10 Testing and teaching

1. The things we do for tests

Much of this book has been concerned with how teachers and testers can develop language tests in professional ways. This chapter looks at the effects that tests can have on classrooms. Teachers, as we have observed, have to respond to the demands made by testing regimes and students' desires to pass tests. It is therefore about evaluating the impact that test use may have on teaching and learning, in the broadest sense. The effects of the use of language tests are the measure of the meaning of the test in practice. If the test has been well designed, with its purpose and effect in mind, we might expect to see many positive practical effects for most stakeholders.

We first discuss washback and related research, and the practice of aligning tests and the curriculum to content standards. We then consider what is probably the most widespread responsibility of teachers: preparing students to take externally mandated tests. We also look at how to select tests when resources aren't available to develop them in house.

2. Washback

The present concern with washback began with Messick's (1989: 20) introduction of the notion of consequences into his definition of validity. His conception of validity incorporated both the values that the test endorsed, and the impact that the use of the test had on individuals and institutions. Messick (1996: 241) says that 'washback refers to the extent to which the introduction and use of a test influences language teachers and learners to do things that they would not otherwise do that promote or inhibit language learning'. Alderson and Wall (1993) set out a number of questions that they referred to as 'washback hypotheses', many of which have been subsequently investigated. The most important of these are listed below. A test will influence:

- what teachers teach
- how teachers teach
- what learners learn
- how learners learn
- the rate and sequence of teaching
- the rate and sequence of learning
- attitudes to the content, method, etc. of teaching and learning.

While it seems obvious that washback does exist, it is not quite so clear how it works. For example, it does not appear that it is systematic (Tsagari, 2009). It affects some learners more than others, and some teachers more than others. Nor is washback always negative. Washback is mediated by many other factors, such as the nature of the curriculum, the training background of teachers, the culture of an institution, and the quality of the support and resources available in the teaching context (Watanabe, 2004a). Washback studies have therefore not been able to pick any particular classroom teaching or learning activity and unequivocally state that it is 'caused' by the test (Alderson and Hamp-Lyons, 1996; Green, 2007). Evidence from major studies into the introduction of new tests in education systems with the intention of introducing change generally shows that things do not work as intended; there is no simple relationship between the use of a test and its effects (Wall and Alderson, 1993, 1996). Indeed, with reference to their study of the introduction of a new test into the Sri Lankan educational system, Wall and Alderson (1996: 219) conclude that:

> *the exam has had impact on the content of the teaching in that teachers are anxious to cover those parts of the textbook which they feel are most likely to be tested. This means that listening and speaking are not receiving the attention they should receive, because of the attention that teachers feel they must pay to reading. There is no indication that the exam is affecting the methodology of the classroom or that teachers have yet understood or been able to implement the methodology of the text books.*

Although empirical studies show that the deliberate use of high-stakes tests to modify teacher behaviours is usually not effective, even over time (Qi, 2004), washback remains an important and highly emotive subject. Evidence or not, many teachers feel the pressure of the test. Sometimes, this pressure is internal to an institution, where a test may become entrenched. Sometimes the administration or groups of teachers who are comfortable with the testing status quo accept a situation that places what seems like a straitjacket on newer members of staff. This has led to a trend to look at washback in the small cultures of institutions, rather than those of national or international tests. For example, Nicole (2008) studied the impact of a local test on teaching and learning in Zurich. Using survey and interview techniques to gather teachers' views of the impact of the test, and classroom observations against which to confirm what she was told, Nicole discovered that the test appeared to encourage greater coverage of skills and content. She also reported evidence of an improvement in teaching methodology. In this case, the participatory teacher-researcher, working with her own colleagues, produced evidence to suggest that the test appeared to have positive washback on teaching.

Teachers can and should study washback in their own professional contexts. Studies of the kind conducted by Nicole can provide washback evidence to militate for change, or to support positive developments. This is professionally valuable at a local level, and can be empowering for teachers in shaping their own testing world. There is little point in spending all the time and resources available to produce an excellent test, following the practical advice in books like this, if the effect on those who matter is not evaluated.

Researching washback requires careful thought, especially if you are investigating washback within a context with which you are extremely familiar, like Nicole. This is mainly because it is very easy not to notice key features of the context that are important to arrive at an informed interpretation. Familiarity can desensitise us to important features of the context (Watanabe, 2004b: 25). The first requirement is, therefore, an attempt to distance yourself from the familiar; to make it unfamiliar, and to find it curious. The next step is to consider the scale of the investigation. Should it be limited to the context of a particular classroom, a particular school, a school district, or an educational system? This is important, because it determines not only the size of any sample you will need to study, but also the kinds of data that might be collected. For example, if an educational system is to be evaluated it is reasonable to look at press reports, curriculum documents, teacher training systems, and so on. This is a matter of focus. Descriptive work is important. This will include the test itself, the institution(s) in which the washback study is taking place, the participants, the intended purpose of the test, and all facets of the teaching environment such as materials, syllabus and learning objectives.

It is also important to describe those aspects of washback in which you are interested. This is similar to construct definition, as discussed in Chapter 4, but it is guided towards answering this question: what would washback look like in my context? This was the question asked by Wall and Alderson (1996: 197–201) when they began looking at changes to the examination system in Sri Lanka. As the introduction of a new test was meant to be 'a lever for change', one of the expected effects was an increase in the time devoted to speaking and listening in the classroom. This therefore became a key focus of their study.

