book driven approach was noted in about one in four of the teachers (28%) interviewed, which indicates the presence of a noticeable trend, and is a cause for concern.

With relation to the exam, the more autonomous teacher describes working selectively with course books for class one, glossing over the sections in them which are exam-focused, in some cases omitting them, or, in other cases, simply using them to signal to learners that this is the kind of task they will meet in the future and that it will be focused on in more detail in class two or three. The course book-driven teacher, by contrast, highlights these tasks from the beginning of the lower secondary school, apparently believing in their importance because they are there, in the book. This fact may also contribute to their believing in the central importance of the exam. In many cases where teachers explained that the exam played a role in teaching throughout the school, the presence of exam-related sections in the course books for classes one and two were cited.

3. How do teachers view the relationship between the exam and the core curriculum?

We have already considered the relationship between the exam and the core curriculum in the section on the Perception of the Exam. There is a dichotomy between teachers who understand that the exam assesses the core, and those who consider that it only assesses part of it. We found some evidence in the data of teachers narrowing the syllabus for this reason, although there were very few instances reported. Most of these limited teaching of speaking, but there were also two instances of teaching of grammar and functions being limited with low level groups.

4. Which aspects of language do teachers consider important in teaching and learning in *gimnazjum*?

Answers to this question were taken from the questionnaire for teachers (n=380). In Table 6.5. we compare the ranking of responses of teachers to a comparable question in the 2008 baseline study and answers from the 2012 Impact Study.

Table 6.5. What is considered important in teaching and learning in lower secondary school: Baseline and Impact studies compared

	Baseline study	Impact Study
Rank	Teachers 2008	Teachers 2012
1.	vocabulary	speaking
2.	speaking	vocabulary
3.	listening	listening
4.	grammar	reading
5.	reading	grammar
6.	writing	writing

We see from Table 6.5 that there appear to be some changes. First, we must stress that care should be taken, because of the fact that these are two different samples of teachers, which vary considerably in size, and consequently any conclusions must be extremely tentative. The Baseline sample is 45 teachers from 25 schools in one region of the country, while the Impact Study sample is 380 teachers from 120 schools throughout Poland. The first change we note is that in the Impact Study speaking is ranked in first place, as opposed to second in 2008. The other change is that grammar has been demoted one place, in favour of reading, which rises one place. The primary importance on speaking in the Impact Study seems to be consistent with the fact that communicative skills featured in the aims for teaching given by 81% of the teachers interviewed. This is in keeping with the main aim of the national core curriculum, which teachers are required to cover. Teachers' beliefs about what is important in teaching at this educational stage seem not to have been noticeably affected by the examination, although we could speculate that reading has risen for this reason.

5. Are all four skills assessed with similar frequency? If not, what reasons do teachers give for this? Is there evidence that this results from the exam?

In Table 6.6 we see a comparison of what teachers said they assessed in 2008 during the Baseline Study and the Impact Study results from 2012. The questions for these two studies were different. In 2008 teachers were asked to mark all aspects of language that they assess from a list, while in 2012 they answered on a 5-point Likert scale about the frequency with which they assess different aspects of language. An additional problem is that the different aspects in 2012 were not all separated, but grouped as "tests of grammar and vocabulary" and "tests of reading and listening". To obtain the percentages given, the positive values *I totally agree* and *I partly agree* have been aggregated and are expressed as a percentage of the responses for that question. The next issue is that these are two different samples, and samples of different size and composition. While all this is problematic, and any trends visible could be questioned for this reason, we notice that there appears to be considerable similarity between the two studies, with the exception of speaking, which has decreased considerably.

Table 6.6. What teachers claim is assessed: aggregated scores in percentages: Baseline and Impact studies compared

	Vocabulary	Grammar	Speaking	Reading	Listening	Writing
2008	100	93	85	79	76	88
2012	96		51	78		88

Let us now turn to evidence from the teacher interviews. In chapter 5 we discussed the problem of trying to determine whether speaking is assessed from what teachers said in response to the question "What do you give grades for?" and noted that answers were often not clear, because of how the teachers spoke about this aspect of assessment. We discounted references to answering questions, giving responses in class, or reading aloud, on the grounds that it was not clear that it was the *skill* of speaking that was being assessed in these cases. Three out of five teachers in the sample (183/301) were deemed to say that they assessed speaking, and only three teachers (1%) specifically stated that they did not. We also noted that there was evidence that speaking was not often assessed, with mentions of this happening once or twice a semester, which lasts around 18 weeks. The factor of the school assessment scheme seemed to play a role here. Where teachers described a weighted system, oral grades tended to be awarded half the weight of written tests. This is a scheme which applies to the whole school across all subjects and to which teachers should conform. Other teachers mentioned a subject assessment scheme for foreign languages, which also seemed to set a lower number of grades for speaking than for other aspects. There is, however, no clear evidence that this is caused by the exam, rather that it is an issue with the logistics of oral testing in class and the time teachers feel this involves. We conclude by suggesting that where the majority of teachers place an emphasis on communicative skills as the aims of their teaching, the fact that speaking is apparently less frequently assessed is worrying. This is increasingly so if the assessment of speaking, has, in fact, decreased in the time since the Baseline Study. Yet we cannot say that this has been caused by the exam.

6. Does the exam affect the planning of teachers' work? If yes, in what ways? What reasons do teachers give for this?

There are many clear examples in the interview data that the exam is having an effect on the work of teachers in lower secondary school. Let us first consider if this could have been caused by the design of the interview question. The question was formulated in the following way: What role does the foreign language exam in play in the planning and choice of materials in teaching English in your school? It is possible that the interviewee could have interpreted this as the interviewer expecting them to say that the exam has a role, and consequently could have produced biased responses that a more neutral dichotomous question such as: Does the exam play a role in planning, or the choice of materials…? might not have. Yet, we did note a considerable number of responses (52%) in the data set, where teachers disagree that the exam is of central importance, which suggests that this is not an issue. Here is a summary of the findings of interview data from three teachers in one such school, coded as school 7.

The first teacher is strongly guided in her planning and work by the course book, which prepares learners for the exam. In several instances, she does however, refer to developing speaking, and places creating a positive atmosphere in class foremost as an aim for the learners. The second teacher, by contrast, seems more motivated by the core curriculum, recent changes to which she refers to several times. While she does mention the exam and wanting to make sure her learners cover everything needed for it, she stresses that getting her learners speaking and using language in real life is most important. She points out that the exam does not test speaking, implying a clash between the core curriculum and exam content, but this is not said overtly. The third teacher, like her colleagues, mentions the exam in passing, but is clearly focused on general language development. When asked about her aim in teaching in *gimnazjum* she comments that

> recently at conferences and other sorts of meetings there's focus on how many points the learners get in the exam. They check things from the point of view of the number of points, because then, you know there's the ranking of the school and so on. But I don't think that only teaching for the test is the most important, because in general the most important thing is learning English, which will help them cope later in different real-life situations, right? In everyday life. So preparing them to use that sort of language, above all spoken language and in everyday life.

The teacher indicates that although there are pressures from others to focus on the exam results (although she does not make it clear who these others are, or whether this is in-school, or from outside), her agenda is for learners to be able to communicate in everyday life. Although not overtly referring to it, she is in fact citing the main aims of the core curriculum.

There seems to be a strong commonality in the responses of the three teachers. They appear aware of the exam and various aspects relating to it, but despite this are motivated primarily by the core curriculum in their work.

This school represents what could be described as a rational and balanced approach to the exam, in terms of planning decisions. While the three teachers are aware of the exam, and of some of the expectations associated with it, they all appear to be teaching according to what they believe is important, which is to develop communicative skills and have given the exam a lesser role. This type of approach, in various guises, is expressed by slightly more than half the teachers in the sample, while 48% mentioned ways in which the exam affected their planning. However, in a separate question, 83% of teachers agreed that the exam had an influence on choice of materials and planning throughout all three years of school, although the main focus of these responses was the influence of the exam on the choice of materials.

The exam was seen to affect the planning of teachers' work in several ways. The first is how the teacher approaches preparation of their students

for the exam. We saw in the descriptions in Chapter 5 how some teachers described a process of gradual adjustment of their learners to various aspects of the exam, which we named "attuning learners to the test". This can be divided into the following areas:

Attuning learners to:
- task types
- test content
- test level
- test requirements, procedures

We noted that many teachers appear to have made use of the Test Syllabus [*Informator*] published by the Central Exam Board and have carefully noted the types of tasks which might appear in the different parts of the exam. We also saw how these task types feature strongly in course books and that analysis of their similarity to the exam had been one of the criteria for selection of books mentioned by teachers. Task types are among the most frequently mentioned aspects of the exam referred to. Teachers rationalize this by wanting to make sure that their learners are well-prepared for the type of tasks they might meet and have strategies for dealing with them. Others speak of wanting to reduce learner stress by helping them understand what task types the tests will contain.

The content of the test, also apparently analyzed by teachers on the basis of the *Informator,* was noted as another criteria for choice of books, but also for course planning. Teachers are concerned to make sure that all elements of the syllabus are covered, so that learners will be ready for whatever elements are in the test. The fact that the exam in 2012 includes two levels, and that the extended level in effect extends what learners need to have practised, beyond what had perhaps been the focus in previous years, seemed to be felt by some teachers as a source of pressure. Indeed, it was cited by some interviewees as justification for starting the exam focus as early as possible. It is the grammatical syllabus that seemed of particular concern. Other teachers felt constrained to push learners to work on more advanced vocabulary to prepare them to deal with more difficult texts. The inclusion of an open writing task for the first time had encouraged teachers to plan the introduction of writing, which some admitted they had neglected previously, on the account of it not being part of the original version of the exam. Teachers seem in agreement that writing is not something which can be left until class three to introduce, and so describe a gradual introduction to open writing tasks, with increasing exam focus as the learners progress up the school. The Test Syllabus is again mentioned in how these writing tasks are formed and how they are assessed, with some teachers using the official rating criteria for marking writing. Some teachers specifically extended their teaching programme to include writing, indicating that the exam has been the reason for this, while some others mentioned beginning to include writing, but at the cost of, most often, speaking.

The level of the extended test is mentioned with some frequency in the data, and as indicated in the previous paragraph, is perceived as requiring teachers to push many of their learners further than before. This is described in relation to the level of the group, as perceived by the teacher, which is another factor mentioned. Where learners were felt to be at the level of the exam, then a particular focus on the exam was not seen as necessary, rather further work on developing the learners' skills. For groups perceived as weak, teachers described focusing on the exam, closely following exam preparation material and doing intense revision, in an attempt to raise the learners' level as much as possible. The reasons given for this are most often those of the 'social conscience' type, with wanting to provide learners the best opportunities for the future the prime justification. Extreme examples make specific reference to narrowing the syllabus to focus exclusively on the exam, or even on only some aspects of the exam, as we saw in the case of Teacher 23/1 who claimed to "throw out grammar" and instead focus on reading and listening with weak learners.

Some teachers refer to extra-curricular classes, aimed at providing additional exam preparation, offered within the Teachers' Charter hours. Many of these focus on learners in class three, but there were also mentions of lessons open to everyone. Deciding to devote these extra hours to this purpose, rather than any other, is taken as another indication of the effect of the exam on planning. Many respondents also indicate a 'deficit' approach, with the teacher wanting to increase class time, in order to make up for what they perceive learners still lack, and in this way raise the level.

The final aspect mentioned in "attuning learners to the test" is the introduction of the test requirements and procedures. While some of this is described as being done in specially dedicated sections of the course book, other teachers describe the use of practice tests with this purpose in mind. Use of mock exams is also mentioned, as these offer learners experience of the whole test and include potentially problematic procedures, such as transfer of answers by the learners from their question booklet to the answer sheet, which must be done within the time allowed. This calls for learners not only to understand what they have to do, but to learn how to check their work, make any necessary corrections according to the instructions, and manage the time. There is mention of use of practice tests and mock exams at all levels in the school, but many teachers speak of a clear exam focus in class three. There appears to be increasing focus on the exam, use of practice tests and test tasks, as the exam date draws nearer, as noted by Watanabe (1996).

The final aspect mentioned of the effects of the exam on planning is in response to the exam results. Teachers report doing careful analysis of the test results, considering how to react and then planning action to improve outcomes. This includes informal, personal responses, where teachers check which items on the core curriculum were problematic for the learners

in the test and consider how they could address this in their plan for the coming year, to the production and implementation of official "corrective programmes" where the response must be documented for the directorship. The actions described in this seem to focus on identifying areas perceived to be in deficit, increasing the intensity of tasks done in, and the time spent on, this area. In other words, the typical corrective programme described narrows the syllabus and promotes teaching to the test. As some teachers point out, this is of dubious merit, particularly as the intended recipients of this programme are not those who had problems in the test, but the next cohort. We will discuss this further when considering school principals.

The question of analyzing the results of the exam appears to be motivated not by the exam itself, but by legislation on pedagogical supervision. From the perspective held by Messick (1996) this, in contrast with all the other aspects discussed in this section, is therefore not strictly washback.

6.2.2. The effects of the examination on learners

In this section we discuss the research questions concerning learners in the light of the findings reported in the previous chapter. Each question will be taken in turn.

1. Which aspects of language do learners consider important in foreign language learning in *gimnazjum?*

In the ranking obtained from questionnaires completed by 4343 learners of class one of lower secondary schools throughout Poland, we saw that speaking and vocabulary were valued as most important, followed by writing. In Table 6.7 below we compare this with findings from the Initial Washback Study of 2009.

Table 6.7. What is considered important in learning in lower secondary school according to learners: Initial Washback and Impact studies compared

	Initial washback study	Impact Study
Rank	2009 (n=229)	2012 (n=4343)
1.	speaking	speaking
2.	vocabulary	vocabulary
3.	reading	writing
4.	listening	reading
5.	grammar	listening
6.	writing	grammar

We must be aware that these are two different samples of different size and composition, with the Initial Study comprising learners from Class 3 and the Impact Study learners from Class 1 lower secondary. There is noticeable

difference in the attitude of learners towards writing, which has moved from the last place to third place in the Impact Study. The only other change is that listening has dropped one place, and is now behind reading, instead of ahead of it. The learners completing questionnaires in 2009 did so *after* the examination, which then only included one level and closed tasks. The 2012 sample completed this information towards the end of class one. While we see that there have been changes, we have no evidence from the learners themselves as to why these changes have occurred, and consequently cannot ascribe them to the exam. However, we can speculate.

If we compare how teachers in the Impact Study ranked these same aspects with the learner ranking (see Table 6.8) we note that writing differs, coming last for the teachers and in third place for the learners. The fact that there are indications of change both over time and in contrast to teacher opinion suggests that real change has taken place. If we look at the factors which could have contributed to this we see that the 2012 sample are learners who will be taking the exam in English in 2014 at the extended level, as they are following the continuation programme. (They began English in primary school. A criterion for inclusion in the BUNJO sample was that they were following this programme.) The extended level includes an open written task. We saw that teachers have declared making changes to their teaching plans to include the teaching of writing and, as they feel this takes time, many have declared that they begin teaching writing in class one. We have also noted that course books include special sections on the exam, even in class one, so we can assume that these sections may also include information about the writing tasks in the exam. It is possible, but without evidence, that it is the change in the exam (to include writing) that has brought about this change in the ranking.

Table 6.8. What is considered important in learning in lower secondary school: Learners and Teachers

	Impact Study	Impact Study
Rank	Learners 2012	Teachers 2012
1.	speaking	speaking
2.	vocabulary	vocabulary
3.	writing	listening
4.	reading	reading
5.	listening	grammar
6.	grammar	writing

However, there is also another possible reason for the change. At the time of the Initial Study 58.6% of households in Poland in 2009 had internet access. By 2012 this had increased to 70.5% of households (OECD, 2017). Online, the primary media for communication is writing and it is possible

that these Year One students may feel writing is more important for this reason. From another part of the BUNJO data (see Paczuska et al., 2014) we know that these same learners have many possibilities of such contact. 97% of them have a computer at home and 95% of them have internet access. 45% declare that they play online games several times a week in English and 23% that they use English to communicate during these games several times a week. We do not have corresponding data for the 2009 learner sample, so again we can only speculate. In sum, writing has been seen to increase in importance between 2009 and 2012 and while it is possible that the exam may be a contributing factor, we have no evidence that this is the case. This is a drawback of the data for the Impact Study being taken from a larger data set, which was not specifically designed to study the impact of the exam. In a dedicated study the learners could have been asked to explain why they felt different aspects of language were important.

2. According to learners are all four skills assessed with similar frequency?

In Chapter 5 we saw that learners felt that tests of grammar and vocabulary were the most frequent forms of assessment, followed by receiving grades for writing. There was more variety in the responses on the frequency of tests of reading and listening, with the largest number of learners giving a neutral response. We suggested that this is an effect of the way the question was formulated and that placing reading and listening together led to less clear responses. This could have been caused by the fact that, for example, one of these two aspects is less frequently assessed than the other. Speaking received more negative than positive responses, although a large number of answers were undecided. Let us now compare the learner responses with what teachers claimed.

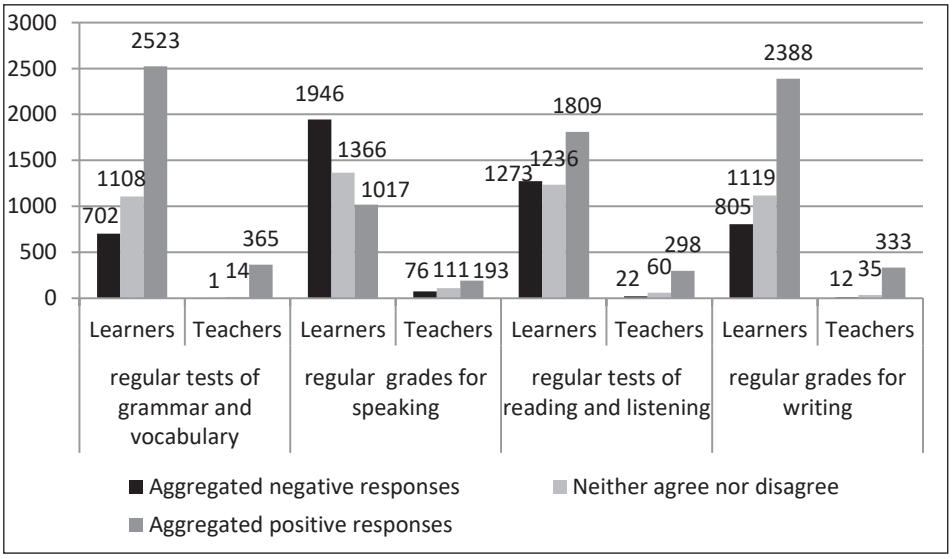
In Table 6.9 below we see that there are differences between learner and teacher perceptions of the frequency with which some aspects are assessed. The most noticeable difference is with speaking, where teachers gave more than twice as many positive (*I totally agree*, *I partly agree*) (51%) than negative responses (*I totally disagree*, *I partly disagree*) (20%), whereas learners stated the opposite, with 45% negative as opposed to 23% positive responses. The percentage of neutral responses is comparable for both groups. There appears to be a strong discrepancy. This may have been caused by a Hawthorne effect in the teacher sample, with teachers aware that, at least in theory, they should assess all aspects of the core curriculum and so may have answered as they felt was appropriate. However, we noted that data from teacher interviews appeared to corroborate the questionnaire data, so this is not necessarily the case. More likely, drawing on what was mentioned in the interview data, the difference hinges on interpretation of the word "regular" in the statement the respondents were given as a prompt (see column 2 of Table 6.9). We noted that teachers mentioned giving grades for speaking less frequently than other aspects of language and sometimes referred to this happening only once or twice a semester. Learners, perhaps

comparing "regular" with how frequently they have tests of grammar and vocabulary, or writing, concluded that speaking is not assessed as often and so gave negative responses. If this is the case, then it is a worrying trend, but we have no evidence that it is caused by the exam.

Table 6.9. Impact Study: Comparison of teachers and learners on how frequently different aspects are assessed in English lessons. Aggregated responses

	regular tests of grammar and vocabulary		regular grades for speaking		regular tests of reading and listening		regular grades for writing	
	Learners	Teachers	Learners	Teachers	Learners	Teachers	Learners	Teachers
Aggregated negative responses	702 16%	1 0.3%	1946 45%	76 20%	1273 29%	22 6%	805 19%	12 3%
neither agree nor disagree	1108 26%	14 4%	1366 31%	111 29%	1236 29%	60 16%	1119 26%	35 9%
Aggregated positive responses	2523 58%	365 96%	1017 23%	193 51%	1809 42%	298 78%	2388 55%	333 88%

The second aspect which shows difference between learners and teachers is in the distribution of responses for "tests of reading and listening". The teachers gave considerably more positive responses than any other category, leading to a positively skewed distribution (see graph 6.1 below).



Graph 6.1. Comparison of teachers and learners how frequently different aspects are assessed in English lessons: Impact Study

The learners, by contrast, gave more neutral and negative responses, although the largest number were positive. This seems to suggest that there is more differentiation in assessment than the teachers themselves admit, providing that we consider the learner responses to be reliable. In the Initial Study we noted considerable intra-group disagreement, suggesting that learners may not always be aware what is being assessed, or may remember selectively. Again, the differences may have been caused by putting reading and listening together in one question, which may have pushed more learners to choose the neutral response if they perceived that one of these aspects was assessed more frequently than the other.

3. What are learner attitudes towards the foreign language exam? What reasons do they give for this?

The largest number of mentions in the learner interviews, when asked about whether they thought the foreign language (FL) exam was important, were positive (88%), with a small number (5%) of negative mentions. If we compare these responses with those from learners in 2009 in the Initial Study (when points from the exam did not yet count towards selection for upper secondary), we note that in 2012 responses are more polarized. In 2009 learners responded on a 5-point Likert scale, where we found that 44% of learners gave positive as opposed to 23% negative responses, while 33% were neutral. It would seem that learner attitudes have become more positive since the first administration of the exam.

