1 What Is Certification?

Abstract The first chapter looks at how English language certification has developed over the past decade, in the light of the massive growth of English as the language of choice for international communication, and the related needs for language assessment. It shows how certification received a boost from the publication of the CEFR at the beginning of the new millennium, offering the opportunity for all boards to validate their exams in line with a common, functional based, description of language competences. Although the major international boards use quite different approaches to implement these assessments (as we shall see in later chapters), we conclude by suggesting that they nonetheless share five common objectives, by attempting to produce tests which are authentic, valid, fair, secure, and which have a positive impact.

1.1 The Scope and Limitations of Language Certification in Assessing English

Language testing fulfils a variety of functions, and it can take many forms. At school, the purpose of a test might typically be to check what a student has learnt (or not learnt) at the end of a teaching unit (in a 'progress' test, set to monitor a list of objectives, or a 'diagnostic' test if it is intended to identify problems); in a language school or in higher education it might be used to decide which class or group the test taker should attend (an 'entrance' or 'placement' test). Prospective employers might need to appoint staff, or universities select students, on the basis of a test which provides evidence of an overall level of competence in the language (a 'proficiency' test). In an increasingly mobile, globalized society, governments may require immigrants to pass a 'citizenship' test, which will include an element of language competence, and which is intended to give some sort of indication of how the test taker has integrated (or will integrate) into their adopted country.

It should be clear from this incomplete list that tests come in all shapes and sizes, that they serve a vast range of purposes, and that they may be more or less important to the test taker. Of course, they are all intended to provide accurate information about the test taker. As a rule, though, as students progress through the educational system and into higher education, or the world of work, or international mobility, the stakes become higher; they have more to lose if they fail the test. Similarly, it is more crucial for the organization which requires the information the test is

intended to provide to be accurate and reliable. It is in this area of 'high stakes' tests that language certification thrives. 'Certification' usually refers to an independent assessment (independent, that is, of the test user, the organization requiring the assessment), which is valid, fair and reliable – three key concepts in testing which Messick (1998) puts together under the overarching umbrella term 'validity'. Crucially, the certifier is a professional organization which specializes in language assessment, and which typically may have developed from an educational institution (such as a university), or a government agency, (such as a cultural institution). Equally crucially, the certification has been through a process of validation, to guarantee the claims made by the certifier about the language competences the test is supposed to measure.

In recent years language certification, particularly for English, has developed enormously. This is largely due to the undisputed role which English now enjoys as the world's lingua franca. Whereas certification for other languages may be understandably bound to a (more or less overt) promotion of the culture, and cultural values, of the native speakers of that language, English certification has gradually moved away from a 'one language, one culture' approach to a policy of promoting the use of English for international communication, and a lifestyle which sees English occupying the free time, as well as the workplace, of today's globally mobile citizens. This is evident from the slogans used on the homepages of certifying agencies, or examining boards (as we shall refer to them in this study). Cambridge English declares that their certification can help test takers "achieve their goals for study, work and life";[1] IELTS is "the high-stakes test for study, migration or work";[2] while TOEFL entices candidates with the invitation to "pursue your dreams and go anywhere with the TOEFL test".[3]

The wealth of materials offered to potential candidates by the boards, through well-maintained websites, and traditional paper-based publicity, are an indication that certification also means big business. TOEFL and IELTS, the principal international tests for access to higher education, both number around 2 million test takers per year; IELTS is the current leader, having reached 2.7 million in 2015,[4] while for the same year TOEFL does not appear to have issued candidate numbers but instead claims that, since its inception in 1964, it has administered "thirty million [tests]

1 Cambridge http://www.cambridgeenglish.org/exams/ (2017-02-01).

2 https://www.ielts.org/ (2017-02-01).

3 http://www.toeflgoanywhere.org/toefl-practice (2017-02-01).

4 https://www.ielts.org/ielts-for-organisations/why-accept-ielts-scores (2017-02-01).

and counting".[5] The certification which has most candidates, however is, TOEIC (Test of English for International Communication), a test of English for the international workplace, administered, like TOEFL, by ETS (Educational Testing Services). TOEIC numbers more than five million tests annually,[6] 1.5 million of which are taken in Japan.