Next, it is important to consider what kind of data would provide the evidence you need to decide whether the washback is working as intended (see Wall, 2005). In the case of the Sri Lankan project, classroom observations in which observers recorded the percentage of time devoted to different skills would generate the evidence needed to make an evaluation. Typical data sources, depending upon the questions asked, include surveys of teachers, supervisors, managers or policy makers; these are often followed up by interviews with selected participants in order to get more fine-grained views that cannot be elicited in questionnaires. The same may be undertaken with language learners. Classroom observations are frequently an important part of washback studies, as what teachers and learners report happens in classrooms does not always tally with reality. Before conducting classroom observations it is important to decide what is going to be observed. Checklists need to be drawn up in advance. If audio recording is an option, pilot studies can be conducted to see whether or not check-lists need to be altered or expanded with new categories before a main study takes place. Follow-up interviews might be conducted with teachers or learners, either with or without playing back a recording, to ask them to reflect on how particular classroom activities are related to their understanding of the test.

Documentary evidence is always important. Teacher-prepared lesson plans, teacher-prepared activities, selected textbooks and other resources, can be analysed to detect

washback. Focus groups may be used to discover whether materials are deliberately developed to teach the skills that teachers think the test measures, or how they adapt published materials to meet their own teaching goals.

Two issues concerning data need to be raised at this point. Firstly, it is always unwise to rely entirely upon one source of evidence in washback studies. Using multiple sources of evidence is termed *triangulation*. When data can be presented from multiple sources and used in a coherent interpretative argument, it gives more weight to the usefulness of the emerging understanding of how washback operates in the context. Secondly, if at all possible, data should be collected before the introduction of a new test, and again after its introduction, using the same methods, and ideally the same teachers in the same institutions. It is not always clear that a particular practice might be followed irrespective of whether a test is used or not, even if teachers claim that it is. It could be that these practices are part of the culture of the institution, or are used because of the content of teacher training programmes. If data is collected before the introduction of a test – in a *baseline study* – it is possible to compare the results of the baseline study with those of a post-introduction study to see what changes come about. Of course, in complex environments like educational systems and schools, it is always possible that changes occur for other reasons. Perhaps there is staff turnover, a syllabus is changed, a new textbook is introduced. But baseline studies do provide another anchor that allows us to try to see which practices may be caused by the use of a test, and which are incidental to it.

This was the practice adopted in what is perhaps the largest language testing washback study conducted to date, which is reported in Wall and Horák (2006, 2008). In the first study Wall and Horák (2006: 3) note that one of the intended effects of the introduction of the TOEFL iBT over 2005–06 was a positive impact upon classroom teaching and learning. The previous pencil-and-paper test had been criticised for encouraging the use of multiple-choice items in classroom teaching, and a neglect of speaking and writing. In order to study whether the test designers' intentions were realised, they conducted a baseline study in a number of institutions in Eastern and Central Europe. Description is central to their study, as they argue that the impact from innovation is determined 'by the interaction of features in the antecedent situation (the context into which the innovation is being introduced) and a number of factors that work together (or against one another) during the process period (the time that the innovation is introduced and being tried out by the users)' (2006: 4–5). Only by undertaking this description can the consequences – the adoption, adaption or rejection of an innovation – be understood. They described existing test preparation classes (materials, methodology and assessment) and the institutional context (policies, practices, resourcing), in preparation for a second study to take place after the innovation (the new test) had been introduced. Wall and Horák also studied the washback intentions of the test designers.

One example in relation to writing is the following:

Writing

Innovation: introduction of multiple writing tasks that include both independent and content-dependent tasks.

Intended effect: move beyond the single independent essay model to a writing model that is more reflective of writing in an academic environment.

Washback intention: specific to the teaching of writing, and positive.

With such specific intentions it is possible to use interviews, observations and analysis of writing materials to chart the impact of the innovation in the classroom. Using a similar approach in a specific institution would allow teachers to consider how they wish to change teaching, and to co-ordinate this with changes in assessment practices.

As part of the data collection it is important to design or adapt appropriate questionnaires or observation schedules. It is not within the scope of this book to discuss questionnaire design, for which the reader is directed to the relevant literature (Brown, 2001; Dörnyei, 2003). However, Wall and Horák (2006, 2008) reproduce all their survey and observation schedules in the appendices to their reports, which the reader is encouraged to study. With regard to collecting data on the changes to the writing test that we considered above, Wall and Horák (2006: 173) used the following classroom observation schedule when visiting writing classes during the baseline study, and to collect data in the follow-up study after the new test had been introduced.

This observation schedule is designed to direct the attention of classroom observers to the key lesson features that are related to an existing test, and which are expected to change when the new test is introduced. Every instrument needs to be sensitive to the context. However, these schedules can be adapted to suit new contexts.

Start time: End time: S work mode: I P G C Medium: Com PP Language: LI L Atmosphere: Skill focus: Int Sing		
WRITING	✓	
Activities: Notes:		
		Generating ideas
		Organizing ideas
		Developing ideas
		Supporting ideas with examples/evidence
		Selecting appropriate vocabulary
		Developing sentence structure
		Writing essays
		Writing an essay–no time limit
		Writing an essay in time limit
		Writing essays–no word limits
		Writing essays with word limits (NT)
		Writing essay based on a listening (NT)
		Writing an essay based on reading (NT)
		Organizing ideas from listening/ reading before writing (NT)
		Writing on topics from ETS pool
		Writing on topics selected by teacher
		Writing on topics selected by student(s)
		Examining ETS scoring scale
		Synthesizing data from 2 or more texts

Fig. 10.1. An observation schedule for writing classes

Investigating washback is primarily qualitative research which involves the careful interpretation of data generated from a number of methodologies. They may also be quite time-consuming to conduct. Nevertheless, washback is critical to the evaluation of the extent to which testing policies have been successful in terms of their intended effect.