About three quarters of the learners (78%) in the Impact Study (2012) felt the FL paper was as important as the other papers in the exam, while about one in five (13%) felt it was less important. A minority (8%) considered that the FL paper was more important than the rest of the exam.

When asked for their opinions about the exam, the majority of mentions by learners could be interpreted as positive, with the largest number referring to the exam as objectively measuring what learners know. Others commented positively on the introduction of the FL exam, stated that it was useful and explained its importance in helping them attain a place at a good upper secondary school. By contrast, the number of negative mentions was much lower (more than three times fewer than the positive mentions). Some of these were emotional responses that, for example, the exam was stressful or worrying, while the largest number of negative responses were that the exam was difficult. A very small number (18 mentions) were against the exam and saw it as unnecessary.

To sum up, learners appear to have positive attitudes to the FL exam, viewing it as a useful, objective measure of what they know and accepting that it is as important as the other parts of the exam. These positive attitudes appear to have strengthened since 2009. The fact that attitudes are strongly positive and that learners view the exam as important may indeed be felt by teachers, some of whom we noted felt obligated to help their learners do as well as possible, because they perceived that their learners were focused

on the exam. In this way the learner responses appear to give credence to what those teachers were saying.

4. What do learners in class one *gimnazjum* know about the exam?

In Chapter 5 we reported, on the basis of interviews with 480 learners, that slightly more than half the learners (58%) from class one lower secondary claimed to know nothing about the exam. Of the remainder, most gave rather general information, most commonly that the exam is at two levels, that learners could choose which language they will take at the extended level, that it includes reading and listening, and that they need to prepare well for it. Very few learners gave more specific information, and some of what was given was incorrect. By contrast, nearly three-quarters (72%) of the learners interviewed said that they had received information about the exam in school, while one in four students (25%) said that no information had been given. This discrepancy in numbers who said information had been given and the number of learners able to say something about the exam is possibly a feature of the age of the learners (13–14), who, while being aware that something had been said about the exam, had not retained what was said. This is typical of young teenagers. What it does seem to indicate is that either these youngsters are extremely forgetful, or more likely, that their teachers have not expended a great deal of time or effort on giving examination information. If we were seeing purposeful and overt teaching to the test then we might expect learners, even in class one, to be very well-informed about the exam. We could, therefore, suggest that this is a sign that washback to the learners in class one in the form of information about the examination is most likely not evident. However, in the questionnaire where learners were asked to give opinions on the statement "We do lots of exercises during class preparing for the *gimnazjum* exam in English" we saw that around half of the respondents (56%) either totally or partly agreed, as opposed to about one in five (17%) who disagreed. This would suggest that work on exam preparation may be beginning in class one, and that a proportion of the learners are aware of this. Approximately a quarter (27%) of the learners neither agreed nor disagreed with the statement, suggesting that they are not aware if such tasks are being done or not. This appears to be consistent with what some teachers were saying about how they start "attuning" learners to the test from early in the school, while others stated that they omitted, or downplayed exam-focus material in the course book at this stage.

In short, it would appear that there is some evidence that exam tasks are being done as early as class one lower secondary school, and that approximately half of the learner sample appear to be able to recognize exam tasks and be aware they have done them in class.

To sum up, we have a little evidence to suggest the presence of washback on learners. Around half of the learners in the questionnaire sample claim to have done "lots of" exam tasks in class one of lower secondary. We see

that attitudes to the exam are strongly positive and that approximately four out of five learners (86%) in the interview sample felt the FL paper was as important, or more important, than the other parts of the national examination. Learners were noted to be aware that the exam was important in helping them to get to a good upper secondary school. We suggest that this awareness and these positive attitudes could be being felt by some teachers and point to those teachers in the interview sample who felt obligated to place the exam as an aim in their work, because their learners felt it important. While half of the learners claim not to know anything about the exam, three quarters of the sample believe information about it had been given in school. Although unable to give more than very general information when asked about the exam, it would appear that learners have become aware of its importance and significance for their futures. We will return to this in the discussion of the mechanism of washback.

6.2.3. The effects of the examination on the work of schools: Evidence from school principals

In this section we turn to a different perspective and discuss the effect of the examination at the level of the school, drawing on data from school principals. We will take each of the research questions in turn.

1. What role does the examination play in planning the work of the school?

We saw substantial evidence in the data from interviews with school principals that the exam plays a considerable role in planning the work of the school. The majority of principals explained the process of analyzing the exam results and what is then done in response to them. How the school treats the results appears to be a continuum ranging from a pragmatic and holistic whole-school response, to a prescriptive principal-led directive which strictly adheres to the Regulation on Pedagogical Supervision. We suggested in the previous chapter that school principals appear to fall into two main types, the school-manager and the principal-pedagogue, with the first type very concerned about exam results as part of a bureaucratic process. To the school-manager type results are numbers which are seen outside the school, which give information about the school's performance, the school's status, the school's rank. Results, consequently, ideally need to improve year on year, or at least remain on a similar level. Results have become anonymous statistics, remote from the educational process, or the learner. It would seem that the extreme type of school-manager promotes an ethos in the school where the exam results drive the planning process. To find evidence to support these hypotheses we looked for evidence at the level of the school, checking to see if there was response from the teachers working under such principals. In School 50, where we noted in the previous chapter that the principal appeared concerned to compare the school's results with those of other schools in the region, we found that all of the

three teachers from that school made several mentions to exam results, to a programme of mock exams and to analysis of results in their interviews. However, none of the teachers made any reference to the exam when stating what was important for them in their teaching, or in what they perceived important for the learners.

In school 116, however, given in the previous chapter as an example of the school-manager style, the two teachers interviewed showed slightly different responses to the exam. One could be described as exam-driven, in that she states preparing learners for the exam is her main aim, for the reason that "we are held to account on it." She does however, also have a second aim, which is to develop communicative skills. The exam plays a very large role, she claims, and states that "it's our priority." She describes in detail the process of using mock exams in class two and three to give information, which aims to "to get the best outcomes in the *gimnazjum* exam." Her colleague, by contrast, does not mention the exam in her aims, but states that the assessment system in school is entirely geared to the exam. Her response to the question on the role of the exam is somewhat enigmatic:

> The *gimnazjum* exam itself is quite a complicated thing isn't it? Because not all the learners have a gift for languages, so it's quite difficult that all the learners have to be prepared to the right level. Every learner has to be at the right level. The *gimnazjum* exam gives them possibility to check themselves at all levels, in all the skills and it gives them the chance for further development and it also sums up all the three years of learning in lower secondary school.
>
> School 116 Teacher 1

As we saw that the principal of her school is rather fixated on the exam results, it may be possible to sense an undercurrent in her text, which might suggest that she is experiencing some difficulty reconciling the prevailing school ethos, which is to get good results, with the learners she is teaching, some of whom, she indicates, find learning English a challenge. While the principal stated that he tries not to evaluate a teacher on the basis of the results their learners get, we could possibly detect a sense of unease in what this teacher is saying, suggesting that the fact that not all her learners might reach "the right level" gives her a cause for concern. Another signal of the centrality of the exam in the school is that both teachers refer to choosing an exam-preparation book for class three and focusing exclusively on the exam in that year. This is a small school, with only two teachers of English, which may also have a bearing on the situation. In a larger institution, with more members of staff, the influence of the school principal is more diffuse. In a small staff room it is possible that this teacher may feel under pressure, possibly exacerbated by the fact that her colleague views the exam as a driving force in her work. We may be seeing school-wide impact in the making, washing down from the school principal.

There would seem to be indications in the data, particularly when viewed at the level of school, with data from the teachers and principal viewed together, that the style of the principal, and the principal's approach to the exam, impact on the work of the school, not only at the general planning level, but also having an influence on the attitudes and planning of the teachers who work there.

If this argument is to be considered valid, then it should also apply to the other type of principal, the principal-pedagogue. In Chapter 5 we cited School 84 as an example of this type of school head. Both of the two interviews with the teachers of English in this school are characterized by very limited references to the examination, which is mentioned in response to the question specifically about it, and by one of the teachers, once, with reference to the course book. Both of the teachers have aims which are to help learners develop their speaking skills and to be able to use the language in everyday situations. They both state that they predominantly assess learner activity in the lesson. In response to the question on the role of the exam, one of the teachers claims that the exam plays a large role, as it is a criteria for their choice of books, and as they have selected an exam preparation book for class 3. However, when asked if the exam affects work throughout the school, both teachers state firmly that this only applies to class three. In classes one and two they have planned carefully, so as to be able to cover the syllabus, and treat class 3 as a time for revision and exam preparation, which is why they chose the exam-focused book. Both the teachers describe how they work with learners who are having difficulty, both in class and in remedial lessons, which they offer under the Teachers' Charter hours. One of the teachers explains how she works to interest and motivate her learners by using new technologies in the lesson, making her lessons attractive and interesting, so that learners do not feel the need to look for extra lessons outside school:

> So that they want to learn in school, because if they have that desire, that motivation, I think that learning English only in school should be enough by itself.
>
> School 84 Teacher 1

In this way we see that the teachers seem to reflect their principal's focus on the learners, with no mention at all of the exam results, or mock exams, in either of the interviews.

There would, therefore, appear to be some suggestion that the ethos created by the principal in the school also applies to the principal-pedagogue. However, when we discuss the mechanisms of washback we will see that this is only one of several factors which seem to contribute to the phenomenon. This suggestion is, of course, a subjective opinion and would need verifying through further research.

The next aspect of the exam which affects the planning of the school is the corrective action the school decides to take following analysis of the exam results. We saw that in almost all cases school principals describe the analysis of the results of the English exam which is carried out and subsequent action which is taken. What is not overtly stated is the reason why this is so. In Chapter 2 we gave information about legislation on quality control in education in Poland, in the form of the Regulation on pedagogical supervision [pol. *w sprawie nadzoru pedagogicznego*] (MEN, 2009), which at the time of this study required all schools to analyze the results of external examinations. We saw that principals responded differently to conclusions from the analysis of the examination results. Some adhered closely to the legislation and required teachers to draw up corrective programmes, which in many cases comprised allocating additional time to areas of the syllabus found to be wanting, doing more exercises on these, testing these aspects more frequently, or spending more time in remedial classes on them. Such action could be interpreted as directing teachers to narrow the syllabus (as additional time has to be found at the cost of another aspect of the course, unless additional hours are allocated for this purpose), or to teach to the test, as the additional exercises or tests done are to be on the basis of those on which the learners in the previous year performed poorly. Some teachers themselves drew attention to the fact that cramming learners and doing practice tasks does not lead to development of the required area. This is more effectively done, they suggest, by focusing on general development of the language and providing plentiful opportunities for practice of language in use.

Many principals reported more pragmatic responses to the analyzed results, on the basis that as each cohort of learners is different, strictly applying a corrective programme based on findings from the previous year's results is inappropriate. Action taken on the external results following this interpretation ranged from reflection and discussion in the subject team to try to find possible reasons for the low performance and the drawing up of plans for preventative action for the future, to detailed contextual analysis which investigated factors in individual learners which could have contributed to the difficulty. In addition, many schools undertook mock exams late in Year 2, the results of which were used to diagnose both exam-readiness of learners and to plan any necessary remedial action. Although this too might be considered teaching to the test, at least it has pedagogical foundation, in that it aims to develop the skills of the test-takers on the basis of areas in which they themselves performed less well. If this diagnosis is approached more generally, in the form of achievement testing relating to areas of the core curriculum, rather than a mock exam, then it cannot be cause for objection. Some principals also pointed out that the way to help learners develop is through offering a rich programme of extra-curricular

activities, which encourage the use of the language and report that this type of action has had positive effects.

To conclude, there is evidence that the actions of some principals, in response to analysis of the results of the external exam, appear to be encouraging narrowing of the syllabus, and over-emphasis on practising test items, in a mistaken belief that this will lead to improvements in achievement. This in turn appears to be in response to their rather literal interpretation of the existing legislation on pedagogical supervision. Other school principals, by contrast, were found to be encouraging co-operation between teachers in subject teams, to work together to discuss, find causes of difficulty and plan together how to work to improve them. This echoes Hattie (2015b), who found that intra-school variability in achievement measured on PISA tests is much greater than the variability between schools and that the largest factor which contributes to this variance is the work of the teacher.

> my claim is that the greatest influence on student progression in learning is having highly expert, inspired and passionate teachers and school leaders working together to maximise the effect of their teaching on all students in their care. There is a major role for school leaders: to harness the expertise in their schools and to lead successful transformations. There is also a role for the system: to provide the support, time and resources for this to happen. Putting all three of these (teachers, leaders, system) together gets at the heart of collaborative expertise.
>
> (Hattie, 2015b: 2)

In encouraging teachers to work together to discover why some groups perform better on the exam than others and find solutions, principals appear to be doing just this.

To sum up, when considering the role the exam plays in planning the work of the school we have seen a variety of outcomes. At a systemic level we have seen solutions offered in terms of allocation of time, resources and extra-curricular classes. We have also noted response to the difficulties experienced by individual learners, or to a challenging school district. Where the school principal was seen to be engaged and working together with the teachers to find a way forward, which we would suggest is the response taken by those we described as the principal-pedagogue, we have the impression that exactly what Hattie is calling for is already taking place.

2. What role does the examination play in evaluating the quality, or effectiveness of English teaching in lower secondary school?

Effectiveness, in the Polish context of educational quality control, is defined as the educational outcomes of a school described in relative terms (Gocławska, 2013: 44). This is understood as the results of the school

in external examinations seen in comparison with the wider population, as measured on a standard nine, or centile scale, or using a value-added measure (ibid.). These concepts were presented in Chapter Two.

In the interview data we saw that many school principals compared their school's results with those of others. In the light of the understanding of effectiveness in the Polish context, this could be considered appropriate. Other schools describe using two measures, at the start of school and the end, to determine whether there had been growth. Still others reported annual measures and comparison of the outcomes on these tests to see if growth was taking place.

An alternative approach noted was to use a more criterion-referenced comparison, as opposed to a norm-referenced comparison which focuses on scores in a population. Here the school takes each standard of the core curriculum and evaluates the performance of the learners on items that measure each standard. This is then used to give information about how well learners are able to, for example, read "to find specific information in a text", or "give suggestions". These schools tended to describe this as diagnosis, or qualitative analysis, and explained the findings as being used by teachers in drawing up schemes of work, or setting goals for teaching.

Other schools made it clear that the analysis of external exam results is only one of a series of measures, which we described as a complex school-based assessment scheme, valuing the external exam as an objective, standardized measure, but in some cases criticizing the amount of information it gave the school for planning purposes. More information, claimed many principals, is obtained from formative assessment done during the three years of school, than from a one-off summative measure at the end of the cycle. To illustrate this, one principal suggested that the exam gives information which should feed forward to the upper secondary school about the abilities of candidates, or the new intake, and should not be used to feed backward to the lower secondary school.

To sum up, schools are expected to measure effectiveness of learning outcomes and the results of the external exam are being used as indicators. This is problematic. The Regulation on Pedagogical Supervision (2009) lists 17 areas in which the quality of the work of a school should be evaluated, only one of which are the results of external exams. The over-emphasis on the results seems to be coming from how "effectiveness" is being interpreted by the inspectorate who carry out external evaluation of schools. It is also reflected in supervision of the work of the inspectorate conducted by the Supreme Audit Office [pol. *NIK*]. In a published report of school supervision carried out in 2011, they stated that there was evidence the evaluation process was ineffectual, citing that

> one quarter of the primary and lower secondary schools that were subject to control attained worse results in the [external examinations] during the period

under investigation… Results of the scores at the end of gimnazjum were worse in one third of the schools than in the previous year.

(NIK, 2011: 7)

This appears to indicate that representatives of the Supreme Audit Office had fallen into the trap of comparing not only raw scores, but of raw scores year on year, which is inappropriate, on the grounds that one exam is not comparable with another, a fact which is clearly stated by the Central Exam Board. Effectiveness should not be measured only on the grounds of the results of a summative examination. Several of the principals themselves called for use of value-added for foreign languages or greater focus on the school's internal formative assessment process, rather than depending on the external examination alone. In chapter 2 we noted that the Regulation (2009) has since been revised (2013), has a less draconian approach to the examination results and places more emphasis on the use of a variety of assessments. It is to be hoped that this is now sufficiently flexible to prevent action that leads to narrowing of the syllabus, or teaching strictly to the test being undertaken. Much greater emphasis can now be placed on internal assessment programmes and formative assessment, which can be seen as a very positive move.

3. Is there evidence of changes in the organization of English in lower secondary school which can be attributed to the exam or the exam results?

Some school principals indicated that changes are made in the allocation of hours, organization of extra-curricular classes, and the allocation of resources for foreign language teaching in response to exam results. Where exam results were felt to be less than satisfactory, we saw that one reaction was to allocate the discretionary hours at the principal's disposal to increase the number of contact hours for English. Other principals allocated Teachers' Charter hours for additional exam preparation, or for remedial classes when specific areas of difficulty were noted. We also noted schools obtaining other funds (such as EU grants), which were used to increase the number of hours, both as part of regular teaching and as extra-curricular lessons. One principal claimed that such a programme had helped to increase the mean score of the exams in his school from the mean score of the voivodeship to a mean score above that level.

There was some evidence, although coming from a small number of schools, that principals try to fund new resources for English in the form of dedicated language classrooms, language labs, or interactive whiteboards. Other responses were decisions to change the teaching programme, or course books, with the aim of trying to improve learning outcomes, in response to exam results. A very small number of principals mentioned encouraging teachers to set up European projects to give learners more opportunities

to use English, increase their understanding of its importance and boost their motivation. They justified this by explaining that as learners use the language more they improve and so the exam results will also increase.

In this section we have discussed the research questions for the Impact Study relating to Teachers, Learners, and School Principals. We now move to an attempt to put what has been learnt into a discussion of how washback at the classroom level, and impact at the level of the school and beyond come into effect and work. For this we draw on information from Chapter Five Theme Four, where we discussed what we described as levers which seem to exert forces on teachers.

6.3. Mechanisms of washback and impact

In this section we aim to apply the findings from the research to attempt to explain our understanding of how washback and impact work in the Polish context. Making a differentiation between the micro- and macro-levels is artificial, as we believe that both comprise one complex, dynamic system. However, as there are a large number of factors at work, it seemed to be easier to explain each separately. We will begin with the classroom micro-level and look at washback.

6.3.1. Mechanisms at work in the micro-level of classroom practice

We have seen that teachers respond to the examination in different ways, as predicted by Alderson & Hamp-Lyons (1996), with some clearly experiencing the process of washback, while others remain apparently unaffected. We now need to consider *why* this is the case and what evidence we have found to support our ideas. Key ideas in the text are highlighted in italics.

One of the levers described by teachers, which they appeared to feel pushes them to focus on the exam is *a sense of being held to account by the system*. Some teachers reported feeling that they themselves were being evaluated on the basis of their learners' results. We found little evidence in the interviews with school principals that this in fact takes place, with much of the analysis of results being done at the level of the school, rather than at the level of the group. Where it was reported that teachers did individual analysis, this was most often qualitative criterion-referenced analysis to obtain information about which areas of the curriculum had been problematic for learners in the exam. While we cannot completely exclude the possibility that some teachers might feel personally pressurized, as we saw one example of a teacher who had witnessed criticism of a teacher of another subject on the basis of exam results at a whole-school staff meeting, it seems unlikely that this is the major factor which causes some teachers to feel they are being called to account for the results. However, if there is a school where

teachers are aware that results will be discussed in the forum of the whole school staff, *they may well seek to avoid any potential unpleasantness by taking action to ensure their learners do as well as possible*. We suggested that the need to analyze the results and find reasons for difficulties appeared to be seen by some teachers as threatening, possibly because they do not understand why they have to do this. *Where the teacher could explain the chain of accountability*, the pressure the school principal is under to report to their supervisors in the local council and the LEA, *the teacher appeared able to rationalize why test results were viewed as important, and accept the process.* Where the teacher simply reported that they were evaluated on the basis of the results, it may be the case that they do not understand the accountability process. One teacher who had accepted that analysis was required, even reported finding the process stimulating and appeared to see it as an opportunity for self-development. Thus, it is not the fact that there is an educational policy which requires the school to analyze and reflect on the exam results *alone* which causes washback to come about, but *how* the teacher *understands* the policy and *interprets* it. If it is not comprehended, then it can bring about negative washback, yet if the teacher understands and interprets it in a way which is consistent with their personal beliefs, then it may even have positive consequences and lead to improved teaching.

It may be that the difference between why some teachers feel held to account and others do not lies in *how closely the teachers identify with the learners and feel responsible for their achievement*. We noted the instance of a teacher who felt that she was taking the examination with her learners. This appears to represent very strong investment of the teacher in the learning process. The teacher's belief was that the results are "a witness" to her work. Thus if the results are poor, she would most likely view this as a personal failure. She appears to have become so identified with the process that her whole personality is engaged. In such circumstances it is possible that such a teacher may even experience a sense of threat at the prospect that the results could be less than satisfactory (the teacher reported feeling enormous stress) and so take evasive action to avoid what she could envisage as a hurtful event. Avoidance of negative outcomes, in the form of less than satisfactory results, could lead her to focus strongly on the test and bring about a washback effect. This appears to be an extreme version of accountability, with the teacher causing herself the stress of feeling responsible for the learners' results, because of her beliefs.

We noted other examples of teachers feeling responsibility towards their learners and so focusing on the exam, but these we suggested are caused by what we described as *the teacher having a 'social conscience'*. Teachers reported wanting to give their learners the best opportunities for the future, and as the exam results are used for selection to upper secondary schools, good results were perceived as giving better chances. Other teachers, working in areas they perceived as having social-economic difficulties, or

with learners coming from less advantaged families, wanted their learners to do well, to give them the possibility to even out their opportunities, thus seeing *an exam focus as providing greater equity*. Here again the sense of accountability is being generated by the teachers themselves, in a sense for the greater good. The washback here seemed to some extent mitigated as many teachers also had another aim in their teaching which was to help their learners develop communicative skills, in keeping with the curriculum aims.