Certification comes with a cost, which varies depending on the type of test, the level, and where it is taken. The higher the level, the more expensive the test. Typically, at the time of writing, in Europe, a 'complete' test (i.e. one which assesses speaking, listening, reading and writing) at a higher intermediate level (B2) will cost around 150 euros. This, too, is a reminder of the 'high stakes' which certification usually implies: most candidates are in their teens or early twenties, the cost of the certification is a not inconsiderable sum, and it is viewed as a form of investment for the future.

The rapid growth in the demand for English language certification in Europe over the past decade has been fuelled, at least in part, by the consequences of the Bologna Process, which began with the 1999 Bologna Declaration and which aimed, among other things, to make European universities more competitive in the world market of higher education, by streamlining courses and ensuring that qualifications were mutually recognized by member states (Reinalda, Kulesza-Mietkowski 2005), but also that courses were accessible to an international student body, which in essence meant offering courses in English (Coleman 2006). This in turn led to universities setting minimal levels of proficiency in English for prospective students, to be certified by recognized examining boards. The implications of English Medium Instruction (EMI) for certification will be discussed fully in chapter 5. However, the need for certification in European universities extends far beyond EMI courses. Many universities now require a minimum certified level of English (B1 or B2) for *all* incoming students, whatever their course of study. This is because it is assumed that they will need English to carry out research, and possibly also to interact with students and staff on mobility programmes, for example with Erasmus.

But what does certification certify? And how do tests vary from one examining board to another, if they are all supposed to be assessing the same skills, as exemplified in the Common European Framework of Reference (CEFR)? The Framework, which by declared intent, provides a reference for the learning, teaching, and assessment of European languages, has rapidly established itself as an unavoidable standard setter for language policy makers, curriculum planners, and examining boards throughout

5 https://www.ets.org/toefl/institutions (2017-02-01).

6 https://www.ets.org/Media/Tests/TOEIC/pdf/TOEIC_sw_sample_tests.pdf (2017-02-01).

Europe and beyond. All certifications in Europe today are linked, in some way, to the Framework. They may be set at a stated level (for example, Cambridge First Certificate and Trinity College ISE 2 are both set at B2): to pass the exam, and obtain the certificate, means demonstrating the language skills which are a feature of that level; or they may not be set at any specific level (TOEFL and IELTS are examples) but the range of scores they produce can be interpreted in terms of the Framework. Thus an overall band of (say) seven at IELTS indicates a low C1 level, while in the TOEFL Internet-based test the same C1 level is indicated by a score of 95 or higher. Those boards which have developed a suite of exams directly from the Framework, such as Trinity College Integrated Skills in English, or calibrated an existing suite to the levels of the Framework, such as the Cambridge ESOL exams, are at pains to indicate the basis on which they make claims about levels.

Recognizing this need, the compilers of the Framework issued the manual *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (Figueras et al. 2005). The manual went through a lengthy piloting process, to which the major exam boards as well as the European testing organizations ALTE and EALTA contributed (Figueras & Noijons 2009), and which was intended, along with the more general aim of "competence building in the area of linking assessments to the CEFR", to increase the transparency of examination providers. For 'examination providers', read providers of certification. For descriptions of validation processes, see Papageorgiou (2007) for the Trinity ISE suite, and Khalifa and French (2008) for Cambridge ESOL. There can be no doubt that the major examining boards have emerged from the validation process with stronger products, certification which is the fruit of research, academically acceptable, and easier to market. At the same time, 'certification' which has not been through a rigid and transparent process of validation against the CEFR risks not being recognized as such, at least in Europe.

The Framework itself reflects current orthodoxy on what it means to know a language. It appeared on the crest of a communicative wave at the beginning of the new millennium, and as a result offers a description of language (any language, or at least, any European language) in functional communicative terms. Knowing a language, for the compilers of the Framework, is about doing things with language, a notion grounded in Austin (1975) and reinvented by the theorists of the communicative approach, such as Widdowson (1978), and Canale and Swain (1980). The Framework lists these functions (or rather, lists examples of functions) as *can do* statements, which are categorized into macro-functional areas of *reception*, *production* and *interaction*. Unsurprisingly, in a communicative approach to language use, *spoken interaction* turns out to provide the longest list of examples of language use.