3. Washback and content alignment

It has long been argued that tests should be totally independent of any method of instruction, or the content of instruction (Latham, 1877: 46). While tests may be aligned to standards (Chapter 8) for reporting purposes, the traditional view has been that success in the test should not be correlated with attending any particular educational institution or programme. There is merit in this position. It holds that the test taps a construct, but the test does not dictate to teachers or learners how this construct is acquired. The role of the teacher is to plan a curriculum that enables the acquisition of the tested abilities. The methodologies and materials may therefore look very different depending upon the people and institutions, even if the learners ultimately take the same test. Some even go so far as to claim that any attempt to abandon this 'separation of power' is to undermine the validity of a test on the grounds that test content and classroom teaching are confounded (Haladyna, Nolen and Haas, 1991). The 'separationists' would argue that classroom teaching should take priority and the test is an independent measure of the achievement of the learners.

The alternative point of view is that tests can drive curriculum change. This argument holds that, when test content and instructional content are closely aligned, there is an opportunity to claim that teachers are covering the necessary material to achieve desired educational goals. This is frequently referred to as curriculum alignment. Koretz and Hamilton (2006: 555) describe this alignment as having taken place when 'the knowledge, skills and other constructs measured by the tests will be consistent with those specified in the [content] standards'. Of course, it has always been the case that test designers should take content into account. Defining the universe of possible content and then sampling from that content is part of defining the test purpose and the domains to which the test scores are relevant. What is different in curriculum alignment is the attempt to completely specify all the content that learners are supposed to cover, and replicate this as far as possible in the assessments. The intention is to use the test to control curriculum, removing it from the professional control of teachers. The 'integrationists' argue that the closer the alignment between test and curriculum, the better the teaching, and the more valid the measurement. For example, Gottlieb (2006: 36) argues:

> *In today's classroom, standards are the cornerstone for accountability. Content standards are the starting point, anchor, and reference for teaching and learning. English language proficiency standards lead to instruction and assessment of English language proficiency, and academic content standards are geared to instruction and assessment of academic achievement.*

The 'integrationists', as I have called them, see the process of aligning content standards, curriculum and tests as the means to measure educational success, and in the process to make teachers and institutions accountable for the learning outcomes. Teachers are therefore increasingly being asked not only to align their tests to performance standards (Chapter 8), but also align both their curriculum and their tests to content standards.

The first point to make about this approach for language testing is that, while 'content' in mathematics or history can contain specific topics, in language they generally do not. Rather, they resemble fairly complex performance standards – indeed, the example we will consider is actually called 'performance standards' rather than 'content standards'. In language standards, the content arranged within a performance level is frequently described in terms of tasks that are considered to be of the right level of difficulty. Thus, for example, the McREL Content Knowledge Standards (2009) contains 24 topics, each with sub-standards set out under five general standards, with entries by grade level. If this complexity were not enough, each sub-standard is broken down into 'benchmarks' – or sub-skills within sub-standards. The matrix created by this classification system is extremely large. To illustrate, within the standard 'writing' for topic 'writing for audience and purpose', at the level of grades 9–12, benchmarks 2 and 3 read:

> *Drafting and revising: uses a variety of strategies to draft and revise written work (e.g., highlights individual voice; rethinks content, organization, and style; checks accuracy and depth of information; redrafts for readability and needs of readers; reviews writing to ensure that content and linguistic structures are consistent with purpose.*
>
> *Editing and publishing: Uses a variety of strategies to edit and publish written work e.g., uses a checklist to guide proofreading; edits for grammar, punctuation, capitalization, and spelling at a developmentally appropriate level; refines selected pieces to publish for general and specific audiences; uses available technology, such as publishing software or graphics programs, to publish written work.*

Teachers are required to create a curriculum that covers all the necessary skills and abilities within each standard at the class level, and construct tests that will assess whether the learners have met the requirements for that level. Alternatively, the creation of the tests is outsourced to a specialist testing agency.

One of the most widely used sets of standards is produced by the WIDA consortium in the United States (WIDA, 2007). Produced for both Spanish and English language learners (and other subject areas), WIDA creates the standards used by many states to comply with accountability legislation. All documentation relating to the consortium and its standards can be downloaded at its website (www.wida.us). The content standards are organised within two frameworks: formative (the process of learning) and summative (the outcomes of learning). They reflect both social and academic aspects of language learning in schools according to sets of proficiency standards (see Chapter 8) for subject areas, and are presented in hierarchical clusters for school grades Pre-K, Grades 1–2, 3–5, 6–8 and 9–12.

As the entire set of standards may be downloaded from the website, we will select a small sample for discussion. This is taken from the academic area of language arts, and relates to the teaching and assessment of writing in grades 9–12. There are further standards for other academic and social areas, and for other skills (listening, reading and writing).

	Level 1: Entering	**Level 2: Beginning**	**Level 3: Developing**	**Level 4: Expanding**	**Level 5: Bridging**
Example Genre: Critical Commentary	Reproduce comments on various topics from visually supported sentences from newspapers or websites	Produce comments on various topics from visually supported paragraphs from newspapers or websites	Summarize critical commentaries from visually supported newspaper, website or magazine articles	Respond to critical commentaries by offering claims and counter-claims from visually supported newspaper, website or magazine articles	Provide critical commentary commensurate with proficient peers on a wide range of topics and sources
Example topic: Note taking	Take notes on key symbols, words of phrases from visuals pertaining to discussions	List key phrases or sentences from discussions and models (e.g. on the board or from overhead projector)	Produce sentence outlines from discussions, lectures or readings	Summarize notes from lectures or readings in paragraph form	Produce essays based on notes from lectures or readings
Example Topic: Conventions and Mechanics	Copy key points about language learning (e.g. use of capital letters for days of week and months of year) and check with a partner	Check use of newly acquired language (e.g. through spell or grammar check or dictionaries) and share with a partner	Reflect on use of newly acquired language or language patterns (e.g. through self-assessment checklists and share with a partner)	Revise of rephrase written language based on feedback from teachers, peers and rubrics	Expand, elaborate and correct written language as directed