In sum, a sense of needing learners to get good results appears to come from two main levers: one external to the teacher, caused by the system and educational policy, and the other internal to the teacher, rising either from their relationship with their learners, or their sense of social responsibility towards them. In the case of the external lever, however, we suggest that the effect may be reduced or transformed, depending on the teacher's understanding of the processes at work and how they interpret these for themselves.

The next factor which can contribute to washback appears to be *how the teacher understands the relation between the exam and the curriculum.* We saw that where the teacher reported that the exam did not test speaking, as they failed to recognize the 'communication' tasks as an indirect test of this construct, then there were instances where the teacher narrowed the syllabus to either reduce or omit the teaching of speaking.

Table 6.10. Factors connected to teachers found to lead to washback

Factors found to lead to washback	Migitating factors
External: sense of being held to account by the system	If accountability process is understood, some teachers appear able to accept and work with it positively
Internal: degree of identification with learners/relationship with learners	Degree of personal distance, pragmatism
Internal: social conscience	Teacher has another teaching aim in their work other than preparing learners for exam
Internal/contextual: good results lead to greater equity for disadvantaged learners	Teacher has another teaching aim in their work other than preparing learners for exam
Understanding of relationship between exam and curriculum	If teacher understands test construct well, no negative effects
Perception of level of group in relation to level of exam	Group perceived as below level of exam led to narrowing of syllabus and/or teaching to the test

Another factor at the classroom level apparently causing a washback effect was *how the teacher perceived the level of the group in relation to the level of the exam*. Where the teacher reported that the learners were at the

exam level or above it, there was no negative effect, but where the teacher reported perceiving the learners as below the level of the exam this was seen to lead either to narrowing of the syllabus, or to teaching to the test.

In this study we found no direct evidence of negative washback to the learners from teachers, but we would like to propose a process which might account for the fact that learner perception of the importance of the exam was seen to increase between the Phase II Initial Washback Study in 2009 and the Phase III Impact Study in 2012. While it is possible that the difference is caused entirely by the fact that the exam became high-stakes in 2012, of interest seems to be the fact that, although only half the learners in class one could give some information about the exam, and three-quarters of the learners claimed that some information in school had been given, yet the majority of learners stated that the exam was important. We saw in both teacher and principal reports that earlier some learners had been considered as not taking the exam seriously, which we suggested was most likely a pragmatic response to its results having no significance. We saw many references to action, taken by both teachers and the principal, to change learner attitudes and persuade them of the importance of the exam, in some cases reported as effective. Is it possible that this a reason why some teachers felt obligated to focus on the exam because of their learners? Is this a vicious circle? The mechanism would appear to work like this:

> Teacher, wanting to motivate learners, tells learners the exam is coming and is important for them (iterative process)> Learners believe what teacher says> Learners start to have positive attitudes to the exam, understand its importance for them> Learners want to get good scores because they perceive test as important> Learners imply to teacher that they want to get good scores, as the exam is important> Teacher feels obligated to include exam preparation as an aim, justifying this by saying that their learners think the exam is important.
>
> Ergo: Teacher initiates the process of washback by (repeatedly) telling learners about the importance of the exam.

We have no clear evidence that this is the case, and it would need to be investigated further. Cheng et al. (2015: 446) suggest something similar: "Students' perceptions of tests are likely to be shaped by the school context, for example by their teacher and their peers."

Teachers also reported that *learner attitudes towards the exam seemed affected by their language level and their plans for the future*. We saw some evidence from teachers in the Phase II Initial Washback Study that *less able learners were positively motivated by the exam to engage more in class and try harder*. We did not have a corresponding question for learners and so are unable to corroborate this. Teachers reported that learners who knew what they wanted to do in future, and realized that this involved getting to a good upper secondary school, were motivated by the exam to work hard. For some other learners, according to the teachers, the exam had become

a goal in itself and this also acted as a motive. A proportion of the learners, reported as being less academic in general, seemed indifferent to the exam. Teachers and principals report a sense that learners have become more motivated since the exam became high-stakes.

This closes the section on factors at work at the classroom level. We now turn to the wider context.

6.3.2. Mechanisms at work in the macro-level of the school and beyond

The main factor at work at the school level in the Polish context seen to contribute to the exam having an impact is educational policy, specifically the requirement arising from the Regulation on Pedagogical Supervision (2009).

We saw that teachers reported having to analyze the results of the exam, with the aim of identifying weak areas, to find possible reasons for these, and make recommendations how to improve the situation. These are then presented to the principal and to the whole school staff. Corrective action is agreed upon, recorded in the proceedings of the meeting and subsequently implemented. Teachers are monitored to ensure that the implementations are followed through. The whole process is documented by the school principal in their annual report to the LEA. We noted that extreme examples of this appeared to lead to narrowing of the syllabus and teaching to the test. School past performance on exam was shown to have an impact on the rigour with which this is followed, apparently influenced by the relationship of the school with the LEA and inspectorate.

We also saw less rigorous versions of the process of analysis of exam results, with discussion in the subject team, consultations with the principal, reflection from the teachers and planning decisions made in response. There was a little evidence of the principal believing teachers would regulate each other, through comparison of the outcomes of their learners on the exam. This was outweighed by reports of encouraging teachers to collaborate, particularly in the foreign language subject team, which could be seen as an unintended positive consequence.

School principals responded in different ways to the requirements of the Regulation with a large number appearing to place more confidence in complex school-based assessment programmes to provide information for diagnostic and planning purposes than in the external exam. Others adhered rigorously to the legislation and this appeared more likely to promote negative consequences.

We suggest that a difference in styles of leadership could contribute to impact, with a more bureaucratic school manager style appearing to have a different perspective on the exam results than the principal-pedagogue. We noted that where the results are viewed as figures, removed from the learner or the context, there appears to be more susceptibility to promote negative

impact in a drive to raise scores and improve the school's ranking. By contrast, the principal-pedagogue appears concerned to increase the outcomes and promote development of each individual learner and particularly those who are disadvantaged. Outcomes for this type of principal seem more likely to be measured in terms of growth, than as exam results. Leadership styles could also impact on school policy and mission, which in turn could focus attention on successful exam results, rather than promoting individual progress.

We suggested tentatively that the leadership style of the principal and their perception of the importance of the exam results impacts on the ethos of the school and on the attitudes, aims and planning of the teachers.

Evidence was noted of the principal's perceived importance of exam results impacting on allocation of hours, introduction of ability-streaming, planning of extra-curricular lessons and activities and allocation of resources. Low scores were also seen to lead to similar actions, some reported to have had good effect.

Table 6.11. Factors found to lead to test impact in schools

Factors seen to cause impact	Impact observed
Educational policy: corrective programmes	Narrowing of syllabus, teaching to test
School past performance on exam	Below average> narrowing of syllabus, teaching to test Above average> good advertisement in community > positive impact on student recruitment
Principal's response to educational policy requirement	Rigorous adherence> negative Adapted response> complex assessment programme> growth scores, plus summative scores, reduced negative impact
Principal's leadership style	School manager> focus on results> negative consequences Principal pedagogue> more focus on individual learners, diminished impact of exam
Principal's attitude to teachers	Promotes collaboration> positive outcomes Promotes competition> potential negative consequence, teaching to test
Principal's perception of importance of exam results	Allocation of hours, extra-curricular plans, allocation of resources in attempt to raise scores Effect on teacher attitudes, aims and planning
Socio-economic factors in school	Increased exam-support offered
Situation of school	Increased exam-support offered

Socio-economic factors in the school, in the number of disadvantaged learners, could lead to exam-focus. Perceptions of socio-economic difficulties

in the school district were seen to lead the principal to offer exam-preparation extra-curricular classes and remedial teaching.

The situation of the school was also seen to play a role in impact, with rural settings at a distance from centres of conurbation where the school is the only one in a community, seen to offer more exam-focus support.

The ranking of the school on the basis of its exam results was perceived as being a good advertisement for the school in the community. This in turn was perceived as having a positive impact on recruitment, viewed as particularly desirable where schools are competing for decreasing numbers of students.

This concludes the section on the perceived impact of the exam at the level of the school and beyond.

6.3.3. Mechanisms as a dynamic system

We suggested in chapter one that washback and impact form aspects of one complex dynamic system where many factors interact together. The Impact Study conducted during this research focuses on an exam a critical point in its evolution, the moment when it becomes a high-stakes exam with consequences for the learners. What we appear to be seeing, however, which could well be an artefact of the research instruments and design, is that the greatest impact is being caused not by the exam itself, but rather by a piece of legislation which requires the school to analyze the exam results and act on the findings. When this study was designed this was not foreseen. In Chapter 2 we saw that it is the uses to which test results are put, rather than the tests themselves, which have the greatest potential impact. Our findings seem to provide a good example of unintended consequences resulting from the uses of results of a national examination. We also saw there that the *intended* consequence of the section of Regulation on Pedagogical supervision on use of exam results, was to improve effectiveness of education, which appeared to have been narrowly defined as being measured in terms of results on the national exam. We noted that this policy when followed rigorously was seen to lead to narrowing of the syllabus and teaching to the test, both classic responses to raise scores in response to pressures of accountability. Teachers and principals in the data set voiced their disquiet at the negative consequences they perceived this was bringing about. We saw alternative responses being taken by principals to mitigate these and it would seem that in the changes subsequently made to the legislation their voices have been heard. One small section in a long document about quality control in school had the power to trigger a chain of events with far-reaching effects on schools, teachers and learners. This is a dynamic system in action. The difficulty is that the change in legislation will not erase the memory of the practices it brought about, or the habits it caused to be learnt. Such events leave traces. It is to be hoped that the response in the form of teaching to

the test and narrowing of focus can be unlearnt and changed. We saw one teacher describing how her attitudes to how to prepare her learners for the exam had changed over time, as she came to a greater awareness of what *not* to do. This gives hope that teachers are able to move forward.

We have suggested that factors in leadership style of the school principal appear to impact on teacher attitudes and planning as regards the exam. We have also seen that both teachers and principals appear to respond to the learners' home background and to the socio-economic character of the school district, in terms of planning both at the classroom and school level.

We saw little to support that it is the exam alone which causes negative impacts, rather it was the understanding teachers had of its construct and the relationship of that construct to the core curriculum which potentially have negative consequences. There were also suggestions that some teachers were not aware of how to prepare learners for the tasks indirectly testing speaking, indicating that it was a lack of teacher skills, rather than the test itself that was contributing to washback (Cheng et al., 2015). Conversely, of course we can argue that if speaking were tested directly this difficulty would not arise and therefore it is the fact that speaking is tested indirectly that leads to negative consequences.

There was a great deal of evidence that washback from the exam was being contributed to by course books (as in Cheng, 2005), but again we saw clear examples that the potential for this may be mitigated by how teachers approach the book and plan their work. This appeared to be related to the teacher's degree of autonomy, with those who asserted that the book did not decide for them apparently less exam-focused in their work. Teachers have a voice in the choice of books and we saw that analysis of the exam requirements featured as a criteria in this. We found that teacher beliefs, attitudes and perceptions contributed to washback or mitigated against it as in other studies (e.g. Turner, 1996; Burrows, 2004).

In short, if the researcher restricts their investigation only to the level of the classroom in studying the consequences of a new exam they are missing important factors which influence planning and decision-making. Seeing the school as part of a complex dynamic system provides a richer picture and helps us understand far more about the mechanisms of washback and impact and how they come about.

This concludes discussion of the data collected during this longitudinal study of the impact of a language examination on the work of teachers and schools.

6.4. Implications of the research

In this section we suggest implications for different stakeholder groups based on the findings from this study.

6.4.1. Implications for teachers

As a point for reflection we offer this quotation from a school principal who quite eloquently expresses a key issue for teachers arising from this study:

> It's very important for me that learners also had language skills, because unfortunately the exam system is in such a form at the moment, that in my opinion it drives teaching to the test and for me that means that the student doesn't necessarily need to know the language to be able to pass the *gimnazjum* exam well. And here I believe the most important is for the learner to be able to use the language communicatively…we have to be clear what we expect because, of course, the core curriculum has certain standards which have to be met, and the aims of the programme are really appropriate, but isn't it the case that the form of the external exams, doesn't it result in us forgetting a little about the things that are key in language teaching, are we not going in the direction of teaching to the test and forgetting about the teaching of the whole, that the learners are able to use it, because we know that today the most important skill is communication.
>
> Principal School 119

For the teacher to work effectively and avoid negative consequences of the external exam they need to fully understand the core curriculum, its aims and requirements. Next there is a need to analyze the construct of the external exam, that is what it is testing, and understand how this is done. This is particularly important where part of the construct is being tested indirectly, as in the case of speaking in the *gimnazjum* exam. Once the test construct has been internalized, then the next action is to consider how to plan teaching so as to meet the aims of the curriculum, with the perspective that at the end of the cycle of teaching learners will be tested externally. The curriculum takes precedence, as in the context of the Polish school covering it is a legal requirement. The fact that speaking is tested indirectly does not mean that speaking should not be taught. On the contrary, the most effective way to prepare learners for communicative tasks which are written is for them to engage actively in spoken use of the language, as reported by many of the teachers interviewed in the Impact Study.

It is important for the teacher to have a rational approach towards the exam results, to leave emotion aside, and depend on thorough course planning, appropriate choice of teaching materials and a well-thought out, coherent formative assessment plan. It is also necessary to consider carefully when and how learners should be introduced to the exam and the approach to take to it. While the course book is helpful in this process it should be approached maturely, as a tool to support the teaching process, rather than passively allowing it to determine what is taught.

Teachers need good formative assessment practices and particularly how to deal with the assessment of speaking, which was seen as an area of

concern in this research. Speaking appeared to be less frequently assessed both in teacher and learner data. As becoming able to communicate in everyday situations is the main curriculum aim, then this should be being assessed. As part of assessment the teacher also needs skills to analyze test results, which we have seen involve use of descriptive statistics and item analysis. Also needed is the ability to analyze what a test item is testing in terms of the curriculum, which also implies the ability to analyze the aims of course book tasks, or those from other sources.

The teacher needs to be aware that it is the school principal who leads the process relating to the external exam. Some sensitivity is needed to ascertain their attitudes to the exam and exam results, and to be aware of possible pressures from the context to which the principal may be subject. A teacher in a school does not function in isolation, but as a member of a community which functions according to regulations and procedures. Being open and trying to understand the beliefs and attitudes of others towards the exam, its results, how to prepare learners for it and how to deal with pressures, real and perceived, will lead to more effective work. It may mean having to compromise and find acceptable alternatives. It should be expected that this will be a continuous learning process.

6.4.2. Implications for school principals

We found that the school principal plays a key role in determining the attitude of the school towards the external exam. It would appear that there is a need to remember that exam results are subject to a large number of variables both in the individual learner and in the context. Mean scores of schools on external exams flatten these differences and leave behind a figure which, in terms of diagnostic information, is of no practical purpose. The information obtained from one-off summative assessment, like the external exam, provides only a picture of performance on a measure on a given day. While using the external results for criterion-referenced analysis, or descriptive statistics at the level of the group, or analysis of the achievement of individuals may be informative, it should be compared with other data obtained from formative assessment and teachers and not treated in isolation. This analytic process has rich potential for staff development and the growth of "collaborative expertise" (Hattie, 2015). It should not, however, be used to create competition between members of staff. For improving the effectiveness of education, understood as the growth of the language proficiency of each learner, a well-planned programme of both teaching and formative assessment is needed. Teachers need support and encouragement in this. Full understanding of both the curriculum and the examination are needed, which are topics for whole school staff development sessions. There is a powerful temptation to teach to the test, but as this data show, this is not the best path to take. Being aware of this as a possible tendency and helping

teachers work to find alternative approaches is a subject for whole-school staff discussion. Part of this should include how to inform learners and their parents, or carers, about the school's approach to the examination and its role in the teaching, learning process. By taking such a holistic approach it is to be hoped that pressures on the school staff can be mitigated and a positive learning environment created. As Poland enters another period of educational reform and new external examinations at the end of the basic school in 2019, experience learnt from the last few years is relevant and valuable.

6.4.3. Implications for the learner and their family/home context

The aim of language learning in lower secondary school in the context we have investigated is to develop the ability to be able communicate in English. This means being able to use the language in situations both in school and in everyday life. These aims are fully compatible with the aims of the external examination, although the way in which these skills are tested may at first sight be surprising. The exam includes a section which tests communicative skills through recognizing how to ask or respond in a given context. To do these tasks the learner needs to imagine themselves in the situation and, through, for example, conducting an imaginary role play, choose the best option. Learning how to do this comes from real practice of language in use, speaking and responding in pairs and groups during the lesson. In the same way developing the ability to write well in English comes through practising writing in real situations.

Communication, whether in speech or in writing, is about making oneself understood to another or others, which requires having sufficient control over the language resource so that the message is clear. For this the learner needs control over aspects of grammar, functions of language and vocabulary. These are the tools which are used to create effective communication. They should not become the aim of learning by themselves, as they need to be placed in a context and used. Rote learning of lists of words, or the ability to complete gaps in a sentence with appropriate grammatical forms, is not enough to be able to use language in context to convey information.

Development of language skills is best supported by extensive access to the language in both written and spoken form, not only in school, but also outside it. This can be supported at home by encouraging the youngster to read in English, both online and print materials, and to listen and watch films and TV in the original language. This means allowing them access to appropriate web sites, or helping the young person find books and magazines to read. Most useful at this level are simplified materials for language learning. Extensive access to English and use of the language are more supportive of general language development than having the child do countless practice tests. Practice tests teach the learner the tricks of how

to do exam tasks, but will not necessarily help raise their language level or improve their exam scores.

6.4.4. Implications for local education authorities and the inspectorate

Results from external examinations should not be used as a sole measure of the quality of teaching and learning in a school. They must be seen in context and viewed at the level of classes and individuals, rather than as mean scores for the school. They need to be compared with internal measures taken by the school, teachers' reports and observations on learner progress. External exam results are not comparable year on year, as exams differ in content and difficulty and each cohort of learners has different strengths and weaknesses, which come from a host of contextual factors. Evaluating a school by comparing the results of one year's exams with the next is therefore inappropriate. Over-emphasis on demanding that schools raise exam scores has been seen in this research to lead to negative consequences, pushing the school to practices which mean that the whole of the core curriculum is not being covered ("narrowing the syllabus"), or that teachers spend lesson time preparing learners to do test tasks, rather than working on the development of language skills and communicative skills in particular, in other words not achieving the curricular aims. Any practice which leads to a school being discouraged from fulfilling its statuary obligations should be changed. Publication of ranking of schools in terms of external examination results should be accompanied by clear explanation of what such tables do *not* show, how such tables are intended to be responsibly used and what contextual information for parents is more informative in helping them and their children select schools. This, for example, would include the school's policy statement, teaching programme with information about specialized classes on offer, information about extra-curricular activities and individual support available for learners.

6.4.5. Implications for language policy makers

This research has shown the sensitivity of the situation in schools to the effects of language policy, and illustrated how easily policy may lead to unintended negative consequences, resulting from the impact of use of the results of an external examination. It is to be hoped that the review of literature offered in Chapter Two may inform educational policy makers of what has been found in other contexts and lead to reflection on how best to plan for more positive consequences.

As Poland begins a reform of curriculum and examinations it is to be hoped that the information presented here leads to mature reflection on the role of external examinations in schools. We have seen that the situation is complex, with many factors interacting with each other. Apportioning

blame on any of the parties involved would not be helpful in improving the situation, and nor would holding the examination itself responsible for the subsequent outcomes. We have seen that there is the need for all those involved, the principal, teachers, learners and their parents or carers, to have a clear and profound understanding of the intentions of both the curriculum, the examination, and the relationship between the two. Failure to fully comprehend this leads to negative consequences. This would suggest that information to all those stakeholders is key in the process of avoiding the negative impact of a new examination. Information needs to be designed for each target audience in ways which are understandable for them. Broad spectrum information is unlikely to be effective. Making information available in multimedia formats (such as short films and filmed presentations with visual materials) as well as print may make it more accessible. The provision of such information needs appropriate funding and a sufficient number of staff to prepare it.

We have seen that teachers are strongly influenced by the course books they use and so would suggest that there is a need to involve the educational publishers actively in the process of disseminating information about the curricular reform and the new examination, working together with them to ensure information is accurate and appropriate. It may be the case that the expertise of these publishers could be harnessed to help make the form and message content of the new test syllabus [*Informator*] more accessible for teachers, as it would appear that a proportion of teachers have difficulty interpreting the information in the existing document. Again, the addition of multimedia information to print formats may make it more accessible, particularly to younger, or less experienced, teachers. On the basis of the Phase II Initial Washback Study we would also recommend making sample test papers available at the same time as publication of the test syllabus to reduce speculation among teachers, and learners.

In terms of the content of the English examination, we have seen that there appear to be some issues with face validity in the case of tasks testing communication which need to be addressed. In terms of the test construct it would also appear that there is a case for arguing for extended cover of communication, as this was seen to be somewhat under-represented in terms of the proportion of the number of tasks and allocation of points in comparison with tasks testing reading. This could perhaps also be addressed by extending the number of communicative tasks that are tested through listening. There would appear to be a need to include, at least, some semi-open tasks based on the testing of productive use of language functions in context.

We would recommend that if it continues to be the case that the examinations are not comparable year on year, as the tradition of publishing full analysis of the test outcomes with commentary would seem to suggest, then very clear messages need to be made to all stakeholders, and most

particularly to the LEAs, the inspectorate and the Supreme Audit Office, that exam results should not be compared year on year to measure effectiveness. While this information is included in the current *Informat*or it is not prominent.