Language certification has harnessed itself to the Framework in its approach to language competences, and, to some extent, in the use of terminology to describe those competences ('written production', 'spoken interaction', etc.). Thus speaking is no longer seen as a monolithic skill, such as the prepared monologue required in a pre-scientific age of testing (Spolksy 1976), but ranges across a range of competences; a good test needs to elicit samples of language which reflect this range of competences. How examining boards actually do this, however, varies greatly from one board to another. Thus assessment of speaking might be carried out through a range of formats, such as:

- paired speaker format, in which candidates speak to each other, are prompted by a facilitator and scored by a non-participating (but physically present) examiner (Cambridge ESOL).
- one-to-one speaking, in which candidates converse with a physically present examiner (Trinity GESE and ISE).
- one-to-one speaking, in which candidates speak with an interlocutor/facilitator; the exam is recorded and scored later (City and Guilds ESOL).
- responses to taped material delivered over the Internet (TOEFL, Pearson PTE Academic).

It seems reasonable to assume that individual candidates may be more comfortable with one format, and less so with others. Some students may prefer responding to prompts on the internet to interacting with a live examiner, and vice versa; in a paired interaction, some candidates will feel more comfortable talking to peers than to an examiner, while others may fear their own score may be compromised by the performance of by their partners, and so on. The paired interaction format has been the subject of extensive research (O'Sullivan 2002, Norton 2005, Brooks 2009), suggesting both advantages and problems, making it possible for boards to flag paired interaction as more 'authentic' (i.e. closer to real life language use) than individual interaction with the examiner, or, conversely, to promote their preference for individual interaction as more 'controlled', as in the following rationale for the traditional one-to-one format of the City and Guilds ESOL test of speaking: "Candidates are examined individually and converse only with the examiner and not another candidate, resulting in a controlled environment in which candidates can perform at their best".[7]

This type of variability, one might assume, will impinge on the candidate's performance, and compounded with other factors (such as test content, scorer reliability, conditions of administration), will lead to very

7 http://www.cityandguilds.gr/en/ESOLqualifications/oraltestsISESOL/Pages/isesol.aspx (2017-06-27).

different results for the same candidate, depending on the certification chosen. However, generic framework-related certifications are allegedly all testing the same things - competences described and exemplified in the CEFR - at the same level. Unsurprisingly, although individual boards have carried out a lot of research into their own tests, there is a dearth of comparative research, in which different tests and test results are compared. Such a study would be difficult and costly to organize, and is unlikely to be in the interests of the boards (who would see their exams branded as 'easier' or 'more difficult' as a result, and would need to realign their marketing strategies as a result). But for potential test-takers, it is important to realize that there are differences between tests, which are immediately perceptible in the test structures, and to choose carefully the one which is best suited to them.

To a large extent, these structural differences reflect an approach to assessment which the examining board may have nurtured over a long period of time, and which has become part of a house philosophy. The first Cambridge exam (Cambridge Proficiency) was delivered in 1913; all three candidates failed. This might have been due to the fact that the exam lasted for twelve hours, and required knowledge of French grammar (for a translation task), as well as phonetics and English literature. Later versions of the exam moved away from grammar translation to a more structural view of language (with a strong focus on sentence level syntax), and then, from the late seventies onward, to a "gradual shift […] away from structural approaches to language teaching towards approaches which involved using language as a means of communication" (Weir et al. 2013). Trinity College London began life assessing the performing arts, offering qualifications in music (from 1878) and then drama. It held its first exams in English as a foreign language in 1938, and since then has continued to promote a performance-based approach to assessment, with a the main focus on production, rather than testing knowledge of rules. The provider of City and Guilds ESOL certification began its business in the same year, 1878, and has a history of issuing vocationally and technically orientated certification, ranging from "Beauty Therapy to Business, Construction to Conservation and Digital Technology to Tourism".[8] General English language certification (the International ESOL suite) is only a part of this operation, a complement to the vocational and technical certification, and thus projecting an image of relevance to the world of work.