Table 10.1 Standards for formative writing, language arts, grades 9–12 (WIDA, 2007: 59)

	Level 1: Entering	Level 2: Beginning	Level 3: Developing	Level 4: Expanding	Level 5: Bridging
Example genre: Critical commentary	Reproduce critical statements on various topics from illustrated models or outlines	Produce critical comments on various topics from illustrated models or outlines	Summarize critical commentaries on issues from illustrated models or outlines	Respond to critical commentaries by offering claims and counter-claims on a range of issues from illustrated models or outlines	Provide critical commentary on a wide range of issues commensurate with proficient peers
Example topic: Literal and figurative language	Produce literal words or phrases from illustrations or cartoons and word/phrase banks	Express ideas using literal language from illustrations or cartoons and word/phrase banks	Use examples of literal and figurative language in context from illustrations or cartoons and word/phrase banks	Elaborate on examples of literal and figurative language with or without illustrations	Compose narratives using literal and figurative language

Table 10.2 Standards for summative writing, language arts, grades 9–12 (WIDA, 2007: 61)

Even at a particular grade level (9–12) it is assumed that there are five levels of progression (the actual performance standards), and in the left-hand column we are given example genres and topics, with particular abilities in cells. Once again, the entire matrix to cover all content, by age range and ability level, is mind-blowingly large.

One of the problems with language content standards is that they are not targeted at performance in particular domains (for example, whether there is sufficient language to act as a tour guide in a defined context), but whether test takers have acquired 'language' without reference to any domain at all, across all skills, genres, domains and contexts. This is cross-referenced with other skills, such as use of software for publishing and the ability to work in a team (check performance with partners). However, in defence of the approach, it must be said that most content standards are designed for use in schools. It is arguably the case that language learning in this context is much more general, and less targeted, than adult language learning for specific purpose.

Nevertheless, the scale and complexity of content standards raise the question of the relationship between them and any form of a test. By definition, any form will sample from content standards rather than contain everything, and so the link between score meaning and the claim that learners have 'mastered' the content standards are tenuous at best. If the claim is from a test to score the content standards as a whole, a key validity claim is that from a tiny sample the score meaning can be generalised to a very

large universe of potential content. This is the enduring fundamental problem with any content-based approach to considering the validity of score meaning (Fulcher, 1999) that we discussed in Chapter 4. In other words, in any content alignment study there will be a question over whether the content standards are comprehensive enough at all performance levels, and even if there is agreement that they are generally useful for purpose, any form of a test will always be found to under-represent the content.

Before progressing to methodology, it is important to highlight a point of contrast with the content standards discussed here, and the Common European Framework of Reference. While the *Manual* (Council of Europe, 2009) recommends the analysis of content between the CEFR and tests, the CEFR does not contain much in the way of content and, where it does, it is not organised according to levels. This makes content comparison purely arbitrary in a way which it is not with models like WIDA. The CEFR invitation for readers to 'consider' what content might occur at each level is an invitation for invention, not comparison. Perhaps the lesson to be drawn is that under-specification of content is as much of a problem as over-specification for integrationists. I would argue that the cause of both problems is a lack of specific purpose for language use, and for intended score meaning. In short, they are models, not frameworks.

This aside, we turn to the methodology of content alignment. As with alignment to proficiency standards and setting cut scores, all methodologies rely upon expert judgement. Content standards can be used directly in developing curriculum, and 'check-lists' can be developed to ensure that curriculum content and teaching materials map on to content standards.

'Horizontal alignment' is the process of mapping tests and assessments to content standards. As we have argued, no single assessment can measure everything in a set of content standards. The first part of the process is to develop a test specification (see Chapter 5) drawing on the content standards as the domain of interest.

When a test form is compiled, the aim is to include at least one item for every content standard, although this is frequently not possible. The test specifications represent the first piece of alignment evidence. The next stage in the process is for a team of subject experts to compare the content of a number of test forms to the content standards, using a framework for the analysis. The most widely used methods are those of Achieve (www.achieve.org) and Webb (1999). We will describe the Achieve method below, as variants can be used by teachers in their own institutions, whereas the Webb method is more complex.

Before starting the alignment study, teachers are familiarised with the standards contents and the tests, using methods similar to those described in Chapter 8. Ideally, they should also take the test forms which are being used in the alignment study. Once they are fully familiar with all the documentation, the alignment study is conducted in three steps. (Note that the terms 'standard' and 'objective' are used interchangeably, although in some content standards one or the other is used as a superordinate, and the other to describe clusters of related target behaviours.)

Step One. Looking at each item/task individually, each teacher identifies the content standard objective(s) that the item measures, and applies a content standard code to

the item (this is sometimes called 'categorical concurrence'). The question that is being asked is whether the test items measure objectives in the content standards, and only objectives in the content standards. The aim is to ensure that learners are only being tested on what they will be covering in the curriculum. In the Achieve method, three judgements are made: firstly, confirmation that the item measures an objective within the content standards (code each item with the appropriate code(s) from the content standards); secondly, estimate content centrality; and thirdly, estimate performance centrality. Content centrality is whether the item clearly and explicitly measures the standard. Performance centrality is a judgement about whether the cognitive complexity of the item is similar to the cognitive complexity of the objective required in the content standards. This usually requires the judges to focus on functions of the standards, such as 'describe', 'compare' or 'analyse'. For both of these a grading system like the following is adopted:

0 = inconsistent
1A = not specific enough (standard or objective is too broad to be assured of item's strong alignment)
1B = somewhat consistent (item assesses only part, and the less central part, of a compound objective)
2 = clearly consistent.