6.4.6. Implications for universities and other institutions providing teacher education and development

We have seen in this study that teachers need strong assessment skills which include the ability to analyze test results using descriptive statistics and item analysis. It is to be recommended that this be included in BA teacher education programmes. In the interests of staff development within these institutions we would suggest that a similar practice of analyzing the results of internal examinations in practical language become standard, with discussion of outcomes and contextual factors offering material for improvement of teaching programmes and teacher development. University staff could therefore develop assessment literacy, collaborative expertise and deeper understanding of the challenges classroom teachers face in school.

This research has illustrated that broadening the context of the study to a macro-perspective substantially alters the information which is obtained, helping to clarify where the sources of various behaviours may lie. Focusing too narrowly could have led to incorrect assumptions. Pedagogical research in school contexts is of key importance and needs to be encouraged. Of particular benefit would be inter-disciplinary studies which allow researchers from different fields such as educational psychology and sociology to work together with applied linguists. This research also points to the potential benefit of a large-scale study in providing new insights into phenomena. Such studies are beyond the scope of the individual researcher and emphasize the need for co-operation between institutions to create projects.

6.5. Suggestions for further research

This section will be divided into two sections, first suggestions arising from the findings of this study aimed at any researcher in interested in washback and/or impact and second suggestions particularly aimed at researchers in Poland.

We have seen difficulties in this research with the logistics of organizing a study of the washback from and impact of a national examination. In order to capture the complexity of the phenomena we have seen that it is necessary to explore the context thoroughly from the perspective of a range of stakeholders. In terms of design there would therefore appear to be two options. First, to create a multi-disciplinary team, as suggested above, and aim at a larger sample, or second, to conduct a series of smaller-

scale contextualized studies of schools, perhaps benefitting from the type of design employed by Wall & Horák (2011). The sample would ideally be stratified to include schools in different centres of population, both rural and urban, as well of schools of different sizes. If a sample representative for the country is not possible for reasons of scale, then it might be worth considering a sample representative for an administrative region. Such case studies could allow the washback and impact mechanisms to be fully investigated in context.

Also relevant for design is an important point about choice of instruments. It was originally intended to compare information obtained from teachers and learners in the Impact Study with data from observation of lessons, but the scope of this work had to be reduced to make it feasible. In a smaller scale study, observation would be advised at all phases, and not just in the baseline, as was the case here. We have seen that confronting data from teacher questionnaires and interviews with their classroom practice seen firsthand reveals discrepancies.

The next area of consideration are the choice of stakeholders a study should include and the forms of instruments to use to obtain data. This research drew mainly on surveys and interviews with teachers, learners, and school principals. Stakeholders which were identified as important, but who were not included, were the LEA and the inspectorate. It would appear that including representatives from these authorities may provide important perspectives in a study of impact. Parents were not included here either and future studies could consider extending the scope to examine the impact of the home environment on the formation of learner attitudes towards the exam in question. Another group not directly investigated, but found to be influential in the role of disseminating and interpreting information about the new exam, were educational publishers. This would also seem to be a fruitful line of enquiry.

Future studies would be advised to engage learners more fully in the process than in this research. It may be that small groups of focal learners with an experienced teacher of this age group as facilitator could provide richer data than one-to-one interviews. It may also be the case that giving open questions in an online environment to be responded to in writing offer a workable alternative. On a cautionary note, we have seen that asking learners to reflect on learning which has taken place over a whole school year produced information which showed a lack of intra-group reliability. For young teenagers a year was clearly too long a period for focus. Particularly confused was the data from learners on classroom assessment and this would require an alternative approach in a future study.

The study itself has created several areas worth further consideration. Of particular interest for investigation appears to be the suggested role played by the leadership style of the school principal on teacher attitudes towards, and perception of, the examination. This is a difficult area for research

and would require strong ethical standards to be maintained. Contact with teachers online rather than face to face in school, where they may feel uncomfortable and inhibited, could be a useful approach to data collection.

As the course book featured strongly in teacher evidence about washback in the classroom, analysis of the treatment of the examination in course books could provide useful insights, especially if supplemented by discussions with the editor, author(s) or publisher. A study focused exclusively on how teachers approach exam preparation in planning their scheme of work in different year groups, or with learners of different levels of ability, could also be very interesting, as these were found to be factors contributing to washback. Discovering more about teacher decision-making processes in planning could inform us more about the washback mechanism. Teachers reported including aspects of the examination in their schemes of work, and assessment plans, both of which are another potentially useful source of data.

Another area of interest may be to investigate the sources of teachers' attitudes towards the exam in question, which might entail researching official information given by the exam provider (as in Wall & Horák, 2011), information disseminated by publishers, social media and public media. This would appear to be interesting in the months before the first administration of a new exam and at the time when the first results are published. It would be worth comparing how different teachers respond, particularly those new in the profession with those of, say, ten or more years' experience.

This study found different responses to the exam in teachers according to the level of the learners they were teaching and depending on how the teacher perceived those learners' ability in relation to the exam. It would be interesting to see if this is borne out in other contexts.

In sum, while this research has contributed to our knowledge of washback mechanisms, there is still much left to explore.

For researchers in the Polish context there are many opportunities created by the educational reform commencing in the school year 2017/2018. The return to a two-tier primary/secondary system eliminates the lower secondary school, which was the focus of this research. In 2019 a new examination, the Year 8 exam is planned. Such radical change, taking place with relatively little forewarning, is a very fruitful topic for study. A good starting point would be careful analysis of all new documents such as the new core curriculum (2017), the new test syllabus (2017), new course books and other information available on the new examination. It would seem that in-depth rich-context studies of schools would be needed to capture the highly complex nature of this reform, as inevitably the changes in the exam will be entangled with responses to the changes in the school system.

Teachers of the new classes 7 and 8 in the reformed basic school in the interim period of school years 2017/2018 and 2018/2019 face the enormous

challenge of preparing their learners for a new exam at level A2(+) in two years. Under the previous system these learners would have had three years to reach this level (*gimnazjum* classes 1–3). How teachers choose to deal with this very specific situation in the face of a new examination is of special interest, particularly in the case of their response to groups where learners find English a challenge.

In the first two phases of this project it was noted that teachers were concerned with learners having a limited lexical resource and so it would be of interest to investigate vocabulary size among learners of later classes of primary school, perhaps related to learning outcomes as measured by the new examination, to see if there is empirical support for this. Classroom assessment of speaking was found to raise questions in Phase III the Impact Study and seems to be a good area for further study. The ways schools analyze the results of external examinations and how the team of foreign language teachers respond to them is another worthwhile focus. Within teacher education, what and how assessment and testing is taught and the applicability of this knowledge and skills in the school context is an area where findings could have useful and practical application.

It should be noted that the educational reform envisages not only a new Year 8 exam, but also changes to the *matura.* At the time of writing, plans for reform of higher education propose that universities may introduce entrance tests if they feel the need, which, if it comes into effect, may also have the potential for washback.

In short, the current situation offers very many possibilities for research and a great deal remains to be discovered.

6.6. Closing remarks

The report of this research has been long in the making. Since the collection of data for the Impact Study, education has moved on, policies have changed, and the three-year lower secondary school, which was the subject of this study, will shortly become a thing of the past. With it the external examination in English phases out, to be replaced by a new examination, the Year 8 exam, in 2019. As a result, unexpectedly, this book becomes a historical record of a period in Polish foreign language education. However, it is very much to be hoped that has been found out in the process of this longitudinal research, lasting from 2008 to 2012, will serve to inform the educational reform which is just beginning.

There appear to be many lessons to be learnt, as the implications given here suggest, for all those who are touched by national mandatory examinations. It would seem that it is not necessary for a school to fall into the trap of teaching to the test, if certain guidelines are followed. Yet as the school is not a hermetic entity isolated from other influences, guidelines are also needed

in other institutions, those responsible for examinations, those responsible for assurance of educational quality, those which educate teachers, and those which decide policy affecting schools. We have seen that how individuals respond to an external examination is subject to a large number of factors, both inside themselves in the form of their beliefs, attitudes and perceptions, but also outside themselves in response to relationships in school, to learners and teachers, and to the wider context. The school as a community has been seen to be strongly affected by educational policy, and a part of this has been adverse.

The process of the impact of an external examination is not a simple linear one of cause and effect, nor does it happen all at once, but in stages, differentially according to the parties involved, over time. It appears sensitive to a large number of factors. Revising an examination will not automatically change anything and it should not be expected to. We have seen that teachers, learners and their families need to have a thorough understanding of the demands of an exam, and its relation to the curriculum, and to work together to find rational ways to achieve curricular targets, which will then be tested. Schools are, above all, places that provide opportunities for learning and individual development. An external examination is simply one way, among many others, of gathering information about what has been done. It is time to recognize this, emphasize the prime importance of the process of learning, and promote a mature and rational approach to examinations.

Streszczenie

Egzaminy odgrywają ważną rolę w systemie edukacji. Badania przeprowadzone w różnych kontekstach na całym świecie pokazują, że egzaminy doniosłe (ang. *high-stakes*) mogą mieć wpływ na różne aspekty nauczania i uczenia się. Dokonano rozróżnienia pomiędzy pojęciem wpływu (ang. *impact*), który służy do opisu ogólnych efektów egzaminu w szerszym kontekście dotyczącym szkoły lub systemu edukacyjnego, oraz pojęciem efektu zwrotnego egzaminu (ang. *washback*), zastosowanego do opisu efektów egzaminu przekładających się na nauczanie i uczenie się. Jednakże badania pokazują, że nie wszyscy uczniowie i nie wszyscy nauczyciele podlegają wpływowi egzaminów doniosłych, lub że podlegają im w różnym stopniu. Wskazuje się, że stopień tego wpływu jest związany z postawami i poglądami samych uczniów i nauczycieli. Aktualnie niewiele wiadomo na temat mechanizmów wyjaśniających dlaczego niektóre osoby w pewnych warunkach mogą podlegać wpływowi, podczas gdy inne im nie podlegają.

W polskim systemie edukacji, zewnętrzne egzaminy kończące etap edukacyjny w gimnazjum rozpoczęły się w 2002 roku i miały na celu kierowanie i przyjmowanie młodych uczniów do właściwej szkoły ponadgimnazjalnej. W roku 2009 po raz pierwszy języki obce stały się częścią egzaminu. Ta zmiana postawiła nowe wyzwania nauczycielom, uczniom i szkołom oraz stała się źródłem interesujących pytań badawczych. Czy nowy egzamin będzie miał wpływ na nauczanie i uczenie się języków obcych w gimnazjum? Jeśli tak, jaką przyjmie formę?

Niniejsza książka opisuje badanie podłużne przeprowadzone w gimnazjach, które rozpoczęło się od opisu stanu wejściowego, przedstawiającego sytuację w szkołach przed wprowadzeniem egzaminu z języka obcego. Badanie kontynuowano przez okres pięciu lat, zbierając dane po pierwszym egzaminie, kiedy wyniki nie były jeszcze brane pod uwagę przy rekrutacji do szkoły ponadgimnazjalnej, oraz ponownie po trzech latach kiedy egzamin po raz pierwszy został wykorzystany w celach rekrutacyjnych.

Książka składa się z sześciu rozdziałów, z których trzy stanowią tło teoretyczne badania. W pierwszym rozdziale zawarto definicje pojęcia efektu zwrotnego egzaminu (ang. *washback*) oraz zaprezentowano modele pokazujące jak ewoluowało rozumienie tego pojęcia. Następnie dokonano przeglądu literatury opisującej efekt zwrotny egzaminu w odniesieniu zarówno do przeprowadzonych badań dotyczących egzaminów międzynarodowych, jak i państwowych egzaminów doniosłych w kontekstach szkolnych. Kolejna część rozdziału definiuje pojęcie wpływu (ang. *impact*), biorąc pod uwagę różne określenia tego pojęcia oraz pokazuje modele proponowane przez różnych autorów. Rozdział kończy przegląd badań na temat wpływu egzaminów.

Rozdział drugi skupia się na sposobach użycia testów. Bada związek pomiędzy wpływem egzaminu a jego trafnością, przedstawiając jak jest on postrzegany przez różnych badaczy. W części drugiej rozważany jest problem siły sprawczej testów i ich wykorzystania do dokonywania selekcji w społeczeństwie. To prowadzi do dyskusji na temat etyki w testowaniu i refleksji dotyczącej odpowiedzialności za rezultaty egzaminu. Sekcja ta zamyka się ogólnym zarysem kwestii dotyczących sprawiedliwości i słuszności testów. Kolejna część przygląda się związkowi pomiędzy wynikami testów a pomiarem jakości pracy szkół, gdzie przytoczono dwa przykłady dotyczące testów oraz odpowiedzialności i rozliczalności (ang. *accountability*) – pierwszy ze Stanów Zjednoczonych, a drugi z Anglii. W ostatniej części tego rozdziału opisana jest rola, jaką odgrywają egzaminy zewnętrzne w polskim systemie edukacji.

Rozdział trzeci śledzi ewolucję państwowego egzaminu zewnętrznego z języka obcego na zakończenie gimnazjum w Polsce, umieszczając go w kontekście historycznym. Przedstawiono dokładną analizę egzaminu, jego uzasadnienie i formę oraz związek z podstawą programową i treściami testu. Zaprezentowano wyniki egzaminów osiągnięte za pierwsze trzy lata, a następnie zawarto informacje o powodach modyfikacji egzaminu w 2012 roku oraz opis dokonanych zmian.

Druga część książki bada wpływ zewnętrznego egzaminu z języka angielskiego przeprowadzanego na zakończenie gimnazjum. Rozdział czwarty wyjaśnia motywy podjęcia się niniejszego badania. Opisuje typ projektu badawczego oraz omawia możliwe trudności związane z używaniem danych jakościowych. Następnie prezentuje trzy fazy badania wraz z informacjami o zastosowanych instrumentach badawczych, wybranej próbie badawczej oraz procedurach.

Rozdział piąty omawia dane uzyskane na trzech etapach badania: badania stanu wejściowego, badania wstępnego efektu zwrotnego egzaminu oraz badania wpływ egzaminu. Wyszczególniono również ograniczenia w badaniu.

Rozdział szósty omawia wyniki w świetle pytań badawczych. Rozważane są efekty egzaminu w odniesieniu do nauczycieli i uczniów (efekt zwrotny egzaminu) oraz pracy szkół (wpływ). Opisano działające mechanizmy, na poziomie mikro – w klasie – oraz na poziomie makro – w szkole i poza nią. Wypracowane sugestie adresowane są do różnych interesariuszy podlegających wpływowi egzaminów oraz zaproponowano kierunki dalszych prac badawczych.

Zaobserwowano, podobnie jak i w innych badaniach, że egzamin ma wpływ na wybór materiałów nauczania. Pojawił się jednak nowy aspekt – wydaje się, że nauczyciele podejmując decyzje dotyczące intensywności skupienia się na egzaminie uzależniają je od poziomu zaawansowania swoich uczniów.

Na stosunek nauczycieli do egzaminu ma wpływ szereg czynników, w tym samo postrzeganie egzaminu, czynniki kontekstowe i społeczne oraz stosunek do podręcznika.

Zauważono, że uczniowie mają pozytywny stosunek do egzaminu z języka obcego, traktując go jako przydatne i obiektywne narzędzie pomiaru ich wiedzy, oraz równie ważny element jak pozostałe komponenty egzaminu gimnazjalnego. Znaleziono niewiele dowodów na istnienie efektu zwrotnego egzaminu na uczniów.

Są istotne dowody wskazujące na istnienie wpływu egzaminu na pracę szkół. Wyniki egzaminów wywierają wpływ na przydział godzin nauczania, decyzje dotyczące wyposażenia szkoły oraz prowadzenie zajęć pozalekcyjnych. W odpowiedzi na wyniki egzaminów, dyrektorzy wskazywali na opracowywanie ogólnoszkolnej koncepcji dotyczącej nauczania języka angielskiego. Na te decyzje ogromny wpływ ma Rozporządzenie MEN w sprawie nadzoru pedagogicznego. Wykazano, że dyrektorzy w bardzo różny sposób reagują na rezultaty egzaminów, co jak się przypuszcza ma związek ze stylem zarządzania szkołą.

References

AERA, APA & NCME, (1999). *The Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Albee, A. (1999). Assessing impact: some current and key issues. Caledonia Centre for Social Development, Scotland. Retrieved from http://www.caledonia.org.uk/pia.htm.

Alderson, J. C. (2004). Foreword. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in Language Testing: Research Contexts and Methods* (pp. ix-xii). Mahwah, NJ: Lawrence Erlbaum Associates.

Alderson, J. C. & Wall, D. (1989). *The Sri Lanka O Level Evaluation Project. First interim report.* Lancaster: University of Lancaster.

Alderson, J. C. & Wall, D. (1990). *The Sri Lanka O Level Evaluation Project. Second interim report* Lancaster: University of Lancaster.

Alderson, J. C. & Wall, D. (1991). *The Sri Lanka O Level Evaluation Project. Third interim report* Lancaster: University of Lancaster.

Alderson, J. C. & Wall, D. (1992). *The Sri Lanka O Level Evaluation Project. Fourth and final report*. Lancaster: University of Lancaster.

Alderson, J. C. & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*, 115–129

Alderson, J. C., Clapham, C. & Wall, D. (1996). *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press.

Alderson, J. C., & Banerjee, J. (2001). Impact and washback research in language testing. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley (Eds.), *Experimenting with Uncertainty: Essays in Honor of Alan Davies* (pp. 150–161). Cambridge, UK: Cambridge University Press.

Alderson, J. C. & Hamp-Lyons. L. (1996). TOEFL preparation courses: A study of washback. *Language Testing, 13,* 280–297.

Alderson, J. C. & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*, 115–129.

Allen, P., Cummins, J., Mougeon, R. & Swain, M. (1983). *Development of Bilingual Proficiency: Second Year Report.* Toronto, Ontario: The Ontario Institute for Studies in Education.

Allwright, R. & Bailey, K. (1991). *Focus on the Language Classroom*. Cambridge: Cambridge University Press.

Amrein, A. L. & Berliner, D. C. (2002). High stakes testing uncertainty and student learning. *Educational Policy Analysis Archives, 10*(18). Retrieved from http://epaa.asu.edu/epaa/v10n18

Amrein, A. L. & Berliner, D. C. (2002c). An analysis of some unintended and negative consequences of high-stakes testing. Tempe, AZ: Education Policy Studies Laboratory, Arizona State University. Retrieved from http://www.asu.edu/educ/epsl/EPRU/documents/EPSL-0211–125-EPRU.pdf.

Anastasi, A. (1968). *Psychological Testing*. New York: MacMillan.

Anastasi, A. & Cordova, F. A. (1953). Some effects of bilingualism upon the intelligence test performance of Puerto Rican children in New York City. *Journal of Educational Psychology, 44*, 1–19.

Anderson, J., Muir, W., Bateson, D, Blackmore, D. & Rogers, W. (1990). *The Impact of Provincial Examinations on Education in British Columbia: General Report*. Victoria: British Columbia Ministry of Education.

Andrews, S. (1994a). The washback effect of examinations-Its impact upon curriculum innovation in English language teaching. *Curriculum Forum 4*, 44–58.

Andrews, S. (1994b). Washback or washout? The relationship between examination reform and curriculum innovation. In D. Nunan, R. Berry & V. Berry (Eds.). *Bringing about Change in Language Education: Proceedings of the International Language in Education Conference 1994* (pp. 67–81). Hong Kong: University of Hong Kong.

Andrews, S. (2004). Washback and Curriculum Innovation. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in Language Testing: Research Context and Methods* (pp. 37–52). Mahwah, NJ: Lawrence Erlbaum Associates.

Andrews, S., Fullilove, J. & Wong, Y. (2002). Targeting washback- a case-study. *System 30*, 207–223.

Antaki, C., Billig, M., Edwards, D. & Potter, J. (2007). Discourse analysis means doing analysis: A critique of six analytic shortcomings. In J. Potter (Ed.). *Discourse and Psychology: Volume 1 Theory and Method* (pp. 331–347). London: Sage.

Atkinson, J. W. (1964). *An Introduction to Motivation*. Princeton, N. J.: Van Nostrand.

Atkinson, J. W., & Feather, N. T. (1966). *A Theory of Achievement Motivation*. New York: Wiley.

Au, W. (2007). High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis. *Educational Researcher, Vol. 36*, No. 5, 258–267.

Austin, J. (1962). *How to Do Things with Words*. Oxford: Oxford University Press.

Bachman, L. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. (2004). *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press.

Bachman, L & Palmer, A. (1996). *Language Testing in Practice.* Oxford: Oxford University Press.

Bachman, L & Palmer, A. (2010). *Language Assessment in Practice.* Oxford: Oxford University.

Bachman, L. & Purpura, J. (2008). Language assessments: Gatekeepers or door openers? In B. Spolsky & F.Hult, (Eds.) *The Handbook of Educational Linguistics*. (pp. 456–468). Oxford: Blackwell.

Bailey, K. (1996). Working for washback: a review of the washback concept in language testing, *Language Testing 13*, 257–279.

Bailey, K. (1999). *Washback in Language Testing.* TOEFL Monograph Series MS 15, Princeton, NJ: ETS.