In contrast, the provider of TOEIC and TOEFL, Educational Testing Service (ETS), is an American institution founded in 1947, at the height of the structural linguistics era associated in the US with Leonard Bloomfield. This was reflected from the start in the tests, with the overarching concern

8 http://www.cityandguilds.com/qualifications-and-apprenticeships#fil=uk (2017-03-03).

for accuracy of measurement, and use of new technology, leading to the computer-based test (1998) which was then rapidly superseded by the Internet-based test (2005). TOEFL refers to itself as "the most researched" language test in the world, quoting more than 150 test reports of "rigorous research".[9]

More recent tests include Pearson Test of English (PTE Academic) and the Ireland-based Test of Interactive English (TIE). To distinguish itself from existing certification, the PTE draws attention, as the first feature of the test on the home page, and as the first point to be made in an introductory video, to the rapid reporting of test results: test takers will usually have their results in five days. This reflects not just the growth in need for certification, but the ever increasing need for university applicants, job seekers, and other test takers to provide evidence of their level in English at short notice. The Test of Interactive English requires candidates to carry out three pre-test preparatory tasks (reading a novel, following a news story, and carrying out an 'investigation'), which are then discussed with the examiner. These are supported by a 'logbook' which the candidate brings to the exam, a feature which owes something to the European Language Portfolio, developed in tandem with the CEFR, and whose aims are to motivate learners and to provide a record of linguistic and cultural skills acquired (Stoicheva, Hughes and Speitz 2009).

This brief overview of the approaches taken by six different boards should give an initial glimpse into the kind of variability, and the range of task types, a potential candidate may be faced with. At the same time, all boards share at least five common concerns, which are reflected in the frequent updates to tests, based on the development of new technologies, and research into language testing and assessment, and the nature of language itself. Updating tests are also of course a marketing strategy in an increasingly important global market. The five shared concerns are that tests should

1. assess 'real' or 'authentic' English
2. be recognized as valid
3. be fair and inclusive
4. be secure
5. have a positive impact

9 https://www.ets.org/c/mrm/ets00068/ (2017-03-03).

1.2 The Primary Shared Concerns of Examination Boards

1.2.1 Authenticity

Examining boards typically claim that their tests certify real, realistic, or real-life English. Cambridge English informs would-be test takers that "the Speaking test is taken face to face, with two candidates and two examiners. T]]]]his creates a **more realistic** and reliable measure of your ability to use English to communicate".[10] Trinity College London, in the foreword to the teacher's handbook for ISE, explains that "this integrated approach reflects how skills are used together in **real-life** situations".[11] IELTS, on a webpage for students, claims that its "content reflects **real-life** situations around the world".[12]

This emphasis on the authentic nature of tasks and language has a double purpose: it reassures teachers and students that the underlying approach is a communicative one, and it also sends a message to potential recognizing institutions that successful test takers will be able to use the language appropriately in the educational or work environment in which they may find themselves as a result of obtaining the certification. In the language assessment literature 'authenticity' refers to the degree to which a task reflects features of language use in real life, in what has come to be known as the target language use (TLU) domain (Bachman, Palmer 2010, 33). For Green (2014, 228) there are at least two types of authenticity: situational authenticity, "the fidelity with which real life tasks are reproduced in an assessment task", and interactional authenticity, "the extent to which an assessee engages the same mental processes in an assessment task as in target language use in the world beyond assessment". Bachman and Palmer (1996) consider authenticity to be a fundamental quality in a good test, along with *usefulness*, *reliability*, *construct validity* and *interactiveness*.

But authenticity and real life are not the same thing. Fulcher (2015) warns against the dangers of circular reasoning when making claims about test content, such that

> The test has *authentic* content. So: The test is valid because it measures *real-life* language use. (8)

10 http://www.cambridgeenglish.org/exams/first/exam-format (2017-06-27) (emphasis added).