The following example is taken from Rothman *et al.* (2002: 13):

[Passage read is 'When I Heard the Learn'd Astronomer' by Walt Whitman]

This poem is best classified as which of the following?

A. A sonnet
B. Epic poetry
C. Lyric poetry
D. A ballad

Relevant standard: 'The learner will analyze, synthesize, and organize information and discover related ideas, concepts, or generalizations.'

As knowledge of literary types does not appear to be a central part of this standard, the item received 1A for content centrality. The item received 0 for performance centrality as it only requires learners to identify the type of poem, whereas the function words of the standard are 'analyze', 'synthesize' and 'organize'.

Step Two. For each item, each teacher makes a judgement about how difficult or challenging the item is. This is in fact two judgements – one about the source of the difficulty, and the second about the level of difficulty. Source of difficulty is coded as 0 = inappropriate source, or 1 = appropriate source. An inappropriate source of difficulty is anything that is not construct relevant, such as a problem with the item, or a failure

on the part of the item writer to produce a good item. The level of difficulty is a simple yes/no decision for each item, assessed as to whether it is of a suitable level for the target learners in terms of the concepts the item employs, and how cognitively demanding it *is thought to be.*

Step Three. The balance and range of the test are considered. Balance is a judgement about whether groups of items that measure a particular standard focus upon the most important aspects of that standard, rather than on peripheral abilities. Range is not a judgement at all, but is the proportion of total standards/objectives from the content standards that are measured by at least one item, as calculated in step one. Rothman *et al.* (2002: 20) argue that a result of .67 or higher is considered to be a good range, while values of .5 and higher are acceptable.

Reports based upon such content alignment studies provide the evidence with which schools can show that they have adequately taken into account content standards in their curriculum, and that the achievement tests they have devised to measure student outcomes measure the same content. In this way, accountability is imposed through the alignment of both teaching and assessment to the external standards. It is a complex and time-consuming task, which is frequently outsourced to professional testing companies. But perhaps this is an intentional effect of the desire to construct an independent accountability machinery?

4. Preparing learners for tests

It is important to distinguish between two types of preparation (Popham, 1991). The first type is designed to familiarise learners with the item types on the test, the kinds of instructions they will encounter, and give them practice in working within time constraints. If it is a computer-based test, preparation will also include becoming familiar with the interface and navigating through the test. The purpose of this kind of test preparation is to ensure that the learners do not spend time and effort having to work out what they should be doing during the test. Relating this to the ever-present question of validity, this type of preparation reduces the chance that scores will be affected by their unfamiliarity with any aspect of the test; it therefore increases the validity of score meaning by removing a potential source of construct-irrelevant variance. The second kind of test preparation is designed to increase the score of the test taker by instilling test-taking techniques that focus upon the test items, rather than improving the learner's ability on the constructs in question. For example, by spending time looking at the options in multiple-choice items to discover how frequently the longest option is likely to be the correct response, or attempting to match lexical items in the stem to synonyms in the key (see Chapter 6). Haladyna *et al.* (1991: 4) refer to the effects of such preparation as 'test score pollution', and claim that these practices are unethical.

It is not surprising that Haladyna *et al.* also consider it to be unethical to base a curriculum on a test, and to align standards, curriculum and assessments. The reason they

give for their separatist stance is that tests are developed and piloted using samples drawn from populations who are not given targeted training in the test content. It is claimed that increasing curriculum alignment results in excessive test preparation that changes the meaning of test scores, and hence reduces validity (Shepard, 1990). Basically, the argument is that if you know what's on the test and you teach to it, it is hardly surprising that the learners get higher scores. But this undermines the validity of the score interpretation. Hamp-Lyons (1998) also questions the ethicality of test preparation practices that focus upon using past test papers (other than for test familiarisation), criticising particularly the production of teaching materials that merely copy test content.

This is most troubling for language teachers who have learners in their classes who wish to pass a test, and have a clear idea of what they should be doing in order to achieve this goal – even if they are mistaken. The dilemma arises from what we have observed in previous chapters; namely, that language testing has become a high-stakes, high-value activity. Test scores and test certificates provide access to many of the good things in life, and so teachers are expected to secure examination passes.

Although it may not make our lives easier, at least we can find some consolation in the fact that it has always been so. Miyazaki (1981) tells us of the long years of test preparation required in ancient China, starting in the home at around the age of 3, with formal education beginning at the age of 7. Much of the learning consisted of recitation and memorisation. From these very early times we also learn that the test results were treated not only as a measure of the success of the test taker, but also the skill of the teacher: 'If education is not strict, it shows that the teacher is lazy' (1981: 15). This provides us with the other part of the test preparation dilemma. The accountability agenda of politicians holds teachers responsible for outcomes, and as we have seen above, alignment to standards is part of this agenda. And so questionable test preparation practices emerge, as we saw in Chapter 9. In ancient China there were also 'quick fixes' to get higher scores. Miyazaki (1981: 17) says:

> *Despite repeated official and private injunctions to study the Four Books and Five Classics honestly, rapid-study methods were devised for the sole purpose of preparing candidates for the examinations. Because not very many places in the classics were suitable as subjects for examination questions, similar passages and problems were often repeated. Aware of this, publishers compiled collections of examination answers, and a candidate who, relying on these publications, guessed successfully during the course of his own examination could obtain a good rating without having worked very hard ... Reports from perturbed officials caused the government to issue frequent prohibitions of the publication of such collections of model answers, but since it was a profitable business with a steady demand, ways of issuing them surreptitiously were arranged, and time and again the prohibitions rapidly became mere empty formalities.*

History repeats itself, and policy makers do not learn. Curriculum alignment leads to teaching to the test, and as soon as test preparation practices and materials focus upon the test, teachers are rounded upon for doing precisely what they have been encouraged

to do. Similarly, accountability leads to the publication of 'league tables' for schools, most of which are tied to rewards and sanctions of some kind. The public now enjoy looking up particular schools to see where they are in the hierarchy of performance. Teachers respond to this by trying to increase test scores by any means possible, given limited time and resources.