Baker, E. L. & Linn, R. L. (2004). Validity issues for accountability systems. In S. Fuhrman & R. Elmore (Eds.). *Redesigning Accountability* (pp. 47–72), New York: Teachers College Press.
Bakhtin, M. (1981). *The Dialogic Imagination*. Austin, TX: University of Texas Press.
Banerjee, J. 2000. Using English language screening tests in your institution. Paper presented at the BALEAP Professional Issues Meeting Nottingham Trent University, 20 May.
Banerjee, J. & Tsagari, D. (Eds.) (2016). *Handbook of Second Language Assessment.* Berlin, Boston: de Gruyter Mouton.
Bartmańska, A. (2007). Wpływ egzaminów zewnętrznych na kształtowanie postaw moralnych uczniów, nauczycieli i egzaminatorów. *Edukacja* 1, 22–30.
Battle, E. (1965). Motivational determinants of academic task persistence. *Journal of Personality and Social Psychology, 2*, 209–218.
BBC Trending.(2016). ‚Traditionally submissive Muslim women' say who us? Retrieved from http://www.bbc.com/news/blogs-trending–35403106.
Berwick, R. and Ross, S. (1989). Motivation after matriculation: Are Japanese learners of English still alive after exam hell? *Japan Association for Language Teaching Journal 11*, 193–210.
Binet, A. (1905). New methods for the diagnosis of the intellectual level of subnormals. *L'Année Psychologique*, *12*, 191–244. Translated by E. Kite, (1916) in *The development of intelligence in children*. Vineland, NJ: Publications of the Training School at Vineland.
Bishop, J. H. (1998). The effect of curriculum-based external exit exam systems on student achievement. *Journal of Economic Education, 29*, 171–182.
Bishop, J. H. (2000). Curriculum-based external exit exam systems: Do students learn more? How? *Psychology, Public Policy, and Law, 6*, 199–215.
Bishop, J. H. (2001a). A steeper, better road to graduation. *Education Next,* 1(4), 56–61.
Bishop, J. H. (2001b). *Why do students learn more when achievement is examined externally?* Retrieved from http://media.hoover.org/documents/ednext20014unabridged_bishop.pdf.
Black, P. & Wiliam, D. (1998). 'Inside the black box'. Raising Standards Through Classroom Assessment. *Phi Delta Kappan 80* (2), 139–144.
Blake, J. E. (2012). High-Stakes Testing: A (Mis)construed, Normalizing Gaze. *International* Journal of Educational Policies, 6, 5–23.
Blake, J. (2008). *The Consequential Effects of High-Stakes Testing On Teacher Pedagogy, Practice and Identity: Teacher Voices Disrupt The A Priori*. Unpublished Dissertation, University of Tennessee.
Block, D. (2000). Problematizing interview data: Voices in the mind's machine? *TESOL Quarterly 34*, 757–763.
Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas Accountability System. *American Educational Research Journal, 42*, 231–268.

Booth, D. (2012). *Exploring the Washback of the TOEIC in South Korea. A Sociocultural Perspective on Student Test Activity*. Unpublished PhD thesis. University of Auckland, New Zealand.
Bourdieu, P. (1991). *Language and Symbolic Power*. Cambridge: Polity Press.
Bourdieu, P. & Passeron, J. (1976). *Reproduction in Education, Society and Culture*. London: Sage.
Brickman, W. W. (1946). Preparation for the Regents' Examination. *School and Society*, *64*, 263.
Brim, O. G., Jr., Glass, D. C., Neulinger, J., & Firestone, I. J. (1969). *American Beliefs and Attitudes about Intelligence*. New York: Russell Sage Foundation.
Broadfoot, P. (1996). *Assessment Schools and Society*. London: Methuen.
Brookover, W. B., Shailer, T., & Paterson, A. (1964). Self-concept of ability and school achievement. *Sociology of Education, 37*, 271–378.
Brunfaut, T. (2014). A lifetime of language testing: An interview with J. Charles Alderson. *Language Assessment Quarterly*, 11, 103–119.
Butler, J. (1990). *Gender trouble: Feminism and the Subversion of Identity*. New York and London: Routledge.
Butler, J. (1993). *Bodies that Matter: On the Discursive Limits of Sex*. New York and London: Routledge.
Burrows, C. (2004). Washback in classroom-based assessment: A study of the washback effect in the Australian Adult Migrant English Program. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in Language Testing: Research Contexts and Methods* (pp. 113–128). Mahwah, NJ: Lawrence Erlbaum Associates.
Byrnes, H. (2005) Perspectives, No Child Left Behind, *Modern Language Journal,* 89, 248–282.
Cambridge English. (2016) *Principles of Good Practice. Research and Innovation in Language Learning and Assessment.* Cambridge: Cambridge English Language Assessment. Retrieved from http://www.cambridgeenglish.org/images/22695-principles-of-good-practice.pdf.
Canale, M. (1983). From communicative competence to language pedagogy. In J. C. Richards& R. W. Schmidt (Eds.) *Language and Communication*. (pp. 2–27), London: Longman.
Canale, M. & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics 1*, 1–18.
Cannell, J. J. (1988). Nationally normed elementary achievement testing in America's public schools: how all fifty states are above the national average. *Educational Measurement: Issues and Practice,* 7(2), 5–9.
Cannell, J. J. (1989). *The "Lake Wobegon" Report: How Public Educators Cheat on Standardized Achievement Tests.* Albuquerque, NM: Friends for Education.
Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? Across-state analysis. *Educational Evaluation and Policy Analysis, 24*, 305–331.
Centralna Komisja Egzaminacyjna. (2007). *Informator o Egzaminie Gimnazjalnym z Języka Angielskiego Przeprowadzanym od Roku Szkolnego 2008/2009*. Warszawa: CKE.

Centralna Komisja Egzaminacyjna. (2009). *Osiągnięcia Uczniów Kończących Gimnazjum w Roku 2009. Sprawozdanie z Egzaminu Gimnazjalnego 2009.* Retrieved from https://www.cke.edu.pl/images/stories/Wyniki_09/raport_gimnazjum_2009.pdf.

Centralna Komisja Egzaminacyjna. (2010). *Osiągnięcia Uczniów Kończących Gimnazjum w Roku 2009. Sprawozdanie z Egzaminu Gimnazjalnego 2010.* Retrieved from https://www.cke.edu.pl/images/stories/001_Gimnazjum/spr_gimn_2010.pdf.

Centralna Komisja Egzaminacyjna. (2011). *Osiągnięcia Uczniów Kończących Gimnazjum w Roku 2009. Sprawozdanie z Egzaminu Gimnazjalnego 2011.* Retrieved from https://www.cke.edu.pl/images/stories/0001_Gimnazja_2011/spr_gim.pdf.

Chalhoub-Deville M. (2009). The intersection of test impact, validation, and educational reform policy. *Annual Review of Applied Linguistics, 29*, 118–131.

Chalhoub-Deville, M. and Deville, C. (2005). A look back at and forward to what language testers measure, In E. Hinkel (ed.), *Handbook of Research in Second Language Teaching and Learning.* Mahwah, NJ: Erlbaum.

Chalhoub-Deville, M. & Deville, C. (2006). Old, borrowed and new thoughts in second language testing. In R. L. Brennan (Ed.) *Educational Measurement* (4th ed.) (pp. 517–530), Westport, CT: American Council on Education & Praeger Publishers.

Chapman, D. W., & Snyder, C. W. (2000). Can high-stakes national testing improve instruction: Re-examining conventional wisdom. *International Journal of Educational Development, 20,* 457–474.

Chappell, P., Bodis, A. and Jackson, H. (2015). *Impact of Teacher Cognition and Classroom Practices on IELTS Test Preparation Courses in the Australian ELICOS Sector*. IELTS Research Reports Online Series; 2201–2982; 2015/6. Retrieved from http://search.informit.com.au/documentSummary;dn=880020729503598;res=IELHSS

Chapelle, C. A., Enright, M. K. & Jamieson, J. M. (Eds.) (2008). *Building a Validity Argument for the Test of English as a Foreign Language™*. Mahwah, NJ: Lawrence Erlbaum.

Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong. *Language Education, 11*(1), 38–54.

Cheng, L. (1998). Impact of a public English examination change on students' perceptions and attitudes toward their English learning. *Studies in Educational Evaluation, 24*(3), 279–301.

Cheng, L. (1999). Changing Assessment: Washback on Teacher Perspectives and Actions. *Teaching and Teacher Education, 15*(3), 253–271.

Cheng, L. (2004). The washback effect of a public examination change on teachers' perceptions toward their classroom teaching. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in Language Testing: Research Contexts and Methods*. Mahwah, NJ: Lawrence Erlbaum Associates, 146–170.

Cheng, L. (2005). *Changing Language Teaching through Language Testing: A Washback Study.* Cambridge: Cambridge University Press.

Cheng, L. (2008). The key to success: English language testing in China. *Language Testing 25*, 15–37

Cheng, L., & Curtis, A. (2004). Washback or Backwash: A Review of the Impact of Testing on Teaching and Learning. In L. Cheng, Y. Watanabe & A. Curtis (Eds.), *Washback in Language Testing: Research Contexts and Methods* (pp. 3–17). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Cheng, L., S. Andrews & Y. Yu (2011). Impact and consequences of school-based assessment in Hong Kong: Views from students and their parents. *Language Testing* 28, 221–250.

Cheng, L., Sun, Y., Ma, J. (2015). Review of washback research literature within Kane'sargument-based validation framework. *Language Teaching 48*, 436–470.

Cheng, L., Watanabe, Y., & Curtis, A. (Eds.). (2004). *Washback in language testing: Research contexts and methods*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cicourel, A. V. (1964). *Method and Measurement in Society*. New York: Free Press.

Cizek, G. J. (2001), More Unintended Consequences of High-Stakes Testing. *Educational Measurement: Issues and Practice. 20*, 19–27.

Clotfelter, C. T., & Ladd, H. F. (1996). Recognizing and rewarding success in public schools. In H. F. Ladd (Ed.), *Holding schools Accountable: Performance-based Reform in Education* (pp. 23–63). Washington, DC: The Brookings Institution.

Cook, D. L. (1955). An investigation of three aspects of free-response and choice-type at the college level. *Dissertation Abstracts 15*, 1351.

Coughlan, S. (2013). Pisa tests: UK stagnates as Shanghai tops league table. Retrieved from http://www.bbc.com/news/education–25187997 on 3.12.2013

Coulter, C. (2012). Aer Lingus apology over language tests for passenger. *Irish Times* 12 March 2012.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Council of Europe. (2009). *Relating language examinations to the Common European Framework of References for language : Learning, Teaching, Assessment (CEFR). A Manual.* Strasbourg: Language Policy Division.

Crandall, V. J. (1963). Achievement. In H. W. Stevenson (Ed.), *Sixty-second Yearbook of the National Society for the Study of Education, Part I, Child Psychology.* Chicago: University of Chicago Press.

Cronbach, L. J. (1971)Test validation. In R. L. Thorndike (Ed.). *Educational Measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L. J. (1984) *Essentials of Psychological Testing*. Fourth Edition. New York: Harper and Row.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L.J. (1989).Construct validation after thirty years. In, R.E Linn (Ed.), *Intelligence: Measurement, Theory, and Public Policy*. Urbana, IL: University of Illinois Press.
Cronbach, L.J. & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin 54* (4), 281–302.
Crooks, T.J. 1988. The impact of classroom evaluation practice on students. *Review of Educational Research 58* (4), 438–481.
Curtis, F. D. (1944). Testing as a means of improving instruction. *Science Education, 28*, 29–31.
Darling-Hammond, L. (2007). Race, inequality and educational accountability: The irony of "No Child Left Behind." *Race, Ethnicity, and Education, 10*, 245–260.
Davies, A. (Ed.). (1968). *Language Testing Symposium: A psycholinguistic approach*. Oxford: Oxford University Press.
Davies, A., (1997). Demands of being professional in language testing. *Language Testing 14*, 328–339.
Deville, C. & Chalhoub-Deville, M. (2011). Accountability-assessment under No Child Left Behind: Agenda, practice, and future. *Language Testing, 28*, 307–321.
Docherty, C., Gratacos Casacuberta, G., Rodriguez Pazos, G. and Canosa, P. (2014). Investigating the impact of assessment in a single-sex educational setting in Spain. *Cambridge ESOL : Research Notes* 58, 3–14.
Dolata, R. & Sitek, M. (Eds.) (2015). *Raport o Stanie Edukacji 2014. Egzaminy Zewnętrzne w Polityce i Praktyce Edukacyjnej.* Warszawa: Instytut Badań Edukacyjnych.
Dörnyei, Z. (2005) *The Psychology of the Language Learner*. Mahwah, NJ: Lawrence Erlbaum Associates.
Ebel, R. L. (1965) *Measuring Educational Achievement*. Englewood Cliffs, N. J.: Prentice-Hall.
Eckstein, M. & Noah, H. (1993). *Secondary School Examinations: International Perspectives on Policies and Practice.* New Haven, CT: Yale University Press.
Educational Testing Service (2002). *ETS Standards for Quality and Fairness*. Princeton, NJ: ETS.
Educational Testing Service (2003). *ETS Fairness Review Guidelines*. Princeton, NJ: ETS.
EALTA. (2016). *Guidelines for Good Practice in Language Testing and Assessment*. Retrieved from http://www.ealta.eu.org/guidelines.htm.
El-Ebyary, K. (2009). Deconstructing the complexity of washback in relation to formative assessment in Egypt. *Cambridge ESOL Research Notes. Vol. 35*, 2–5.
Ellis, M. (2009a). Próba opisania warsztatu pracy nauczyciela języka angielskiego w szkołach gimnazjalnych. In M. Pawlak, M. Derenowski & B. Wolski (Eds.) *Problemy Współczesnej Dydaktyki Języków Obcych* (pp. 321–329). Poznań-Kalisz. PWSZ Konin. UAM Poznań.

Ellis, M. (2009b). Wpływ egzaminu gimnazjalnego z języka angielskiego na proces nauczania i uczenia się. *Neofilolog 33*, 185–196

Ellis, M. (2010). Podstawa programowa i podręcznik: najbardziej typowe wyzwania w pierwszej klasie gimnazjum. In K. Droździał-Szelest, (Ed.). *Materiały Edukacyjne w Nauczaniu Języków Obcych- Teoria i Praktyka* (pp. 9–16). Gorzów Wielkopolski: Państwowa Wyższa Szkoła Zawodowa w Gorzowie Wielkopolskim.

Ellis, M. (2013). Beliefs and attitudes about language learning and teaching among learners in lower secondary school. *Neofilolog 40*/2, 183–203.

Ellis, M. (2014). English lessons in lower secondary schools in Poland: Reflections based on learner voices. *Neofilolog 42*/1, 25–38.

Ellis, S., Hawkey, R. & Docherty, C. (2014). Iterative impact studies of Progetto Lingue 2000. *Cambridge English Research Notes. Vol. 58,* 15–20.

Eurich, A. C. (1931). Four types of examinations compared and evaluated. *Journal of Educational Psychology, 26,* 268–278.

European Commission/EACEA/Eurydice. (2015). *Assuring Quality in Education: Policies and Approaches to School Evaluation In Europe. Eurydice Report.* Luxembourg: Publications Office of the European Union.

Feather, N. T. (1963). Persistence at a difficult task with alternative task of intermediate difficulty. *Journal of Abnormal and Social Psychology, 66*, 604–609.

Feather, N. T. (1966). Effects of prior success and failure on expectations of success and subsequent performance. *Journal of Personality and Social Psychology, 3*, 287–298.

Feldt, L. S., & Brennan, R. L.(1989). Reliability. In R. L. Linn (Ed.). *Educational Measurement* (3rd ed.). New York: American Council on Education and Macmillan.

Ferman, I. (2004). The washback of an EFL national oral matriculation test to teaching and learning. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in Language Testing: Research Contexts and Methods* (pp. 191–210). Mahwah, NJ: Lawrence Erlbaum Associates.

Figlio, D & Loeb, S. (2011). School Accountability. In E. Hanuszek, S. Machin & L. Woessman (Eds.). *Handbooks in Economics*, Vol. 3, (pp. 383–421). The Netherlands: North-Holland.

Fish, J. (1988). *Responses to Mandated Standardised Testing.* Unpublished PhD Thesis. University of California, Los Angeles.

Flock, A. J. M., & Saggar, U. (1968). Academic performance with, and without, knowledge of scores on tests of intelligence, aptitude and personality. *Journal of Educational Psychology*, 59, 395–401.

Foucault, M. (1995). *Discipline and Punish: the Birth of the Prison,* (2nd Vintage ed.) A. Sheridan (trans.). New York, NY: Vintage.

Freeman, D. (1996). 'To take them at their word'. Language data in the study of teachers' knowledge. *Harvard Educational Review, 66*, 732–761.

Frederiksen, J.K. & Collins, A. (1989). A system approach to educational testing. *Educational Researcher*, *18* (6), 27–32.

Frohlich, M. Spada, N. & Allen, P. (1985). Differences in the communicative orientation of language classrooms. *TESOL Quarterly 19*, 27–56
Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization, *Language Assessment Quarterly,* 1, 253–66.
Fulcher, G. (2009). Test use and political philosophy. *Annual Review of Applied Linguistics* 29, 3–20.
Fulcher, G. & Davidson, F. (2007). *Language testing and assessment. An Advanced Resource Book.* Abingdon: Routledge.
Fulcher, G. & Davidson, F. (2008). Tests in Life and Learning: A deathly dialogue. *Educational Philosophy and Theory 40*, 407–417.
Fullan, M. (2001). *The New Meaning of Educational Change* (3rd ed.). London: Cassell.
Fullan, M. & Park, P. (1981). *Curriculum Implementation: a Resource Booklet.* Toronto, Ontario: Ontario Ministry of Education.
Fullan, M. with Stiegelbauer, S. (1991). *The New Meaning of Educational Change* (2nd ed.) London: Cassell.
Gee, J. (1995). *Social Linguistics and Literacies: Ideology in Discourses*. (2nd ed.). London: Falmer Press.
Gipps, C. (1999). Socio-Cultural Aspects of Assessment. *Review of Research in Education,* 24, 355–392.
Glaser, B. G. (1978). *Theoretical Sensitivity: Advances in the Methodology of Grounded Theory*. Mill Valley, CA: Sociology Press.
Glidewell, J. C., & Stringer, L. A. (1967). *Early Detection of Emotional Illness in School* St. Louis, Mo.: County Health Department, Division of Research and Development.
Glover, P. (2006). *Examination Influence on How Teachers Teach: A Study of Teacher Talk.* Unpublished PhD thesis, University of Lancaster.
Glover, P. (2014). Do language examinations influence how teachers teach? *International Online Journal of Education and Teaching. 1*(3). 197–214. Retrieved from http://iojet.org/index.php/IOJET/article/view/48
Głowacka B. (2007), Poza dostarczone informacje, *Języki Obce w Szkole 3*, 79–90.
Goertz, M., & Duffy, M. (2003). Mapping The Landscape of High-Stakes Testing and Accountability Programs. *Theory into Practice*, 42, 4–11.
Goffman, E. (1981). *Forms of Talk.* Oxford: Basil Blackwell.
Goldman, L. (1961). *Using Tests in Counseling*. New York: Appleton-Century-Crofts.
Goldberg, G. L., & Roswell, B. S. (1999/2000). From perception to practice: the impact of teachers' scoring experience on performance-based instruction and classroom assessment. *Educational Assessment, 6*(4), 257–290.
Gosa, C. (2004). *Investigating washback: A Case Study Using Student Diaries*. Unpublished PhD thesis, Lancaster University, Lancaster.
Goslin, D. A. (1967). The social impact of testing. *Personnel and Guidance Journal, 45*, 676–682.
Gribble, C., Blackmore, J., Morrissey, A. & Capic, T. (2016). *Investigating the Use of IELTS in Determining Employment, Migration and Professional Registration Outcomes in Healthcare and Early Childcare Education in Australia.* IELTS

Research Reports Online Series. Retrieved from http://search.informit.com.au/documentSummary;dn=011159594009259;res=IELHSS

Green, A. (2003). *Test Impact and English for Academic Purposes: A Comparative Study in Backwash between IELTS Preparation and University Pre-Sessional Courses.* Unpublished PhD thesis. University of Surrey, Roehampton.

Green, A. (2006). Watching for washback: Observing the influence of the International English Language Testing System academic writing test in the classroom. *Language Assessment Quarterly, 3*, 333–368.

Green, A. (2007a). *IELTS Washback in context*. Cambridge: Cambridge University Press

Green, A. (2007b). Washback to learning outcomes: A comparative study of IELTS preparation and university pre-sessional language courses. *Assessment in Education* 14, 75–97.

Gu, X. & Saville, N. (2012). Impact of Cambridge English: Key for Schools and Preliminary for Schools – parents' perspectives in China. *Cambridge ESOL : Research Notes* 50, 48–56

Hahn, S.L., Stassen, T., & Reschke, C. (1989) Grading classroom activities: Effects on motivation and proficiency. *Foreign Language Annals, 22*(3), 421–252.

Haladyna, T. M., Nolen, S. B., & Haas, N. S. (1991). Raising standardized achievement tests scores and the origins of test score pollution. *Educational Researcher, 20*(5), 2–7.

Hamp-Lyons, L. (1997). Washback, impact and validity: ethical concerns, *Language Testing 14*, 295–303

Hamp-Lyons, L. (1998). Ethical test preparation practice: the case of TOEFL, *TESOL Quarterly 32*, 329–337.

Hamp-Lyons, L. (2000). Social, professional and individual responsibility in language testing. *System 28*, 579–591

Hamp-Lyons, L. (2007). The impact of testing practices on teaching. In J. Cummins & C. Davison (Eds.), *The International Handbook of English Language Teaching* (pp. 487–504). Norwell, MA: Springer.

Haney, W. (2000). The myth of the Texas miracle in education. *Education Analysis Policy Archives*, *8*(21). Retrieved from http://epaa.asu.edu/epaa/v9n2.html.

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management,* 24(2), 297–327.

Hargreaves, A. (1994). *Changing Teachers, Changing Times: Teachers' Work and Culture in the Post-modern Age.* London: Cassell.

Hargreaves, E. (1997). The Diploma Disease in Egypt: Learning, Teaching and the Monster of the Secondary Leaving Certificate. *Assessment in Education: Principles, Policy & Practice, 4*(1), 161–176.

Harley, B., Allen, P., Cummins, J., & Swain, M. (Eds.) (1990). *The Development of Second Language Proficiency*. Cambridge: Cambridge University Press.