11 From the forward to the Teachers Guide to ISE. URL http://www.trinitycollege.com/site/?id=3196 (2017-03-03) (emphasis added).

12 https://www.ielts.org/about-the-test/how-we-develop-the-test (2017-03-04) (emphasis added).

Take for example examiner-candidate interaction in speaking tasks, in which the candidate is required to react to an input, perhaps by showing surprise, giving advice, or expressing sympathy. This role-playing function is a feature of many speaking tests, and it is designed to sample a range of everyday communicative functions. However, depending on their cultural background, candidates may feel more or less inhibited about taking the initiative when interacting with an examiner (Fulcher, Reiter 2003) than they would in real life, when interacting with peers. But objections on the grounds of inauthenticity could be made throughout virtually any test, from multiple choice tests of receptive skills, to the content of reading and listening texts, to test administration, such as second hearings of listening texts, and time limits for written production. It would be difficult not to agree with Spolsky (1985) when he comments:

> Setting authenticity as a criterion raises important pragmatic and ethical questions in language testing. Lack of authenticity in the material or method used in a test weakens the generalizability of results. Any language test is by its very nature inauthentic, abnormal language behaviour, for the task is not to give so much as to display knowledge. With examinees who do not know or who are unwilling to play by the rules of the game the results of formal tests will not be an accurate and valid account of their knowledge. (31)

What is true of any language test is potentially more so of certification, especially generic certification which, intended for an international market, has no one clearly defined TLU domain. In-house tests in schools or universities are developed with a clear test taker profile in mind; progress tests produced by publishers for their course books have an explicit syllabus to check. But free-standing international certification needs to look elsewhere for its certainties, and it does so by investing in test validity and reliability, through research and development, through pilot studies and test taker feedback, to guarantee responsible management of high-stakes testing.

1.2.2 Validity and Validation

Traditionally, a test has validity if it measures what it is supposed to test. Thus a test of grammar does not necessarily give any information about (say) a learner's ability to speak, even though exam results may suggest a correlation between the two traits (Fulcher 2003, 203). Similarly, a test of listening which requires test takers to read the questions is more than just a test of listening. The notion of validity used to be viewed (Cronbach, Meehl 1955) from various perspectives: content validity, concurrent valid-

ity, construct validity, and predictive validity. To this we should add the popular notion of 'face validity', which concerns test takers' (and test users') perceptions that a test feels right, because it is set at the right level, or appears to test the right things. Content validity reflects the degree to which test samples from the stated target language use domain of the test, while concurrent validity is achieved when a second, alternative measurement from an independent source (such as another test, or an expert opinion) confirms the test result. Predictive validity refers to the ability of the test to predict a test taker performance in some future scenario; for example, TOEFL and IELTS both claim predictive validity in that they are intended to demonstrate to potential higher education institutions how well test takers will be able to operate successfully in English in that institution. Interestingly, their certificates are date stamped; the predictive validity is guaranteed for two years only. The central, underlying validity, however, is construct validity, a 'construct' being the underlying skill, or skills, which a test intends to measure. This may be an abstract skill such as 'reading', or a hypothetical trait or sub-skill such as 'reading for gist', which is operationalized through the test tasks.

Today, largely due to the work of Messick (1975, 1989), a researcher for Educational Testing Services (the organization responsible for TOEFL and TOEIC), construct validity has come to be seen as the core validity of a test, a super-ordinate notion or underlying framework which embraces all other features of validity and extends to notions of fairness and reliability. Messick's much quoted definition of validity has become the consensus view on validity (see D'Este 2012, 65 and Fulcher 2015, 107):

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

Validity and validation are of course closely connected: validation is defined as "the collection of evidence which supports the validity of the inferences that may be made on the basis of assessment results" (Green 2014, 242); in other words, the confirmation of the meaning of test scores. Thus, when examining boards make claims about their tests, such as the level of the CEFR at which they are set, or the underlying skills which they are supposed to measure, then these claims need to be validated. Most examining boards devote considerable human and financial resources to these procedures.