The term 'cramming' for test preparation introduced in the nineteenth century, and survives to this day:

> *Those who afford this kind of preparation are often called crammers. Now so far as this term implies any opprobrium it is unjustly applied; a market has been opened for a particular kind of fabric, the stouter and costlier stuffs are thereby rendered less saleable, and the mill owners must meet the popular demand or close his mills. People are hardly aware of how thoroughly the educational world is governed by the ordinary economical rules. While employing the motives of gain and advancement most profusely, the public seems to find fault with teachers and pupils for being influenced by these considerations ... they make learning a marketable commodity and then complain that it is grown for the market.*
> (Latham, 1877: 6–7)

These market forces compel teachers to devise the test preparation practices to meet demand. And thus,

> *The tutor must consider not what studies or what kind of teaching will do him [the learner] most good, but what studies will yield the highest aggregate in the given time, and he must teach his pupil each subject not with a view to call out his intelligence, but with a view to producing the greatest show on a stated day; for instance he must teach him a language by some sort of Ollendorff process, which shall address itself to the ear and the memory, rather than by a method which involves any grammatical analysis.*
> (Latham, 1877: 5–6)

It was also acknowledged from the earliest times that test preparation was pretty boring, and distracted learners from the real task of learning. The famous statistician Karl Pearson wrote a biography of Sir Francis Galton, who was one of the first 'scientific' testers. He reports that in 1889 Galton had argued the best way to avoid learners cramming for tests is to have lots of different kinds of tests and assessments, to vary the content as much as possible, and to ensure that what is tested cannot be easily memorised. This is still good advice in the twenty-first century, and one which many test developers try to follow. Cognitively demanding questions that require the application of knowledge and problem solving are particularly desirable.

One way of reducing the amount of test preparation is, of course, not to use tests for teacher and institutional accountability as well as learner achievement. While decoupling the two is not likely to happen for political reasons, there are excellent educational reasons for doing this when possible. Mansell (2007) provides an excellent account of what happens to education when the testing system is used by policy makers for the

'hyper-accountability' that we briefly discussed in Chapter 1. The tests are used for more purposes than they can reasonably bear, and the pressure to find short cuts is increased immensely. It leads not only to poor test preparation practices, but also encourages cheating on coursework, such as dictating answers, and allowing copying from the internet without sanction. With particular reference to language test preparation, Mansell cites practices such as teaching set phrases that can be reproduced in multiple contexts (speaking or writing) without thought (the burglar's 'swag bag' technique; 2007: 85), and the memorisation and scripting of written and oral work (2007: 89–93). He reports a senior official from an examination board as saying that without these techniques teachers simply won't be able to get the 'grim kids' to pass any tests.

This is where the unethical side of test preparation begins to get very sad indeed. While testing is primarily meant to be about meritocracy and giving people the chance to make the best of themselves, hyper-accountability can reduce real learning opportunities and damage equality of opportunity. The 'grim kids' mentality is a symptom of just how damaging the unintended consequences of some testing practices have become. Some politicians do understand this, however. I was struck by this passage from Obama (2006: 163):

> *While I was talking to some of the teachers about the challenges they faced, one young teacher mentioned what she called the 'These Kids Syndrome' – the willingness of society to find a million excuses for why 'these kinds' can't learn; how 'these kids come from tough backgrounds' or 'these kids are too far behind.' 'When I hear that term, it drives me nuts,' the teacher told me. 'They're not "these kinds". They're our kids'. How America's economy performs in the years to come may depend largely on how well we take such wisdom to heart.*

A narrow focus on test preparation therefore not only threatens the validity of the assessment, it undermines the entire rationale for having tests: providing genuine opportunity, and meritocratic access to education and employment. It actively disadvantages those who are already disadvantaged. In addition, there is no evidence to suggest that test preparation (other than as familiarisation) has any positive impact on test scores at all. When it comes to language learning, quick fixes have always been expensive delusions. The research evidence suggests that learner success in tests depends on the quality of the teaching, teacher training and factors associated with the educational system and resources (e.g. Alderson and Hamp-Lyons, 1996). A study by Robb and Ercanbrack (1999) using TOEIC showed that most gains when undertaking test preparation came from using English, rather than the test preparation *per se*. Similarly, investigating IELTS preparation, Comber (1998) and Bialy (2003) studied the effect of test preparation programmes in China, and discovered that, while test preparation did have a marginal impact on scores, this was not as high as teaching the language. Among other studies of the effect of test preparation on language test scores the findings are remarkably similar (Read and Hayes, 2003; Elder and O'Loughlin, 2003; Zhengdong, 2009). While Brown (1998) found that test preparation did increase test scores on the IELTS writing sub-test, the 'preparation' turned out to consist of a course in writing

skills with timed writing practice. This kind of preparation is much closer to teaching than the kinds of activities normally associated with construct-irrelevant test preparation. Similarly, in a meta-analysis of studies looking at the impact of test preparation on the US SAT test scores, Powers (1993) found no significant impact.