Harris, D. & Herrington, C. (2006). Accountability, standards, and the growing achievement gap: Lessons from the past half-century. *American Journal of Education, 112*(2), 209–238.

Hart, R., Moro, M. & Roberts, J. E. (2012). Date of birth, family background, and the 11 plus exam: short- and long-term consequences of the 1944 secondary education reforms in England and Wales. *University of Stirling Economics Discussion Papers 2012–10*. Retrieved from http://www.management.stir.ac.uk/research/economics/working-papers.

Hattie, J. (2015a). *What Doesn't Work in Education: The Politics of Distraction*. Pearson.

Hattie, J. (2015b). *What Works Best in Education: The Politics of Collaborative Expertise*. Pearson.

Hayes, B., & Read, J. (2004). IELTS test preparation in New Zealand: Preparing students for the IELTS academic module. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in Language Testing: Research Contexts and Methods* (pp. 97–112). Mahwah, NJ: Lawrence Erlbaum Associates.

Hawkey, R. (2006). *Impact Theory and Practice. Studies of the IELTS test and the ProgettoLingue 2000. Studies in Language Testing 24*. Cambridge: Cambridge University Press.

Hawkey, R. (2006b). Teacher and learner perceptions of language learning activity. *English Language Teaching Journal 60,* 242–252.

Hawthorne, L. (1997). The political dimension of English language testing in Australia. *Language Testing, 14* (3), 248–260.

Head, K. & Taylor, P. 1997. *Readings in Teacher Development*. Oxford: Heinemann ELT.

Henrichsen, L. (1989). *Diffusion of innovations in English language teaching: The ELEC effort in Japan 1956–1968*. New York: Greenwood Press.

Heubert, J. P., & Hauser, R. (Eds.). (1999). *High-Stakes: Testing for Tracking, Promoting, and Graduation*. Washington, D. C: National Academy Press.

Heyneman, S. P. (1987). Use of examinations in developing countries: Selection, research, and education sector management. *International Journal of Education Development* 7, 251–263.

Herman, J. L., Dreyfus, J., & Golan, S. (1990). *The Effects of Testing on Teaching and Learning* (CSE Tech. Rep. No. 327, Grant No.OERI-G–86-0003). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing. Retrieved from http://www.cse.ucla.edu/products/reports/TR327.pdf.

Hill, K. T., & Sarason, S. B. (1966). The relation of test anxiety and defensiveness to test and school performance over the elementary years: A future longitudinal study. *Monograph of the Society for Research in Child Development, 31*, 1–76.

Hillocks, G. (2002). *The Testing Trap: How State Writing Assessments Control Learning*. New York: Teachers College Press.

Hills, D. A., & Williams, J. E. (1965). Effects of test information upon self-evaluation in brief educational-vocational counseling. *Journal of Counseling Psychology, 12*, 275–281.

Holstein, J. A. and J. F. Gubrium. (1995). *The Active Interview*. London: Sage.

Holstein, J. A. and J. F. Gubrium. (2003). 'Introduction' in J. A. Holstein and J.F. Gubrium (Eds.). *Inside Interviewing: New Lenses, New Concerns*. (pp. 3–30).London: Sage.
House of Commons, Children Schools and Families Committee. (2008) *Testing and Assessment. Third Report of Session 2007–2008. Volume 1.* London: The Stationery Office Ltd.
Hughes, A. (1993). *Backwash and TOEFL 2000*, unpublished manuscript, Reading, University of Reading.
Huhta, A., Kalaja, P. & Pitkänen-Huhta, A. (1996) Discursive construction of a high-stakes test: the many faces of a test-taker. *Language Testing 23*, 326–350.
Hursh, D. W. (2007). Exacerbating inequality: The failed promise of the No Child Left Behind Act. *Race, Ethnicity, and Education, 10*, 295–308.
Hutchings, M. (2015). *Exam Factories? The Impact of Accountability Measures on Children and Young People.* National Union of Teachers. Retrieved from NUT.teachers.org.uk.
Hymes, D. (1972) On communicative competence. In J. B. Pride & J. Holmes (Eds.). *Sociolinguistics, Selected Readings.* Harmondsworth: Penguin.
ILTA. (2000). *Code of Ethics*. Retrieved from http://www.iltaonline.com/page/CodeofEthics
Jacob, A. and M. Kirst (1999). New Demands and Concepts for Educational Accountability:
Striving for Results in an Era of Excellence. In J. Murphy and K. Seashore Lewis (Eds.), *Handbook of Research on Educational Administration*, San Francisco, CA: Jossey-Bass.
Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TOEFL 2000Framework: A Working Paper* (TOEFL Monograph No. MS–16). Princeton, NJ: ETS.
Jarząbek, A.D. (2007), Kilka refleksji o egzaminie z języka obcego w gimnazjum, *Języki Obce w Szkole 5*, 67–70.
Jarząbek A. D. (2009). Nauczyciele języków obcych w świetle badań kontekstowych towarzyszących pilotażowemu egzaminowi z języka obcego w gimnazjum. In M. Pawlak, M. Derenowski & B. Wolski (Eds.) *Problemy Współczesnej Dydaktyki Języków Obcych* (pp. 331–344). Poznań–Kalisz. PWSZ Konin. UAM Poznań.
Joint Committee on Testing Practices, (1988), *Code of Fair Testing Practices in Education* Washington, DC.
Jones, M. G., Jones, B. D., & Hargrove, T. Y. (2003). *The Unintended Consequences of High Stakes Testing*. Lanham, MD: Rowman and Littlefield.
Kaftandjieva, F. (2007). Quantifying the quality of linkage between language examinations and the CEF. In C. Carlsen & E. Moe (Eds.), *A Human Touch to Language Testing* (pp. 33–43). Oslo: Novus Press.
Kagan, J. (1968). On cultural deprivation. In D. C. Glass (Ed.), *Biology and Behavior: Environmental Influences*. New York: Russell Sage Foundation.
Kane, M. (1990). *An Argument-based Approach to Validation.* ACT Research Report. Iowa City, Iowa: American College Testing Program.

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin*, *112*, 527–535.
Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319–342.
Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*, 31–41.
Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives, 2*, 135–170.
Kane, M. T. (2006). Validation. In Brennan, R. L. (Ed.), *Educational measurement,* (4th ed.) (pp. 18–64). Washington, DC: American Council on Education/ Praeger.
Kane, M. (2010). Validity and Fairness. *Language Testing 27*, 177–182.
Kane, M. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement 50*, 1–73.
Kane, M. (2016). Explicating validity. *Assessment in Education : Principles, Policies and Practice*, 198–211.
Kane, M., Crooks, T. & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*, 5–17.
Kane, T. & Staiger, D. (2002). *Improving School Accountability Systems*. National Bureau of Economic Research, NBER Working Paper, 8156.
Khalifa, H., Khabbazbashi, N., Abdelsalam, S. & Elmahdy Said, M. (2014). An investigation into the impact of a workplace English language programme in Egypt. *Cambridge ESOL : Research Notes* 58, 20–26.
Khalifa, H., Nguyen, T., &Walker, C. (2012). An investigation into the effect of intensive language provision and external assessment in primary education in Ho Chi Minh City, Vietnam *Cambridge ESOL : Research Notes* 50, 8–19.
Khalifa, H., Papp, S., Valero, R. & Videl, F. (2014). Measuring the effectiveness of Teaching Knowledge Test (TKT): Mexico case study. *Cambridge ESOL : Research Notes* 58, 50–57.
Kim, J. S., & Sunderman, G. L. (2005). Measuring academic proficiency under the No Child Left Behind Act: Implications for educational equity. *Educational Researcher, 34*(8), 3–13.
Kirkland, M. (1971). The effect of tests on students and schools. *Review of Educational Research, 41*(4), 303–350.
Kohn, A. (2000). Burnt at the high stakes. *Journal of Teacher Education*, *51*(4), 315–327.
Koretz, D. (1996). Using student assessments for accountability. In National Research Council. *Improving America's Schools: The Role of Incentives*. (pp. 171–196) Washington, DC: The National Academies Press.
Koretz, D. (2008). *Measuring Up: What Educational Testing Really Tells Us*. Cambridge, MA: Harvard University Press.
Kowalczyk-Rumak, E. (2012). Ewaluacja w szkole-najistotniejsze zmiany w nadzorze pedagogicznym. *Trendy 2012/3.* Retrieved from http://www.bc.ore.edu.pl/Content/363/Trendy+3+z+2012_E.+Kowalczyk-Rumak.pdf

Kunnan, A. (2000). Fairness and Justice for all. In A. J. Kunnan (Ed.), *Fairness and Validation in Language Assessment* (pp. 1–14). Cambridge: Cambridge University Press.
Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European Language Testing in a Global Context* (pp. 27–48). Cambridge: Cambridge University Press.
Kunnan, A. J. (2008). Towards a model of test evaluation: Using the Test Fairness and Wider Context frameworks. In L. Taylor & C. Weir (Eds.), *Multilingualism and Assessment: Achieving Transparency, Assuring quality, Sustaining Diversity* (pp. 229–251). Cambridge: Cambridge University Press.
Kvale, S. (1994). Ten standard objections to qualitative research interviews. *Journal of Phenomenological Psychology, 25* (2), 147–173.
Ladd, H., Walsh, R. (2002). Implementing value-added measures of school effectiveness: Getting the incentives right. *Economics of Education Review. 21* (1), 1–17.
Lam, H. P. (1993). *Washback-Can it be Quantified? A Study of the Impact of English Examinations in Hong Kong.* Unpublished MA thesis, University of Leeds, UK.
Lam, H. P. (1994). Methodology washback—an insider's view. In D. Nunan, R. Berry & V. Berry (Eds.), *Bringing about Change in Language Education: Proceedings of the International Language in Education Conference 1994* (pp. 83–102). Hong Kong: University of Hong Kong.
Lantolf, J., & Pavlenko, A. (2001). Second language activity theory: Understanding second language learners as people. In M. Breen (Ed.), *Thought and Actions in Second Language Learning: Research on Learner Contributions* (pp. 141–158). London: Longman.
Lapayese, Y. V. (2007). Understanding and undermining the racio-economic agenda of No Child Left Behind: Using critical race methodology to investigate the labor of bilingual children. *Race, Ethnicity, and Education, 10*, 309–321.
Lee, J. (2006). *Tracking Achievement Gaps and Assessing the Impact of NCLB on the Gaps: An In-depth Look into National and State Reading and Math Outcome Trends*. Cambridge, MA: Harvard Civil Rights Project.
Lemke, J. (1995). *Textual politics: Discourse and Social Dynamics*. London: Taylor and Francis.
Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.
Lewkowicz, J., Zawadowska-Kittel, E. (2008). Impact of the new school leaving examination in Poland. Cambridge: ESOL *Research Notes, 34*, 27–31.
Lewkowicz, J., Zawadowska-Kittel, E. (2011). Classroom-based assessment – the Polish experience. In D. Tsagari & I. Csepes, (Eds.) *Classroom Based Language Assessment.* (pp. 147–167), Frankfurt am Mein: Peter Lang.
Li, X. (1990). How Powerful Can a Language Test Be? The MET in China. *Journal of Multilingual and Multicultural Development, 11*(5), 393–404.
Ligęza, A. Analiza danych dotyczących wymagania „Analizuje się wyniki sprawdzaniu, egzaminu gimnazjalnego, egzaminu maturalnego i egzaminu potwierdzającego kwalifikacje zawodowe"… In G. Mazurkiewicz (Ed.), *Jakość*

Edukacji. Dane i Wnioski z Ewaluacji Zewnętrznych w Latach 2010–2011. (pp. 19–26.) Kraków: Ośrodek Rozwoju Edukacji.
Ligęza, A. & Franczak, J. (2012) Jak analizuje się wyniki egzaminów zewnętrznych w polskich szkołach? Raport z wyników ewaluacji zewnętrznej. Retrieved from http://www.npseo.pl/action/start/227
Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4–16.
Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district results to national norms: The validity of the claims that "everyone is above average." *Educational Measurement: Issues and Practice*, *9*(3), 5–14.
Lipska, M. (2005). Wykorzystanie wyników egzaminów zewnętrznych w procesie sprawowania nadzoru pedagogicznego przez dyrektora szkoły. Wybrane aspekty. In *Egzamin Biuletyn Badawczy Centralnej Komisji Egzaminacyjnej 4*/2005. (pp. 25–33) Warszawa: CKE.
Lisiecka, Z. (2005). Wykorzystywanie wyników egzaminów zewnętrznych w procesie podnoszenia jakości pracy szkoły. Wyniki badań ewaluacyjnych. In *Egzamin Biuletyn Badawczy Centralnej Komisji Egzaminacyjnej 4*/2005. (pp. 61–66) Warszawa: CKE.
Lopriore, L. (2002). The teaching of EFL in the Italian Context: Issues and Implications. *CAUCE, Revista de filologia y su didactica* 25(2), 203–223.
Lloyd-Jones, G., Neame, C. and Medaney, S. (2012). A multiple case study of the relationship between the indicators of students' English language competence on entry and students' academic progress at an international postgraduate university [online]. In: *IELTS Research Reports Volume 11*, 2012, (2nd ed.) (pp. 1–54), Melbourne: IDP: IELTS Australia and British Council.
Lumley, T. & B. Stoneman (2000). Conflicting perspectives on the role of test preparation in relation to learning? *Hong Kong Journal of Applied Linguistics* 5, 50–80.
Luttrell, W. (Ed.). (2010). *Qualitative Educational Research: Readings in Reflexive Methodology and Transformative Practice*. New York, NY: Routledge.
Kiss-Gulyas, J. (2001). Experiencing the Examination Design, Content, Materials and Procedures. In J. G. Egyud, I. A. Gal & P. Glover (Eds.), *English Language Education in Hungary, Part III: Training Teachers for New Examinations* (pp. 40–58). Budapest: The British Council.
Madaus, G. F. (1985). Test Scores as Administrative Mechanisms in Educational Policy. *The Phi Delta Kappan, 66*(9), 611–617.
Madaus, G. F. (1988a). The distortion of teaching and testing: High stakes testing and instruction. *Peabody Journal of Education. Vol. 65*(3), 29–46.
Madaus, G. F. (1988). The influence of testing on the curriculum. In L. N. Tanner (Ed.), *Critical Issues in Curriculum: Eighty-seventh Yearbook of the National Society for the Study of Education* (pp. 83–121). Chicago, IL: University of Chicago Press.
Madaus, G. F. & Russell, M. (2010). The Paradoxes of High-Stakes Testing. *Journal of Education, 190,* 21–30.

Mahler, C., & Smallenburg, H. (1963). Effects of testing programs on the attitudes of students. teachers, parents, and community. In W.G. Findlay (Ed.), *The Impact and Improvement of School Testing Programs. Sixty-second Yearbook of the National Society for the Study of Education, Part II*. (pp. 103–125) Chicago: National Society for the Study of Education.

Mason, R. & Ross, A. (2016). Theresa May's grammar schools plan faces cross-party backlash. *The Guardian* 7 August. Retrieved from https://www.theguardian.com/education/2016/aug/07/senior-tories-likely-to-resist-theresa-mays-grammar-schools-agenda.

Mason, R. & Sherwood, H. (2016a). Cameron 'stigmatizing Muslim women' with English language policy. *The Guardian*. Monday 18 January. http://www.theguardian.com/politics/2016/jan/18/david-cameron-stigmatising-muslim-women-learn-english-language-policy.

Mason, R. & Sherwood, H. (2016b). Migrant spouses who fail English test may have to leave UK, says Cameron. *The Guardian*. Monday 18 January. https://www.theguardian.com/uk-news/2016/jan/18/pm-migrant-spouses-who-fail-english-test-may-have-to-leave-uk.

Matuszczak, K. & Wasilewska, O. (2015). Wyniki egzaminów zewnętrznych w pracy szkoły. In R. Dolata, R. & M. Sitek. (Eds.) *Raport o Stanie Edukacji 2014. Egzaminy Zewnętrzne w Polityce i Praktyce Edukacyjnej* (pp. 221–262), Warszawa: Instytut Badań Edukacyjnych.

Markee, N. (1997). *Managing Curricular Innovation*. Cambridge: Cambridge University Press.

Mathis, W. J. (2010). *The 'Common Core' Standards Initiative: An Effective Reform Tool?* Boulder, CO, and Tempe, AZ: Education and the Public Interest Center & Education Policy Research Unit. Retrieved from http://epicpolicy.org/publication/common-core-standards.

Mazurkiewicz, G. (2011), Ewaluacja w nadzorze pedagogicznym. Model i system wartości, In B. Niemierko & M. K. Szmigiel (Eds.), *Ewaluacja w Edukacji. Koncepcje, Metody, Perspektywy. Materiały z XVII Krajowej Konferencji Diagnostyki Edukacyjnej*, (pp. 311–318), Kraków: PTDE.

McCallum, B., Gipps, C. McAlister, S. & Brown, M. (1995). National curriculum assessment: Emerging models of teacher assessment. In Torrance, H. (Ed.) *Evaluating Authentic Assessment: Problems and Possibilities: in New Approaches to Assessment.* (pp. 88–104), Buckingham: Open University Press.

McNamara, T. F. (1996): *Measuring Second Language Performance*. London and New York: Longman.

McNamara, T. F. (1999) Language testing: users and uses. In B. Spolsky & R.E. Asher (Eds.) *Concise Encyclopaedia of Educational Linguistics.* (pp. 724–728). Oxford: Elsevier Science.

McNamara, T. F. (2001). Language assessment as social practice: challenges for research. *Language Testing 18*(4) 333–349.

McNamara, T. F. (2011). Managing learning: Authority and language assessment. *Language Teaching, 44*, 500–515.

McNamara, T. F. & Roever, C. (2006). *Language Testing: the Social Dimension.* Oxford: Blackwell.

Mehrens, W. (1998). Consequences of Assessment: What is the Evidence? *Education Policy Analysis Archives Volume 6*, 1–30.
Mehrens, W. A., & Kaminsky, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless or fraudulent? *Educational Measurement: Issues and Practice, 8*(1), 14–22.
Merkhofer, B. E. (1954). *College Students' Study Behavior*. Unpublished PhD thesis, University of Chicago.
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist,* 50: 741–9.
Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241–256.
Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 33–45). Hillsdale, NJ: Lawrence Erlbaum.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 12–103). New York: Macmillan.
Meyer, G. (1935). An experimental study of the old and new types of examinations. *Journal of Educational Psychology*, 26, 30–40.
Milanovic, M. & Saville, N. (1996). *Considering the Impact of Cambridge EFL exams*. Internal Report. Cambridge: Cambridge ESOL.
Milecka, B. (2014a). Jak analizuje wyniki egzaminów zewnętrzne w polskich szkołach. Raport z wyników ewaluacji zewnętrznej. In G. Mazurkiewicz & A. Gocławska (Eds.) *Jakości Edukacji. Dane i Wnioski z Ewaluacji Zewnętrznych Prowadzonych w Latach 2012–2013*. (pp. 187–196), Krakow: Wydawnictwo Uniwersytet Jagiellońskiego.
Milecka, B. (2014b). W jaki sposób i jakie analizy wykorzystuje się w polskich szkołach do doskonalenia procesów edukacyjnych. In G. Mazurkiewicz & A. Gocławska (Eds.) *Jakości Edukacji. Dane i Wnioski z Ewaluacji Zewnętrznych Prowadzonych w Latach 2013–2014.* (pp. 213–228). Krakow: Wydawnictwo Uniwersytet Jagiellońskiego.
Milroy, J. & Milroy, L. (1999). *Authority in Language. Investigating Standard English* (3rd ed.) London: Routledge.
Ministerstwo Edukacji Narodowej (1990) *Języki Zachodnio-Europejskie. Program Zatwierdzony do Realizacji w Klasach I–VIII od Roku Szkolnego 1990/1991.*
Ministerstwo Edukacji Narodowej. (1998). *Reforma Systemu Edukacji. Projekt*. Warszawa: Wydawnictwa Szkolne i Pedagogiczne.
Ministerstwo Edukacji Narodowej (1999) Rozporządzenie Ministra Edukacji Narodowej z 15 lutego 1999 w sprawie podstawy programowej kształcenia ogólnego, (Dz.U. 1999.14.129.).
Ministerstwo Edukacji Narodowej. (2007). Rozporządzenie Ministra Edukacji Narodowej z dnia 23 sierpnia 2007 r. zmieniające rozporządzenie w sprawie podstawy programowej wychowania przedszkolnego oraz kształcenia ogólnego w poszczególnych typach szkół.