1.2.3 Fairness

All tests are meant to discriminate. For example, a test set at B2 of the CEFR needs to separate those test takers who can perform at the level of B2 or above from those who can't. But the test must not discriminate for the wrong reasons, by adopting (consciously or unconsciously) criteria unrelated to the language skills ostensibly being assessed. If it did, it would be unfair. Most investigations of test fairness see it as an aspect of test validity (Messick 1989, as noted above, but see also Kunnan 2000, and Xi 2010), although it clearly has implications for test reliability, too: a test which is unfair may be consistent in the (unfair) measurements it yields, but it will not provide a reliable assessment of the skill(s) it claims to measure.

In *Language testing: the social dimension* (2006) McNamara and Roever approach the issue of fairness from Messick's distinction between construct under-representation – when a test fails to assess the test taker as completely as it should – and construct-irrelevant variance, when a difference in test scores between candidates is not due to a difference in the skill(s) being measured, but to some other factor(s). Construct-irrelevant variance is thus one aspect of unfairness; when it is systematically built into a test, it becomes test bias, and systematically harms one group of test takers when compared with another. This might happen when a test appeals (say) to the socio-cultural knowledge of one group, making it easier for that group to pass the test, when compared to another group; when in fact this knowledge is not part of the test construct.

As the title of the volume suggests, McNamara and Roever take a wide view of fairness issues, and are particularly concerned with the use made of test results:

> biased tests harm all stakeholders because students might get exempted from language programs although they would benefit from them, others do not get admitted to a program in which they would excel, universities or employers reject perfectly qualified applicants and accept less qualified ones, and society is deprived of potentially excellent doctors, lawyers, language teachers, or electricians and must make do with mediocre ones. (McNamara, Roever 2006, 82)

This wider view is of paramount importance to examining boards, who have to tread very carefully to avoid producing tests which might have cultural bias, and it is enshrined in the principles of good practice of testing associations, such as ALTE (Association of Language Testers in Europe) and EALTA (European Association for Language Testing and Assessment), and individual examining boards. Fairness is the first quality described in the 2016 Cam-

bridge English *Principles of Good Practice* document.[13]

The notion of fairness extends to that of inclusion of candidates with disabilities, such as the visually impaired or hard of hearing, or with learning difficulties such as dyslexia. All examining boards issue policy statements about how they will attempt to accommodate special needs candidates; not to do so would expose them to possible legal action (in the UK, for example, on the basis of the 2010 Equality Act). Typically, candidates with reading or speech difficulties will be allowed extra time, blind candidates may be assigned readers, and the hard of hearing may be able to lipread texts which are used in listening tests, and which are read to them by a 'live' reader. These accommodations require the deployment of extra resources, and so need to be arranged well in advance of the exam, and disabilities will need to be confirmed by medical records.

Conversely, examining boards are at pains to stress that accommodations do not lead to an unfair positive discrimination, by treating some candidates more leniently than others. As the Cambridge exams website puts it, "Once special arrangements have been made, candidates with hearing difficulties or speech difficulties are assessed in exactly the same way as other candidates; they are not marked 'more leniently' because they have difficulty hearing or speaking".[14]

Another aspect of fairness concerns transparency, which essentially refers to the amount of information the examination board releases to test takers about the way in which they will be (or have been) assessed. This can take the form of specimen papers, sample responses, and rationales behind scores, which boards post on their websites, and which we will take a close look at in the next chapter. More general help and advice to candidates, in the form of worksheets and videos, is also usually available. However, it is unlikely that any examining board is completely transparent about the way in which it trains its raters, sets standards, weeds out underperforming items, or adjusts scores when it realizes that the level was not exactly the declared target level, i.e. when a task or an item, or a whole test, turns out to be easier or more difficult than anticipated.

For the candidate, however, it is probably true to say that the most useful information that needs to be provided by the examining board concerns the structure of the test itself; understanding how the test is structured, and how much time the candidate will have for each section, is essential, given the elaborate structure of most tests. To be unaware of how the exam works will lead to valuable time being lost as candidates try to figure out what they are supposed to be doing.