From theoretical, ethical and practical points of view, we can therefore say that test preparation, other than for familiarisation, is not only a waste of time, but actively detrimental to learners and the educational system. I doubt, however, that the argument and evidence will have a significant impact upon the test preparation industry, which will always seek to find the quick, easy, ways to improve scores even if there is no significant change in competence. Latham (1877: 149) sums up the damage that this practice has to learners and society:

> *If we damage the general standard of truthfulness by leading young men to glory in having outwitted Examiners, and seemed to be what they are not ... then we lose far more morally than we gain in any other way.*

The best advice to teachers whose role is to prepare students for tests is therefore to teach the language using the most appropriate methodologies and materials to achieve communicative competence for the learners to be able to function in the target domains and contexts. Looking at tests and test items should be a minor part of any course, with the goal of familiarisation so that learners know what to do. Once this has been achieved, using test materials has no further value.

5. Selecting and using tests

Teachers and other language professionals are also frequently required to select 'off the peg' tests to use in their institutions. This is frequently the case where the time and resources to develop a local test are not available. This is almost always a second-best solution for local assessment needs where external certification is not required, as it is very difficult to find a test that does precisely what you wish it to do.

The criteria for selecting a test should be drawn up before starting the search. These will differ according to the purpose for which you need the test. However, we are able to generate a number of generic criteria that can be used as a starting point.

A. Test purpose

What decisions do you wish to make?

These may be very specific, or you may wish to classify them more generally as placement, achievement, proficiency, diagnostic or aptitude testing.

Was the test designed to make these decisions?

What evidence is provided by the test developer to justify the use of the test for this purpose?

B. Test taker characteristics

What are the key characteristics of the population to whom you will give the test?

These may include age, gender, educational level, language proficiency, first language background, particular educational or career goals, and so on.
Was the test designed for this population?
Is the test at the right level of difficulty?
What evidence is provided by the test developer to show that the test was piloted using a sample of learners similar to those with whom you are working?

C. Domain of interest

Do you wish to predict performance in a particular domain, such as a particular occupational field?
Is the test content relevant to the domain of prediction?
What evidence is provided by the test developer that the domain has been adequately defined and sampled?

D. Constructs of interest

What particular constructs do you wish to assess?
Does the test assess these constructs?
What evidence is provided by the test developer that these constructs have been included in the test specifications?

E. Reliability

How reliable do you wish the results to be?
How is the test scored?
How are scores reported?
What evidence is provided by the test developer regarding reliability, including information on the standard error of measurement or other estimates of potential sources of error?

F. Validity

What evidence is there that we can correctly make inferences about constructs from the scores?
Does the test provider make research reports available?
Does the test provider make claims about the usefulness of scores to constructs or domains that are not part of test purpose?

G. Parallel or equated forms

Do you need multiple forms of a test in order to maintain security?
Are multiple forms available?
What evidence is provided by the test developer that forms have been developed to ensure that scores do not vary significantly across forms?

H. Test administration and practicality

What resources, including time and budget, do you have available?
Can the test be administered and scored using available resources?
Do you have the funds to purchase the test and continue using it over the projected time scale?

I. Impact

What impact or washback do you expect the use of the test to have on your institution, and on classroom teaching and learning?

Is the format and content of the test likely to produce the intended impact?

What evidence is provided by the test developer to suggest that the intended impact is likely?

Many more criteria may be established. For example, if you wish scores to be reported on a particular set of standards, you will need to select a test for which a standard-setting exercise has been undertaken. You would then need to know how sensitive the test was to the standard, how many cut scores had been established, how dependable these were, which standard-setting technique had been used, and how well the standard-setting study had been undertaken. The more criteria you establish, the more effort will be needed to come to a conclusion. However, all the effort you put in will be rewarded.

The second step is amassing the information you will need to make the necessary judgements against each of the criteria. This will include descriptions of the tests, samples, descriptions of the test purpose and design processes, summaries or full texts of relevant research, and information on administration and cost. These are usually freely available from the websites of professional test development agencies.

The third step is trawling through the information to get the answers to the key questions, and to evaluate each test against the established criteria. This is best done by a small study group rather than individuals, so that different views on the quality of information and studies are taken into account. It is advisable for a short report that sets out questions, answers, and evidence to support decisions. If the selected test does not work as expected, this document may form the basis for a review of the decision, and the selection of a new test.

If you have difficulty finding a test for your particular purpose, two popular tools are available to help you.

Tool 1: The Foreign Language Assessment Directory (Center for Applied Linguistics, 2007)

Available at: http://www.cal.org/CalWebDB/FLAD/

Description: A searchable database of tests that allows you to search by name of test, language, US grade level, proficiency level (on the ACTFL scale), intended test uses, and skills to be tested. The Center for Applied Linguistics also provides an online tutorial (Center for Applied Linguistics, 2009) that guides language professionals in the selection and use of language tests using the database.

Tool 2: ETS TestLink

Available at: http://204.50.92.130/ETS_Test_Collection/Portal.aspx

Description: A searchable database of more than 2500 tests, only some of which are language tests. It is possible to search by test title, language or skill. The database provides a summary description of each test, information on the author (where

available) and the test provider. This information can be used to trace availability and further information.

The general question we are trying to address is: Is the use of this test valid for the intended purpose? The validation of a testing procedure is concerned with ensuring that the test score means what it claims, and that this meaning is both relevant and useful for any decision we intend to make about the test takers. Validation is never an 'all or nothing' affair. What we need to know is whether the evidence is good enough to support the use of the test, and whether it provides us with the necessary warnings about the inevitable uncertainty that comes with any test.