Ministerstwo Edukacji Narodowej. (2009). Rozporządzenie z dnia 7.10.2009 w sprawie nadzoru pedagogicznego (Dz. U. nr 168, poz. 1324).
Ministerstwo Edukacji Narodowej. (2012). Rozporządzenie z dnia 19.11.2009 w sprawie placówek doskonalenia (Dz. U. nr 200, poz. 1537, z późn. zm.).
Ministerstwo Edukacji Narodowej. (2013). Rozporządzenie z dnia 10.05.2013 zmieniające rozporządzenie w sprawie nadzoru pedagogicznego (Dz. U. poz. 560).
Ministerstwo Edukacji Narodowej. (2015). Rozporządzenie z dnia 6.08.2015 w sprawie wymagań wobec szkół i placówek, informacji ze sprawdzaniu oraz egzaminów.
Ministerstwo Edukacji Narodowej. (2015). Rozporządzenie z dnia 27.08.2015 w sprawie nadzoru pedagogicznego.
Ministerstwo Edukacji Narodowej i Sportu. (2004). Rozporządzenie z dn. 23.04.2004 r. w sprawie szczegółowych zasad sprawowania nadzoru pedagogicznego, wykazu stanowisk wymagających kwalifikacji pedagogicznych, kwalifikacji niezbędnych do sprawowania nadzoru pedagogicznego, a także kwalifikacji osób, którym można zlecać prowadzenie badań i opracowywanie ekspertyz, (Dz.U. Nr 89 z 2004 r., poz. 845).
Mizerek, H. (2010). Efektywna autoewaluacja w szkole – jak ją sensownie zaprojektować i przeprowadzić?, In G. Mazurkiewicz (Ed.), *Ewaluacja w Nadzorze Pedagogicznym. Autonomia*. (pp. 19–62), Krakow: Wydawnictwo Uniwersytet Jagiellońskiego.
Mizutani, S. (2009). *The Mechanism of Washback on Teaching and Learning*. Unpublished PhD thesis. University of Auckland, New Zealand.
Modu, C. C. (1969). *Affective Consequences of Cognitive Changes*. Unpublished PhD thesis, University of Chicago.
Moore, S., Mahony, P. Stroupe, R. (2012). Perceptions of IELTS in Cambodia: A casestudy of test impact in a small developing country. In J. Osborne (Ed.). *IELTS Research Reports Volume 13*, 2012. (pp. 1–109), Melbourne: IDP: IELTS Australia and British Council.
Morrison, K., & Tang, F. H. J. (2002). Testing to Destruction: A Problem in a Small State. *Assessment in Education: Principles, Policy & Practice, 9*(3), 289–317.
Morrow, K. (1986). The evaluation of tests of communicative performance. In M. Portal (Ed.) *Innovations in Language Testing.* (pp. 1–13), Windsor: NFER/Nelson.
Morrow, K. (Ed.) (2004). *Insights from the Common European Framework.* Oxford: Oxford University Press.
Munoz, A. P. & Alvarez, M. E. (2010). Washback of an oral assessment system in the EFL classroom. *Language Testing* 27,1–17.
Murray, J. C., Cross, J. L. & Cruickshank, K. 2014. *Stakeholder Perceptions of IELTS as a Gateway to the Professional Workplace: The Case of Employers of Overseas Trained Teachers* IELTS Research Reports Online. Retrieved from Serieshttp://search.informit.com.au/documentSummary; dn=400519847610740;res=IELHSS

Murray, J., A. Riazi & J., Cross (2012). Test candidates' attitudes and their relationship to demographic and experiential variables: The case of overseas trained teachers in NSW, Australia. *Language Testing* 29, 577–595.
NEA Research Bulletin, (1962).
Nevo, D. (1989). *Useful Evaluation*. Tel Aviv: Massada.
Nichols, S. L. (2007). High-stakes testing: Does it increase achievement? *Journal of Applied School Psychology, 23*(2), 47–64.
Nicholls, A. (1983). *Managing educational innovations*. London: Allen & Unwin.
Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220.
Nitko, A. (1989). Designing Tests That Are Integrated with Instruction. In Linn. R. (ed.) *Educational Measurement.* New York: ACE/Macmillan, 447–74.
Nichols, S. L., & Berliner, D. C. (2008). Why Has High-Stakes Testing So Easily Slipped Into Contemporary American Life? *The Phi Delta Kappan, 89*(9), 672–676.
Nichols, S. L., & Berliner, D. C. (2007). *Collateral Damage: How High-Stakes Testing Corrupts America's Schools*. Cambridge, Massachusetts: Harvard Education Press.
Nichols, S. L., Glass, G. V., & Berliner, D. C. (2006). High-stakes testing and student achievement: Does accountability pressure increase student learning? *Education Policy Analysis Archives, 14*(1). Retrieved from http://epaa.asu.edu/epaa/v14n1.
Pearson, I. (1988). Tests as levers for change (or 'Putting first things first') in D. Chamberlain, & R. Baumgardner (Eds.). *ESP in the classroom: Practice and Evaluation*. ELT Documents 128. (pp. 98–107), London: Modern English Publications.
Piotrowska, E. & Skulski, B. (2009). *Klucze Sukcesu.* Poland: Macmillan
Obidniak, D. (2007). Egzamin z języka obcego w gimnazjum – czy nowe otwarcie w polskiej polityce językowej?, *Języki Obce w Szkole 5*, 65–67.
O'Day, J. (2004). Complexity, Accountability, and School Improvement. In S. Fuhrman & R. Elmore. *Redesigning Accountability Systems for Education.* (pp.15–46), New York: Teachers College Press.
OECD (2013). *Synergies for Better Learning: An International Perspective on Evaluation and Assessment in Education*. Paris: OECD Publishing.
OECD (2014). *PISA 2012 Results in Focus What 15-year-olds Know and What They Can Do With What They Know*. Paris: OECD Publishing.
OECD (2017). Internet access (indicator). doi: 10.1787/69c2b997-en.
Osler, S. F. (1954). Intellectual performance as a function of 2 types of psychological. *Journal of Experimental Psychology,* 47, 115–121.
Paczuska, K., Kutyłowska, K., Gajewska-Dyszkiewicz, A., Ellis, M. & Szpotowicz, M. (2014) *Język Angielski w Gimnazjum. Raport cząstkowy z 1 etapu Badania Uczenia się i Nauczania Języków Obcych w Gimnazjum*. Warszawa: Instytut Badań Edukacyjnych.
Page, E. B. (1958). Teacher comments and student performance: A seventy-four classroom experiment in school motivation. *Journal of Educational Psychology*, 173–181.

Paris, S. G., Turner, J. C., Lawton, T. A. & Roth, J. L. (1991). A developmental perspective on standardized achievement testing. *Educational Researcher*, 20, 12–20.

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., & Miao, J. (2003). *Perceived Effects of State-Mandated Testing Programs On Teaching and Learning: Findings from A National Survey On Teachers*. Chestnut Hill, Ma: Boston College: Center for The Study of Testing, Evaluation, and Educational Policy.

Pelz, D. C. (1985). Innovation complexity and the sequence of innovating stages. *Knowledge: Creation, Diffusion and Utilization 6* (3), 261–291.

Perrone, M. (2010). *The Impact of the First Certificate of English (FCE) on the EFL Classroom: A Washback Study*. Unpublished PhD thesis, Teachers College, Columbia University.

Perrone, M. (2011). The effect of classroom-based assessment and language processing on the second language acquisition of EFL students. *Journal of Adult Education 40*, 20–33.

Pezda, A. (2009). Gimnazja: angielski miejski i wiejski. Retrieved from http://wyborcza.pl/1,75248,6706874,gimnazja_angielski_miejski_i_wiejski.html.

Pill, J. & Harding, L. (2013). Defining the language assessment literacy gap: evidence from a parliamentary inquiry. *Language Testing, 30*, 381–402.

Plank, S. & Condliffe, B. (2013). Pressures of the season. An examination of classroom quality and high-stakes accountability. *American Educational Research Journal. Vol. 50*/5, 1152–1182.

Pollaczek, P. P. (1952). A study of malingering on the CVS abbreviated individual intelligence scale. *Journal of Clinical Psychology, 8*, 75–8.

Popham, W.J. (1987). The merits of measurement-driven instruction. *The Phi Delta Kappan*, 68(9), 679–682.

Popham, W. J. (1991). Appropriateness of teachers' test-preparation practices. *Educational Measurement: Issues and Practice, 10*(4), 12–15.

Popham, W. J. (2005). *Classroom Assessment: What Teachers Need to Know* (4th Ed.). Boston, MA: Pearson Education. 300–310.

Popham, W. J., Cruse, K., Rankin, S., Sandifer, P. & Williams, P. (1985). Measurement-driven instruction: It's on the road. *The Phi Delta Kappan*. 66 (9), 628–634.

Popper, K. R. (1965). *Conjecture and Refutation: The Growth of Scientific Knowledge*. New York: Harper & Row.

Popper, K. R. (1968). *The Logic of Scientific Discovery*. New York: Harper & Row.

Poszytek, P. (2008). *European Standards of Testing Language Proficiency and the Polish "Nowa Matura" Examination in English*, Krakow: Universitas.

Potter, J. & Hepburn, A. (2012). Eight challenges for interview researchers. In Gubrium, J.F. and Holstein, J.A. *Handbook of Interview Research* (2nd edition). (pp. 555–570), London: Sage. Text referred to came from pre-print version (2011) retrieved from https://www.researchgate.net/publication/292653443_Eight_challenges_for_interview_researchers.

Qi, L. (2004). Has a high-stakes test produced the intended changes? In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in Language Testing: Research Contexts and Methods* (pp. 171–190). Mahwah, NJ: Lawrence Erlbaum Associates.

Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing 22*(2), 142–173.

Rao, C., McPherson, K. Chand, K. & Khan, V. (2003). *Assessing the Impact of IELTS Preparation Programs on the General Training Reading and Writing Test Modules.* IELTS Research Report (vol. 4). Retrieved from www.ielts.org/pdf/Vol5Report5.pdf.

Ravitch, D., & Mathis, W. J. (2010). *Review of 'College- and Career-Ready Students.'*

Boulder, CO: National Education Policy Institute. Retrieved from http://nepc.colorado.edu/thinktank/review-college-and-career.

Rea-Dickins, P. & Germaine, K. (1992). *Evaluation*. Oxford: Oxford University Press.

Read, J., & Hayes, B. (2003). The impact of IELTS on preparation for academic study in New Zealand. In R. Tulloh (Ed.), *IELTS research reports 2003* (vol. 4), (pp. 154–205). Canberra: IELTS Australia.

Resnick, L. & Resnick, D. (1992). Assessing the thinking curriculum: new tools for educational reform. In B. Gifford. & M. O'Connor. (Eds.) *Changing Assessments: Alternative Views of Aptitude, Achievement and Instruction.* (pp. 37–75), Boston, MA: Kluwer Academic.

Rodriguez, M. C. (2004). The role of classroom assessment in student performance on TIMSS. *Applied Measurement in Education,* 17(1), 1–24.

Rogers, E.M. (1995). *The Diffusion of Innovations 4*. New York: Free Press.

Rogers, E. M., & Shoemaker, F. F. (1971). *Communication of Innovations: A Cross-cultural Approach*. New York: Free Press.

Rolniczy Magazyn Elektroniczny. (2011). Język angielski główną przeszkodą uczniów z terenów wiejskich w dostępie do edukacji na poziomie ponadgimnazjalnym. styczeń-luty 2011, nr 41. Retrieved from http://www.cbr.edu.pl/rmearchiwum/2011/rme41/dane/5_10.html

Romberg, T. A., Zarinnia, A., & Williams, S. R. (1989). *The Influence of Mandated Testing on Mathematics Instruction: Grade 8 teachers' Perceptions* (Monograph). Madison, WI: University of Wisconsin, National Center for Research in Mathematical Science Education.

Rosenshine, B. 2003. High-stakes testing: Another analysis. *Education Policy Analysis Archives, 11*(24). Retrieved from http://epaa.asu.edu/epaa/v11n24/.

Roulston, K. & Shelton, S. A. (2015). Reconceptualizing bias in teaching qualitative research methods. *Qualitative inquiry*, 1–11.

Ruebuch, B. E. Anxiety. In H. W. Stevenson (Ed.), (1963) *Child Psychology. Sixty-second.*

Yearbook National Society for the Study of Education, Part I. (pp. 460–516), Chicago, IL: National Society for the Study of Education, University of Chicago Press.

Sacks, P. (1999). *Standardized Minds: The High Price of America's Testing Culture and What We Can Do to Change It*. Cambridge, MA: Perseus Publishing.
Saif, S. (2006). Aiming for Positive Washback: A Case Study of International Teaching Assistants. *Language Testing, 23*, 1–34.
Sandlin, A. A. (1959). *Factors Influencing Curriculum in Oregon schools*. Eugene, University of Oregon.
Salamoura, A., French, A. & Emery, J. (2014). Enhancing employment opportunities through English: An investigation into the impact of the SCOPE programme in India. *Cambridge ESOL : Research Notes* 58, 27–39.
Salehi, H., Naseri Karimvand,P., & Kashefian Naeeini, S. (2011). Various aspects of high-stakes school tests: A close look at the phenomenon of washback. Paper presented at 24th International Congress for School Effectiveness and Improvement. January 4–7, Limassol, Cyprus.
Sarason, S. B., Davidson, K. S., Lighthall, F. F., Waite, R. R., & Ruebuch, B. K. (1960). *Anxiety in Elementary School Children.* New York: Wiley.
Sassenrath, J. M., & Garverick, C. M. (1965). Effects of differential feedback from examinations on retention and transfer. *Journal of Educational Psychology*, *56,* 259–263.
Saville, N. (2009). *Developing a Model for Investigating the Impact of Language Assessment Within Educational Contexts by a Public Examination Provider.* Unpublished PhD Thesis. University of Bedfordshire, UK.
Saville, N. (2010). Developing a model for investigating the impact of language assessment. *Cambridge ESOL : Research Notes 42*, 2–8.
Sawicki, M. (2008). Szanowni Państwo! *Egzaminy Nasze Uczniów. Kwartalnik dla Dyrektorów i Nauczycieli 2008*/1Warszawa: CKE. Retrieved from http://www.cke-efs.pl/mod/resource/view.php?id=110.
Schiefelbein, E. (1993). The use of national assessments to improve primary education in Chile. In D. W. Chapman & L. O. Mählck (Eds.), *From Data to Action: Information Systems in Educational Planning.* Paris: International Institute for Educational Planning.
Schmitt, N., Jiang, X., & Grabe, W. (2011). The Percentage of Words Known in a Text and Reading Comprehension. *Modern Language Journal, 95*, 26–43.
Schulz, B. C. (2008). *Teachers' Perspectives of How High-stakes Testing Influences Instructional Decisions and Professionalism.* Athens, GA: University of Georgia.
Searle, J. (1969). *Speech Acts: An Essay in Philosophy of Language*. Cambridge: Cambridge University Press.
Sears, P. S. (1940). Levels of aspiration in academically successful and unsuccessful children. *Journal of Abnormal and Social Psychology*, 35, 498–536.
Sewell, W. H., & Shah, V. P. (1968). Social class, parental encouragement, and educational aspirations. *American Journal of Sociology, 73*, 559–572.
Sitek, M. (2015). Miejsce egzaminów zewnętrznych w systemach edukacyjnych i polityce edukacyjnej. In R. Dolata & M. Sitek (Eds.) *Raport o Stanie Edukacji 2014. Egzaminy Zewnętrzne w Polityce i Praktyce Edukacyjnej.* (pp. 11–29), Warszawa: Instytut Badań Edukacyjnych.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of Research in Education*, Vol. 19 (pp. 405–450). Washington, DC: American Educational Research Association.
Shih, C.-M. (2007). A New Washback Model of Students' Learning. *The Canadian Modern Language Review 64*(1), 135–161.
Shih, C-M. (2009). How tests change teaching. A model for reference. *English Teaching: Practice and Critique*. *8* (2), 188–206.
Shohamy, E. (1992). Beyond proficiency testing: a diagnostic feedback testing model for assessing foreign language learning. *The Modern Language Journal, 76*, 513–521.
Shohamy, E. (1993). The power of tests: the impact of language tests on teaching and
learning. NFLC Occasional Paper. College Park, MD: National Foreign Language Center, University of Maryland.
Shohamy, E. (1998). Critical language testing and beyond, *Studies in Educational Evaluation* 24, 331–45.
Shohamy, E. (2001a). Democratic assessment as an alternative. *Language Testing 18*, 373–391.
Shohamy, E. (2001b). *The Power of Tests: A Critical Perspective on the Uses of Language Tests*. London: Pearson.
Shohamy, E. (2003). Implications of language education policies for language study in schools and universities, *The Modern Language Journal,* 87, 278–286.
Shohamy, E. (2006a). *Language Policy: Hidden Agendas and New Approaches.* London: Routledge.
Shohamy, E. (2007). Language tests as language policy tools. *Assessment in Education*, *14*, 117–130.
Shohamy, E., Donitsa-Schmidt, A., & Ferman, I. (1996). Test impact revisited: Washback effect over time? *Language Testing, 13*(3), 298–317.
Silverman, D. (1973). 'Interview talk: bringing off a research instrument,' *Sociology* 7/1, 32–48.
Smith, G. J. (1952). Influence of failure, expressed hostility, and stimulus characteristics on verbal learning and recognition. *Dissertation Abstracts, 12,* 600.
Smith, M. (1991). Put it to the test: the effects of external testing on teachers. *Educational Researcher 20/5*, 8–11.
Smith, M. L., and Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice*, *10*(4), 7–11.
Smolik, M. (2008). Edukacyjny i pomiarowy 'Miś', czyli o nowym egzaminie gimnazjalnym z języka obcego. *XIV Konferencja Diagnostyki Edukacyjnej, Uczenie się i Egzamin w Oczach Nauczycieli, Opole 2008*. (pp. 184–193) Polskie Towarzystwo Diagnostyki Edukacyjnej. Retrieved from www.ptde.org/mod/page/view.php?id=157.
Smolik, M. (2006) Każdy każdego własną miarą. Kilka uwag o kryteriach oceniania egzaminu ustnego z języka angielskiego na poziomie podstawowym.

In B. Niemierko & M. K. Szmigel (Eds.) *O Wyższą Jakość Egzaminów Szkolnych*, (pp. 116–134), Kraków: Grupa Tomami.
Smolik, M. (2008). *Investigating Scoring Validity. A Study of the "Nowa Matura" speaking Exam in English at the Basic Level*. Unpublished PhD thesis. UMCS. Lublin.
Smolik, M. (2012). Języki obce na egzaminach zewnętrznych: innowacje w latach 2012–2015. *Języki Obce w Szkole*, 22–33.
Snyder, C.W. Jr, Prince, B., Johanson, G., Odaet, C., Jaji, L. & Beatty, M. (1997). *Exam Fervor and Fever: Case Studies of the Influence of Primary Leaving Examinations on Uganda Classrooms, Teachers, and Pupils (Volume One), Advancing Basic Education and Literacy (ABEL) Project.* Washington DC: Academy for Educational Development.
Spratt, M. (2005). Washback and the classroom: The implications for teaching and learning of studies of washback from exams. *Language Teaching Research, 9*, 5–29.
Spyropoulou, (2010). Dr Glenn Fulcher talks to ELT. *ELT News*, June 2010. Athens: Greece
Spolsky, B. (1981). Some ethical questions about language testing. In C. Klein-Braley & D.K. Stevenson (Eds.), *Practice and Problems in Language Testing* (pp. 5–21). Frankfurt,Germany: Peter D. Lang.
Spolsky, B. (2012). Language testing and language management. In G. Fulcher & F. Davidson. (Eds.) *The Routledge Handbook of Language Testing.* (pp. 495–503), Abingdon: Routledge.
Stecher, B., Chun, T. & Barron, S. (2004). The Effects of Assessment-Driven reform on the Teaching of Writing in Washington State. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in Language Testing: Research Contexts and Methods* (pp. 53–72). Mahwah, NJ: Lawrence Erlbaum Associates.
Stiefel, L.A., Schwartz, E., Rubinstein, R., Zabel, J. (2005). *Measuring School Performance and Efficiency: Implications for Practice and Research*. Larchmont, NY: Eye on Education.
Stobart, G. (2008). *Testing Times. The Uses and Abuses of Assessment.* Abingdon: Routledge.
Stoneman, B. W. H. (2006). *The Impact of an Exit English test on Hong Kong Undergraduates: A study Investigating the Effects of Test Status on Students' Test Preparation Behaviors.* Unpublished PhD thesis, Hong Kong Polytechnic University.
Stoller, F. (1994). The diffusion of innovation in intensive ESL programs. *Applied Linguistics,15*(3), 300–327.
Stożek, E. (2012). Czy egzaminy mogą pomóc szkole w rozwoju? *Polityka Społeczna*, nr tematyczny 1, *Polityka Edukacyjna: Szanse i Wyzwania*, 30–33.
Stożek, E., Kędracka, E. & Rappe, A. (2015). Opis poziomów wykorzystania wskaźników EWD przez szkoły, Unpublished article referred to in K. Matuszak & O. Wasilewska (2015).
Sturman, L. (2003). Teaching to the Test: Science or Intuition? *Educational Researcher, 45*(3), 261–273.

Swain, M. (1985). Communicative competence: some rules of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.) *Input in Second Language Acquisition* (pp. 235–253). Rowley, MA: Newbury House.

Szaleniec, H. (2010). Czy egzaminy zewnętrzne wpływają na wewnątrzszkolne ocenianie i politykę edukacyjną? Paper given at XVI Konferencja Diagnostyki Edukacyjnej, Toruń 2010. Retrieved from http://www.ptde.org/mod/page/view.php?id=271.

Talmy, S. (2011). The Interview as Collaborative Achievement: Interaction, Identity and Ideology in a Speech Event. *Applied Linguistics 32*, 25–42.

Taylor, R. G. (1964). Personality traits and discrepant achievement: A review. *Journal of Counseling Psychology, 11*, 76–81.

Taylor, L. (2000). Stakeholders in language testing. *Research Notes* 2, 2–4. Retrieved from http://www.cambridgeenglish.org/images/22642-research-notes-02.pdf

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing 30*(3) 403–412.

Teasdale, A., & Leung, C. (2000). Teacher assessment and psychometric theory: A case of paradigm crossing? *Language Testing, 17*(2), 163–184.

Terry, P. W. (1933). How students review for objective and essay tests. *Elementary School Journal, 33*, 592–603.

Thelen, E. & Smith, L. B. (1996). *A Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge, MA: The MIT Press.

Toulmin, S. E. (2003). *The Uses of Argument*. Cambridge: Cambridge University Press.

Traxler, A. E. (1958). Are the professional test-makers determining what we teach? *School Review, 64*, 141–151.

Trochim, W. (2006). Introduction to evaluation. Retrieved from http://www.socialresearchmethods.net/kb/intreval.htm

Tsagari, D. (2006). *Investigating the Washback Effect of a High-stakes EFL exam in the Greek Context: Participants' Perceptions, Materials Design and Classroom Applications*.Unpublished doctoral dissertation. Lancaster, United Kingdom: Lancaster University, Lancaster.