13 http://www.cambridgeenglish.org/images/22695-principles-of-good-practice.pdf (2017-06-27).

14 Giudance Notes for Special Requirements Speaking Tests. URL http://www.cambridgreenglish.org/help/special-requirements/ (2017-02-10).

1.2.4 Security

In 2014 a documentary for the BBC TV programme *Panorama* revealed large scale cheating at an ETS test centre in London. Candidates had paid three times the real test fee to be guaranteed a successful result in the TOEIC exam, which at the time was one of the Home Office recommended exams for non EU students wishing to extend their visas and remain in the UK. The scam was comparatively simple: all candidates had to do was pay the inflated fee, turn up at the test centre, have their photograph taken to prove that they were there on the day of the exam, and then let a proxy sit the speaking and writing parts of the exam for them. They then had to return to the test centre a week later to sit the reading and listening exam. This was handled differently; candidates were told which answers to select (all items were multiple choice questions) by an invigilator who simply read out the answers to all the people in the room.

ETS unsurprisingly claimed they were unaware of the scam, as did the school whose premises were being used for the test, administered by a bogus 'education agency'. The scandal had major repercussions. The Home Office deported 48,000 students who they believed (on slim evidence) might also have cheated in previous versions of the exam; and ETS were removed from the list of test providers for citizenship and visa extensions.

With high stakes tests, such as those relating to citizenship or job applications, there will always be a security risk, and the scam uncovered by Panorama was not the first example of organized cheating. The size of the operation, however, has forced examining boards to review their arrangements to guarantee test security, the ways in which test centres are appointed and managed, how invigilators are selected, and how candidates' identities are checked. In the current document giving information about the TOEFL iBT (like TOEIC, an ETS test), we read:

> You must present valid and acceptable primary ID. [...] Verification of identity at the test centre may also include
> – thumbprinting
> – photographing/videotaping
> – signature comparison
> – electronic detection scanning devices such as hand-held metal detectors/wands
> – biometric voice identification
> – other forms of electronic confirmation

If you refuse to present ID or to have your ID verified, you will not be permitted to take the test and your test fee will not be refunded.[15]

TOEFL also publish[16] a list of clothes items which candidates may expect to be examined before the beginning of the test, ranging from hats, scarves and headbands, to jackets, cuff links and jewellery.

In 2015 the UK government introduced the SELT (Secure English Language Test), not a new test, but rather a list of approved test centres where tests for citizenship and leave to remain had to be taken; the only approved tests being Trinity College ISE (Integrated Skills in English) and GESE (Graded Exams in Spoken English) and IELTS. Unlike a test delivered entirely on line, such as TOEFL, the Trinity exams and IELTS both include face to face exams, with an oral examiner appointed directly by the examining board, which probably makes them intrinsically safer than an online test administered by local agents.

Nonetheless, with the rapid need for reliable high stakes tests, examining boards have become more security conscious and this has sometimes led to structural changes to exam. In 2015 Trinity College updated its existing four skills exam (ISE), introducing, among other things, a multitext reading to writing activity, and a free-standing listening administered during the oral interview. It also got rid of a portfolio component, which for ten years had been a characteristic feature of the assessment; candidates produced three short written texts in their own time, and discussed them in the oral interview. The portfolios were scored and contributed to the overall assessment of writing. Justifying the disappearance of the portfolio (which had proved popular with test takers no doubt because it be written outside the stressful context of an exam) Trinity College suggested that it "could be more usefully harnessed as a teaching and classroom support tool",[17] and consequently developed a 'portfolio toolkit' resource for teachers. It seems clear, however, that security conditions also contributed to the decision. Of course, an exam with a take-home writing component could not have been chosen as a UK Secure English Language Test.

1.2.5 Impact

'Impact' is usually taken to refer to the effect (whether positive or negative) of assessments on educational systems and society as a whole. It is an extension of the notion of "washback", the effect of assessments on

15 2016-17 TOEFL ibT Test Registration Bulletin, 12. URL `https://www.ets.org/toefl/ibt/about/bulletin` (2017-02-10).

16 *What to Expect*. URL `https://www.ets.org/toefl/ibt/test_day/expect/` (2017-02-12).