In short, it is about the test producer constructing an argument for the use of the test for a specific purpose, and our evaluation of whether the argument is a good one (Kane, 1992, 2006). An argument involves a claim about the meaning of a score. Evidence (or data) is amassed in support of this claim. The reason why the evidence supports the claim is provided in a warrant or justification that argues why the two are linked. The argument can be supported by evidence from other research studies or theoretical arguments, and this is often called the 'backing' for the argument. Finally, good arguments usually take into account alternative explanations – or show that the test score is not affected by construct irrelevant variance (Fulcher and Davidson, 2007: 164–166). Some large-scale testing activities directly address the question of the argument for score meaning and test use (Chapelle, Enright and Jamieson, 2008), but this is the exception rather than the rule. One of the important points to recognise is that the more convincing the argument for the use of a test for a particular purpose, the less convincing the argument becomes for other purposes.

We therefore come full circle in our discussion. No test is good for any purpose at all, as Carroll (1961) stated in his groundbreaking contribution. Teachers and language professionals must still be vigilant against tests that have been designed for one purpose being used for another. Tests cannot be 'reused' or 'repurposed' unless there is a clear retrofit argument to show why the test is relevant to its new purpose, or how it has been adapted to suit the new purpose (Fulcher and Davidson, 2009). Some testing organisations tackle this problem head on, recognising the financial advantages that are gained (Wendler and Powers, 2009), but charge the clients with the primary responsibility to review the test or employ 'experts who best know the new test-taking population or the needs of the score users' (2009: 3) to review the suitability of the test. When selecting an 'off-the-peg' solution, beware of taking the easy option without taking due care.

6. The gold standard

As this final chapter is about tests and their impact upon the classroom, I intend to conclude with a brief consideration of the teacher as an assessor. Testing and assessment theory, and *psychometrics* in particular, have always had what can only be called a 'love-hate' relationship with teachers, as Shepard (1991, 1995, 2000) has noted. One rationale for the use of tests is that teachers are either unreliable, or just downright

capricious, in how they grade students. On the other hand, when a criterion is needed for the evaluation of new scoring mechanisms or test tasks, the scores are correlated with teacher assessments. Human ratings remain the 'gold standard' against which other scores are assessed (Chodorow and Burstein, 2004). This dilemma is not going to go away, and so perhaps it is important to set it out as clearly as possible, and learn to live with it.

In his classic treatment, Ruch (1924: 2), with reference to a previous work by Caldwell and Courtis (1923), lists the reasons why a formal test is to be preferred to teacher assessments as:

1. It is impartial.
2. It is just to the pupils.
3. It is more thorough than older forms of examination.
4. It prevents the 'officious interference' of the teacher.
5. It 'determines, beyond appeal of gainsaying, whether the pupils have been faithfully and competently taught'.
6. It takes away 'all possibility of favoritism'.
7. It makes the information obtained available to all.
8. It enables all to appraise the ease or difficulty of the questions.

The assumption underlying this list is that some teachers are not capable of impartiality, and that some teacher assessment is open to unreliable fluctuation that is influenced by construct-irrelevant factors, such as whether a pupil is 'liked'. Two other themes emerge. If tests are used, they can be made available for others to judge how well the assessment has been carried out, whereas a teacher's judgements are not open to such investigation. The other is that tests present the opportunity to judge not only the learner, but the pedagogical abilities and efforts of the teacher.

But how do we know if a test is a good measure of a construct? We have looked at many ways to look at the validation of test scores and how we interpret them. One of the earliest methods, and one that is still very widespread today, is the comparison of test scores with the expert judgments. And who are the best possible judges? Burt (1923: 199) was in no doubt:

> *There is no standard of comparison which can surpass or supersede the considered estimate of an observant teacher, working daily with the individual children over a period of several months or years. This is the criterion I have used.*

As we saw in Chapter 7, the argument about whether machines are capable of scoring constructed responses – either written or spoken – largely hangs on the degree to which they can show agreement with expert human judges, most of whom are classroom teachers. Many studies seek high correlations with teacher judgements as a source of evidence to justify a validity claim for a test. As Kane (2006) argues, teachers are able to refine their views of a student's ability over time and have access to much more complex data than the language test can collect. While this does not guarantee a fair outcome, Kane (2006: 49) believes:

If a qualified teacher applies appropriate criteria to a student's performance, the results would have a strong presumptive claim on our confidence. This is the kind of evaluation that might well serve as a criterion in validating some standardized test.

Teachers are sometimes seen as part of the problem, and then as part of the solution. Perhaps the answer lies somewhere between the two extremes. It has always been intuitively obvious that sensitive teachers with a deep and extensive knowledge of learners over a period of time have an ability to make valid and reliable estimations of their capacities and abilities. That teachers can make fair and dependable decisions (see Chapter 3) is an assumption upon which all classroom assessment is based, and it can be investigated by teachers themselves within the context of professional development and team working. Teacher assessments can be used in place of, or in conjunction with, more formal assessments (Fulcher, 1991). This is the basis of marking coursework and other alternative forms of assessment. Nevertheless, in high-stakes assessment it is not always fair to place the burden of making judgements upon teachers, especially when they have worked with the learners for a long period of time and have come to care deeply about their progress and future careers. Even when a learner clearly does not meet a standard, it is not appropriate to ask a caring teacher to make this decision and communicate it to the learner. It is much better if difficult decisions are taken out of the teacher's hands. The teacher's role is to do the best for the learner, but not to take the blame if the learner ultimately does not meet a required standard.

In a sense, therefore, the teacher is the 'gold standard'. But it is a standard that can always be improved and monitored through the range of techniques that we have described in this book, from designing test specifications to monitoring dependability of judgements. Tests and assessments of all kinds have a legitimate place. The different types of assessment should not be seen as competing with each other. Each fulfils a different function. Some of these are frustrating for teachers – and always have been. But once we can appreciate the conflicting rationales for the varieties of language tests, when we understand their purpose and history, we are much more able to negotiate their use, and control their impact on our lives.