Tsagari, D. (2007). Review of washback in language testing: How has been done? What more needs doing? University of Lancaster. Retrieved from http://files.eric.ed.gov/fulltext/ED497709.pdf.

Tsagari, D. (2009). Revisiting the concept of test washback: investigating FCE in Greek language schools. *Cambridge ESOL Research Notes* 35, 5–9.

Tyler, R. W. (1966). What testing does to teachers and students. In A. Anastasi (Ed.), *Testing problems in perspective.* Washington: American Council on Education.

Valazza, G. (2015). Impact of TKT in Uruguay 2005–2012. In R. Wilson & M. Poulter. *Assessing Language Teachers' Professional Skills and Knowledge*. (pp. 263–284), Cambridge: Cambridge University Press van Ek, J. A.

& Trim, J. L. M. (1998). *Waystage 1990*. Council of Europe, Cambridge: Cambridge University Press.
Vanclay, F., Esteves, A. M., Aucamp, I. & Frank, D. M. (2015). *Social Impact Assessment. Guidance for assessing and managing the social impacts of projects.* Fargo, ND: International Association for Impact Assessment. Retrieved from http://www.kpesic.com/guidance/social-impact-assessment-guidance-assessing-and-managing-social-impacts-projects.
Varghese, N. (1998). Evaluation vs Impact studies. In V. McKay & C. Treffgarne (Eds.) *Evaluating Impact*, (pp. 47–54), London: Department for International Development.
Vernon, P. E. (1956). *The Measurement of Abilities*. (2nd ed.) London: University of London.
UK government. Home Office. (2016). New A2 English requirement in the family route. Retrieved from https://www.gov.uk/government/news/new-a2-english-requirement-in-the-family-route.
Wall, D. (1996). Introducing new tests into traditional systems: Insights from general education and from innovation theory. *Language Testing, 13*(3), 334–354.
Wall, D. (1997). Impact and Washback on language testing. In C. Clapham & D. Corson (Eds.) *Encyclopaedia of Language and Education, Vol. 7: Language Testing and Assessment,* (pp. 291–302), Netherlands: Kluwer Academic.
Wall, D. (1999). *The Impact of High-stakes Examinations on Classroom Teaching: A Case Study Using Insights from Testing and Innovation theory.* Unpublished doctoral dissertation, University of Lancaster.
Wall, D. (2000). The impact of high-stakes testing on teaching and learning: Can this be predicted or controlled? *System, 28*, 499–509.
Wall, D. (2005). *The impact of a High-stakes Examination on Classroom Teaching. Studies in Language Testing: Vol. 22.* Cambridge, UK: Cambridge University Press.
Wall, D. (2012). Examining Washback: What Do We Know, and What Is There Left To Explore? Presentation given at ALTE 41st Meeting and Conference, Lisbon, 18–20 April 2012.
Wall, D. and Alderson, J. C. (1993). Examining washback: the Sri Lankan Impact study. *Language Testing 10*, 41–69.
Wall, D., & Horák, T. (2006). *The Impact of Changes in the TOEFL Examination on Teaching and Learning in Central and Eastern Europe–Phase 1: The Baseline Study* (TOEFL Monograph Series, MS–34). Princeton, NJ: ETS.
Wall, D., & Horák, T. (2007). Using baseline studies in the investigation of test impact. *Assessment in Education 14*(1), 99–116.
Wall, D., & Horák, T. (2008). *The Impact of Changes in the TOEFL Examination on Teaching and Learning in Central and Eastern Europe–Phase 2, Coping with Change*. Princeton, NJ: ETS.
Wall, D., & Horák, T. (2011). *The Impact of Changes in the TOEFL® Exam on Teaching in a Sample of Countries in Europe: Phase 3, The Role of the Coursebook. Phase 4, Describing Change.* Princeton, NJ: ETS.

Wang, L., Eignor, D., & Enright, M. K. (2008). A final analysis. In C.A. Chapelle, M. K. Enright., & J. M. Jamieson (Eds.), *Building a Validity Argument for the Test of English as a Foreign Language* (pp. 259–318). New York, NY: Routledge.

Wasilewska, O., Rybińska, A. & Muzyk, A. (2014). *Wykorzystanie Ewaluacji Zewnętrznej i Wewnętrznej Przez Szkoły*. Warszawa: Instytut Badań Edukacyjnych.

Watanabe, Y. (1992). Washback effects of college entrance exam on language learning strategies. *JACET Bulletin 23*, 175–194.

Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing, 13*(3), 318–333.

Watanabe, Y. (1997). *The Washback Effects of the Japanese University Entrance Examinations of English : Classroom-based Research*. Unpublished PhD Thesis. University of Lancaster.

Watanabe, Y. (2000). Washback effects of the English section of the Japanese university entrance examinations on instruction in pre-college level EFL. *Language Testing Update, 27*, 42–47.

Watanabe, Y. (2001). Does the university entrance examination motivate learners? A cas study of learner interviews. In Akita Association of English Studies (Eds.) *Trans-equator exchanges: A collection of academic papers in honor of Professor David Ingram* (pp. 100–110) Akita: Japan.

Watanabe, Y. (2004). Teacher factors mediating washback. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in Language Testing: Research Contexts and Methods* (pp. 129–146). Mahwah, NJ: Lawrence Erlbaum Associates.

Watanabe, Y. (2004). Methodology in washback studies. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in Language Testing: Research Contexts and Methods* (pp. 19–36). Mahwah, NJ: Lawrence Erlbaum Associates.

Weir, C. J. (1990). *Communicative Language Testing*. New York: Prentice Hall.

Weir, C. J. (1993). *Understanding and Developing Language Tests.* Hemel Hempstead: Prentice Hall.

Weir, C. J. (2005). *Language Testing and Validation. An Evidence-based Approach*. Basingstoke: Palgrave Macmillan.

Weir, C., & Roberts, J. (1994). *Evaluation in ELT*. Oxford, UK: Blackwell Publishing.

Weiss, C. (1998). *Evaluation*. New Jersey: Prentice Hall.

Wesdorp, H. (1982). *Backwash Effects of Language Testing in Primary and Secondary Education*. Amsterdam: Stichting Centrum voor onderwijsonderzoek van de Universiteit van Amsterdam

Wesdorp, H. (1983). Backwash effects of multiple-choice language tests: Myth or reality? *Practice and Problems in Language Testing, 5*. Arnhem, The Netherlands: CITO.

Wiliam, D. (2010). Standardized testing and school accountability. *Educational Psychologist 45*(2), 107–122.

Willingham, W. W. (1999). A systemic view of test fairness. In S. Messick (Ed.), *Assessment in Higher Education: Issues in Access, Quality, Student*

Development, and Public Policy (pp. 213–242). Mahwah, NJ: Lawrence Erlbaum.
Willingham, W. W. & Cole, N. (1997). *Gender and Fair Assessment*. Mahwah, NJ: Lawrence Erlbaum.
Wrightstone, J. W. (1963). The relation of testing programs to teaching and learning. In N. B. Henry and H. G. Richey (Eds.), *Sixty-second Yearbook of the National Society for the Study of Education, Part II, The Impact and Improvement of School Testing Programs*. Chicago: University of Chicago Press.
Woods, D. (1996). *Teacher Cognition in Language Teaching: Beliefs, Decision-making and Classroom Practice*. Cambridge: Cambridge University Press.
WSiP. (1987). *Program Liceum Ogólnokształcącego. Języki Zachodnio-europejskie (Angielski, Francuski, Niemiecki)*. Warszawa: Wydawnictwa Szkolne i Pedagogiczne.
Xi, X. (2010). How do we go about investigating test fairness? *Language Testing 27*(2), 147–170.
Xie, Q. (2010). *Test Design and Use, Preparation, and Performance: A Structural Equation Modeling Study of Consequential Validity*. Unpublished PhD thesis, University of Hong Kong.
Xie, Q. (2011). Is test takers' perception of assessment demand related to construct validity? *International Journal of Testing 11*, 1–24.
Xie, Q. (2013). Does test preparation work? Implications for score validity. *Language Assessment Quarterly, 10*(2), 196–218.
Xie, Q. & Andrews, S. (2012). Do test design and uses influence test preparation? Testing a model of washback with structural equation modelling. *Language Testing, 30*(1) 49–70
Yan, Q., Gu, X., & Khalifa, H. (2014). Impact of Cambridge English: Key for Schools on young learners' English learning: Voices from students and parents in Beijing, China *Cambridge ESOL : Research Notes* 58, 44–50.
Yan, X., Maeda, Y., Lv, J. & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing, 33*/4, 497–528.
Young, R. (2012) Social dimensions of language testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge Handbook of Language Testing.* (pp. 178–193), Abingdon: Routledge.
Zakaria, N.A., Samad, A.A. & Omar, Z. (2013). Pressure to improve scores in standardized English examinations and their effects on classroom practices. *International Journal of English Language Education 2*, 45–56.
Zhan, Y. & Andrews, S. (2014). Washback effects from a high-stakes examination on out-of-class English learning: insights from possible self-theories. *Assessment in Education: Principles, Policy & Practice 21*, 71–89.

Appendix 1. Baseline Study: Teacher questionnaire

The aim of this questionnaire is to gather information about teaching and learning English in gimnazjum for research purposes. Data collected will be reported without revealing who you are, or the name of your school. Please contact me by e-mail if you would like to know more about the project. Thank you for your assistance! Melanie Ellis, NKJO Zabrze [email address].

1. Please number the following in order of importance (1-6) in your teaching in gimnazjum:

Grammar	Writing
Vocabulary	Listening
Reading	Speaking

2. What do you give grades for? Please tick all that apply

Grammar	Writing
Vocabulary	Listening
Reading	Speaking
Other. What?	

3. Where do the tests you use in your classes come from. Please tick all that apply or add a note.

Your coursebook	Your course Teacher's Book
A grammar book	Materials from the internet
Your own ideas	Tests produced by the coursebook publisher
Other coursebooks	Other. What?

4. How often do your students practice reading? (Tick ONE)

Every lesson… Once a week… When there's one in the coursebook… Rarely…

5. What kind of tasks do you do connected with reading? Tick those you do most frequently.

translation	true/false	multiple choice	fill in the gaps
students write answers to questions		matching	reading aloud

6. How often do your students listen to a recording?

Every lesson… Once a week… When there's one in the coursebook… Rarely…

7. What percentage of the time can they see the text while listening?

100% 80% 60% 40% 20% 0 (never)

8. Do you use any authentic materials with your students for reading or listening?

Yes No

9. If you answered Yes to Q8: do you use any of the following? Tick all that apply.

magazines	newspapers	internet materials	graded readers
podcasts	films	TV programs	radio programs

10. How do you help your learners develop their reading skills? Tick all that apply.

Teach how to use a bilingual dictionary	Give lots of reading practice
Teach how to use an English-English dictionary	Do lots of vocabulary tests
Teach strategies for reading	Regularly use translation
Explain step-by-step how to find the answer to a comprehension question	

11. How do you help students to do a reading comprehension exercise? Tick all that apply

Translate the text with the learners
Translate the questions for the learners if there are problems
Ask good learners to translate the questions first
Ask learners to tell you words they don't understand after reading
Pre-teach key vocabulary
Introduce the topic of the text before reading
Encourage learners to use dictionaries in class
Other. What?

12. How do you help learners with finding the answer to a listening comprehension task?

Show them the tapescript
Play the recording again
Play the problem part several times
Read the transcript aloud (or part of it)
Write problem words or phrases on the board
Other. What?

13. Which coursebook do you use?

14. What percentage of time during the lesson do you spend on the coursebook (or workbook) in an average week?

100% 80% 60% 40% 20% 0

15. What percentage of your course materials during the school year come from the coursebook/workbook/teacher's book?

100% 80% 60% 40% 20% 0

16. For a learner learning English in gimnazjum which aspects of English do you consider the most important. Rate each one 1–5.

Grammar 12345 Vocabulary 12345 Reading 12345 Writing 12345
Speaking 12345 Culture 12345 Pronunciation 12345 Other. What?

17. Are you familiar with the new English test to be introduced in 2009 in gimnazjum class 3?

Yes No

If you answered Yes to Q17 Have you done any practice tests with your learners? Yes/No

18. Do you think the new test will be difficult for your learners? Yes No

If you answered yes. Why?

19. What is your training?

mgr fil.ang. NKJO mgr + CAE + kurs kwal. ped.-met. Other? What?

20. How long have you been teaching English?

How long have you been teaching in gimnazjum?

You are nauczyciel: stażysta kontraktowy mianowany dyplomowany

Do you enjoy teaching English in gimnazjum? Yes No

If you answered no, why?

Appendix 2. Initial Washback Study: Learner Questionnaire (original Polish version)

Ankieta jest dla uczniów kl. III i ma na celu badanie wpływ egzaminu gimnazjalnego z j. angielskiego na proces uczenia się. Wszystkie dane są anonimowe. Bardzo dziękuje za współpracę!

1. *Co twoim zdaniem jest najważniejsze w uczeniu się j. angielskiego w gimnazjum? Proszę przyporządkować ocenę od 1 (najważniejsza) do 6 (najmniej ważna)*

a. gramatyka	d. pisanie	1. ☐	4. ☐
b. słownictwo	e. słuchanie	2. ☐	5. ☐
c. czytanie	f. mówienie	3. ☐	6. ☐

2. *Jaka była twoja ocena z j. angielskiego w ostatnim półroczu? Proszę zaznacz kółkiem* ***jedną z nich.***

6 5 4 3 2 1

3. *Jak oceniasz naukę j. angielskiego? Proszę zaznaczyć kółkiem jeden numer w podanej skali*

/_____/_____/_____/_____/_____/

(bardzo łatwy) 1 2 3 4 5 (bardzo trudny)

4. *Czy lubisz uczyć się j. angielskiego w twojej szkole? Proszę zaznaczyć kółkiem jeden numer na skali.*

/_____/_____/_____/_____/_____/

(bardzo lubię) 1 2 3 4 5 (nie lubię wcale)

5. *Na zajęciach j. angielskiego w szkole w tym roku szkolnym jak często ćwiczyliście czytanie? Proszę zaznacz jedną*

raz w tygodniu ☐ na każdej lekcji ☐ rzadko ☐ według podręcznika ☐

6. *Na zajęciach j. angielskiego w szkole w tym roku szkolnym jak często ćwiczyliście słuchanie nagrań? Proszę zaznacz jedną*

raz w tygodniu ☐ na każdej lekcji ☐ rzadko ☐ według podręcznika ☐

7. *Na zajęciach j. angielskiego w szkole w tym roku szkolnym jak często ćwiczyliście mówienie?*

raz w tygodniu ☐ na każdej lekcji ☐ rzadko ☐ według podręcznika ☐

8. *Na zajęciach j. angielskiego w szkole w tym roku szkolnym jak często ćwiczyliście pisanie?*

raz w tygodniu ☐ na każdej lekcji ☐ rzadko ☐ według podręcznika ☐

9. *Za co dostałeś/łaś oceny z j. angielskiego w tym roku szkolnym? Proszę zaznaczyć wszystkie, które dotyczą:*

ćwiczenia gramatyczne ☐	czytanie ☐	mówienie ☐	praca typu projekt ☐
pisanie wypracowania ☐	słuchanie ☐	test z słownictwem ☐	tłumaczenie ☐

10. *Jak oceniasz egzamin gimnazjalne z j. angielskiego?*

bardzo łatwy ☐ ok. ☐ wymagający ☐ trudny ☐ bardzo trudny ☐

11. *Która część egzaminu była dla ciebie najłatwiejsza? Proszę zaznacz jedną odpowiedź.*

czytanie ☐ słuchanie ☐ reagowanie językowe ☐

12. *Które część egzaminu był dla ciebie najtrudniejsza? Proszę zaznacz jedną odpowiedź.*

czytanie ☐ słuchanie ☐ reagowanie językowe ☐

13. *Jeśli opuściłeś/łaś salę egzaminacyjną przed 90 minutami, to z jakiego powodu?*

a. rozwiązywałem/łam wszystkie zadania i skończyłem/am wcześniej ☐
b. zrobiłem/łam tyle ile potrafiłem/łam, (nie wszystkie zadania testowe) i wyszedłem/wyszłam ☐
c. poddałem/łam się i wychodziłem/łam ☐

14. *Czy egzamin gimnazjalny z j. obcego jest dla ciebie ważny ? Proszę zaznaczyć kółkiem jeden numer na skalę*

/_____/_____/_____/_____/_____/
(bardzo ważny) 1 2 3 4 5 (nie ważny wcale)

15. *Jak myślisz, jaki będzie twój wynik z egzaminu gimnazjalnego z j. angielskiego? Proszę zaznaczyć kółkiem jeden numer w podanej skali.*

/_____/_____/_____/_____/_____/
(nie dobry) 1 2 3 4 5 (bardzo dobry)

16. *Jak przygotowywałeś/łaś się do egzaminu gimnazjalnego z j. angielskiego?*

brałem/łam korepetycje ☐ chodziłem/łam na kursy językowe ☐
z pomocą książki (np. repertorium) ☐ z pomocą materiałów z Internetu ☐
przerobiłem/łam testy próbne ☐ nic dodatkowego, tylko lekcje w szkole ☐
nie przygotowałem/łam się ☐

17. *Będąc po egzaminie, co doradziłbyś uczniom z kl. 2 gimnazjum w odniesieniu do przygotowania się do testu?*

czytaj dużo ☐ zrób ćwiczenia gramatyczne ☐ ucz się słownictwa ☐
ćwicz słuchanie ☐ ucz się systematyczne cały czas ☐
coś jeszcze? .....
.....

Appendix 3. Baseline Study: Teacher questionnaire

The aim of this questionnaire is to gather information about teaching and learning English in gimnazjum for research purposes. Data collected will be reported without revealing who you are, or the name of your school. Please contact me by e-mail if you would like to know more about the project. Thank you for your assistance! Melanie Ellis, NKJO Zabrze [email address].

1. Has your teaching in kl. 3 gimnazjum this school year changed in any way as a result of the new foreign language exam?

Yes No

2. If yes please tick all that apply (If you answered No to question 1 please go straight to 3.)

a) I did practice tests

b) I did more reading tasks

c) I did more listening tasks

e) I did picture description tasks

f) I changed my coursebook

g) I added a new book, especially to prepare for the test

h) I changed the type of reading exercises I usually do

i) I changed the type of listening exercises I usually do

j) We practiced matching examples of language to the Polish function word (e.g Przeprosić "I'm sorry")

Other. What ?.....

3. How would you describe the attitude of the majority of your pupils to the new exam? Tick ONE

a) It really motivated them to learn

b) They showed some interest in it

c) They were indifferent to it

d) They said it was not important to them

4. Did you notice any effect on the less able learners because of the new exam?

a) They asked for help more often

b) They appeared more engaged in lessons

c) They seemed to be trying harder

d) They asked more questions in class

e) Nothing seemed to change

5. How did the most able learners react because of the new exam?

they asked for help more often

they appeared more engaged in lessons

they seemed to be trying harder

they asked more questions in class
nothing seemed to change

6. What was your personal reaction when you saw the exam?

It was as I expected

It was different than I expected

7. What was your reaction to the level of the exam? Circle one number on the scale.
I felt the exam was…

/_____/_____/_____/_____/
1 2 3 4 5 (v. difficult)

8. What, in general, was your learners' reaction to the level of exam? Circle one number on the scale.

They felt the exam was…

/_____/_____/_____/_____/
1 2 3 4 5 (v. difficult)

9. Tick the ONE statement you most agree with:

a) The new foreign language exam in gimnazjum in no way affected my learners or my teaching

b) The new foreign language exam in gimnazjum had a negative effect on my learners and on my teaching

c) The new foreign language exam in gimnazjum had a positive effect on my learners and on my teaching

10. Will you make changes to your teaching next year as a result of the new exam?

Yes No Thank you for your time!

List of tables

List of graphs

List of figures

Contents

Rozprawa dr Melanie Ellis poświęcona jest roli, którą pełni i efektowi zwrotnemu egzaminu z języka obcego na zakończenie gimnazjum w polskim systemie edukacyjnym, życiu szkoły, pracy nauczyciela i ucznia, jak i dyrektora placówki. Podjęcie tego tematu [...] to działanie nowatorskie ze względu na rzadkość podejmowania tak szerokich badań teoretycznych i empirycznych w tej kwestii. Jest ono także niezwykle istotne edukacyjnie z uwagi na cele komunikacyjne, jakie stawiamy w nauce języka obcego. Z uwagi na obecne rekomendacje Rady Europy i Unii Europejskiej, które zalecają wprowadzenie minimum dwóch języków obcych w każdym systemie oświatowym, podjęcie tego tematu jest także działaniem przydatnym społecznie.

dr hab. Ewa Piechurska-Kuciel, prof. UO

Monografia poświęcona jest niezwykle istotnej tematyce, a mianowicie wpływowi, jaki mają na nauczanie, uczenie się oraz pracę szkół egzaminy państwowe z języka obcego. Należy w tym miejscu podkreślić, iż monografia ta jest m.in. rezultatem wieloletnich wnikliwych obserwacji, głębokich refleksji i przemyśleń nad tym, co się dzieje w polskim kontekście edukacyjnym, a także doświadczeń jej Autorki, zarówno jako dydaktyka, jak i badacza procesu dydaktycznego. Monografia dr Melanie Ellis niewątpliwie zainteresuje wszystkich uczestników procesu kształcenia językowego – nauczycieli języka angielskiego, dyrektorów szkół, lokalne władze oświatowe, uczniów, rodziców, osoby odpowiedzialne za politykę oświatową czy wreszcie studentów wyższych uczelni oraz nauczycieli akademickich kształcących nauczycieli języka angielskiego.

dr hab. Krystyna Droździał-Szelest, prof. UAM

Uniwersytet Pedagogiczny
im. Komisji Edukacji Narodowej
w Krakowie

Prace Monograficzne 820

ISSN 0239-6025
ISBN 978-83-8084-105-5