17 ISE FAQs page. URL `http://www.trinitycollege.com/site/?id=3323` (2017-10-02).

teaching and learning. Beneficial washback (or "backwash"[18]) began to be recognized in the 1980s as a fundamental test quality, alongside validity, reliability and practicality, and perhaps the most important quality from a teaching perspective (Hughes 1989). In the 1990s, especially in the wake of Alderson and Wall (1993), empirical research into washback developed rapidly, with some researchers (e.g. Messick 1996) seeing washback as fitting a unified theory of test validity.

In an educational context it is easy to find instances of 'harmful' or negative washback: for example, when 'teaching to the test' implies leaving aside useful skills which might be part of a curriculum document (such as speaking) but which are not included in the test, perhaps for organisational reasons. Similarly, a test which is not perceived to be fair by test takers is likely to impact negatively on their attitude to learning the language. Tests need to be motivating for learners, and it is in the interests of teachers to produce motivating institutional tests, since they will have to live with the consequences.

This is not the case for external certification and examining boards who have no direct knowledge of individual test takers. When an individual fails to achieve a necessary grade on a high stakes test, and has to retake the test, there is usually not much comfort to be gleaned on the boards' websites. Take for example the following advice on the IELTS website,[19] such as the following note, which, after explaining that it is possible to retake the test "as soon as you feel ready to do so" continues

> Before applying for another test, take a moment to consider your options. Your score is unlikely to increase unless you make a significant effort to improve your English.

This claim does not take into account variability associated with conditions of administration, and, crucially, the test takers themselves, who may have had to travel a long distance to get to the test centre, and for whom there may be a high discomfort factor which could affect performance. Unsurprisingly, it seems that boards do not publish statistics about test retakes, but it may be that, in spite of the advice quoted above, retakes sometimes give quite different results for the same candidate.

When it comes to the wider picture of impact, examining boards are careful to project an image which suggests, firstly, that they are aware of the ethical dimension of possible social uses of language assessments, and secondly, that their tests can promote international mobility and make a

18 Hughes uses the term "backwash" because, he says, he can find the term in a dictionary; whereas "washback" is not present in most dictionaries.

19 `https://www.ielts.org/book-a-test/resitting-the-test` (2017-02-11).

contribution to successful international communication. But such claims need to be evidence-based. Writing for Cambridge ESOL, Taylor (2005) puts it thus:

> Today tests are increasingly used for 'high-stakes' gate-keeping and policy-making purposes, as well as to provide targets for and indicators of change; it is therefore even more important for test producers to provide appropriate evidence that the social consequences of using their tests in such ways is beneficial rather than detrimental. Until fairly recently, claims and assertions about the nature and extent of a test's impact were largely based upon impression and assumption. It was relatively simple for producers to claim positive washback for their own tests, or for users to criticise tests on the grounds of negative washback; but it was also too easy for both sets of assertions to go unchallenged. Impact research – such as that conducted by our own organisation – reflects the growing importance of evidence-based approaches to education and assessment which enable policy and practice to be justified in terms of sound evidence about their likely effects.

In recent years major boards have published a number of impact studies, such as Merrifield (2016), which looks at the use made of IELTS exam results by professional organizations, or Khalifa and Vidakovic (2014) which presents impact studies for Cambridge exams in mostly educational settings. Evidence of positive impact can be used for promotional purposes, is made available on websites, and filters through to slogans which highlight the global acceptability of certifications.

A test which can demonstrate a positive impact, in an unstable and increasingly mobile world, is a test which is attentive to change and the opportunities afforded by the development of English as the world's lingua franca. Some tests need to be highly specialized, to capture the skills needed for communication in specific environments or for specific professions, ranging from air traffic control, to maritime communication, to legal professions, in which wrong use of a single word may have devastating consequences. This book is not concerned with this kind of ESP test. Rather, it will take a close look at the biggest and most well-established tests of English for academic purposes especially in the context of Europe, and how they may need to evolve to maintain a high degree of validity and a positive impact on the development of English language as a means of international communication.