The practice of language assessment

Glenn Fulcher
University of Leicester, UK

Introduction

Language teachers and applied linguists are constantly engaged with language testing and assessment. Whether informal or formal, assessment procedures are designed to collect evidence for decision making. In low-stakes classroom contexts the evidence is used to provide formative feedback to learners (Rea-Dickins, 2006), scaffold learning (Lantolf, 2009), evaluate the quality and pace of learning, inform pedagogy, and feed into curriculum design (Shepard, 2000; Hill, 2012). In high-stakes contexts identifiable ‘test events’ produce scores that are used to make life changing decisions about the future of learners (Kunnan, 2013). In research, tests are used to generate data that are used to address questions regarding second language acquisition (Chapelle, 1994).

When we use tests or assessments in each of these contexts we make a number of generic assumptions. These are frequently subconscious, although they do occasionally become the focus of overt scrutiny. The four headline assumptions are:

1. The tasks undertaken are relevant to the decision to be made, or the research question to be addressed.
2. The responses elicited by the tasks are useful in making the decision or addressing the question.
3. The ‘score’ adequately summarizes the responses.
4. ‘Scores’ can be used as necessary but not always sufficient evidence for decisions.
   (Fulcher, 2015, p. 2)

The word ‘score’ is placed within quotation marks because it us used to mean any reductive rendering of a complex performance or set of responses to a number, letter, or verbal description of what we infer a learner can do as a result of our observation of their performance. The verbal summaries may be informal teacher comments or notes, or be phrased in terms of ‘can do statements’ or performance level descriptors, as commonly found in standards or framework documents (Fulcher, 2016).

The assumptions we make in all contexts lead to the claim that testing and assessment practices need to be deliberate. They require careful thought, planning, and documenting. This is true of even informal classroom assessment for learning, where the primary purpose is to provide formative feedback to learners that helps them to see the ‘gap’ between what they can currently do, and their next goal. Thus, Black et al. (2003) suggest that planning in assessment for learning should focus on the design of tasks that involve interaction between learners, and feedback on performance from both peers and teachers that is timely and informative. In addition, classroom management should provide time for learners to reflect on feedback and respond to it. Being deliberate in planning requires a level of assessment literacy that includes the principles and concepts that guide practice, as well as the practical
skills in assessment construction (Fulcher, 2012a). In this chapter we focus on the latter, as there are generic practices that can improve the quality of testing and assessment among all professionals who need to use language tests (Coniam, 2009). For the sake of brevity, from now on I will use the term ‘test’ to refer to any testing or assessment practice, formal or informal, as a superordinate.

**The language test design cycle**

The extent to which we can take the assumptions listed above for granted depends upon the deliberate care with which we construct a test or assessment instrument. We can think of this as the test design cycle, as illustrated in figure 1.

![Test Design Cycle Diagram](image)

**Fig. 1. The test design cycle (Fulcher, 2010, p. 94).**

Mislevy and Riconscente (2005) and Fulcher (2010; 2013) liken the process of building a test by following a design cycle to the work of an architect, a metaphor that has been widely adopted elsewhere (Green, 2014, 26–29). The first thing an architect does is consider the purpose of the building. Once we know the purpose, it is possible to consider its likely uses. From there it is possible to start making other decisions, such as how large the building should be, what materials might be used in its construction, and the internal layout of the structure. This all has to be done within the resource and financial constraints that might apply. It must also meet any regulatory requirements of the industry or state. Planning and building a test is no different. In what follows, I break down the process represented in the test design cycle into four distinct types of activity. While these are conceptually distinct, they are not all temporally separate, as they overlap from the very start of the design process.
Activity 1: From Test Purpose to Test Tasks

The first questions we must ask ourselves are “What is this test/assessment for?” and “What decisions are to be made on the basis of the evidence generated?” As Ingram (1968, p. 70) reminds us: “All tests are for a purpose. A test that is made up without a clear idea of what it is for, is no good”. In other words, the last step in the test cycle determines the initial statement of purpose.

Next, we need to know what the test criterion is. That is, the ultimate ‘real world’ activities that the test takers are expected to undertake. For example, a common criterion within an educational setting for English language learners (ELLs) is learning about mathematics in a second language setting (Chapin et al., 2009). In this case, the criterion would be the domain of mathematical English required to gain access to school content disciplines.

Thirdly, we may need to state what skills, knowledge or abilities (KSAs) a learner would need to achieve the target goal. In language testing KSAs are frequently referred to as constructs, the abstract abilities that underlie successful performance. Normally labelled with abstract nouns, these might include specific levels of fluency in language use, or performance with certain levels of accuracy. When constructs are specified in standards documents, operational illustrations are frequently provided to give indications of the task types that might be used to evaluate constructs. Thus, the Common Core Standards for ELLs in the United States list one construct as the “integration of knowledge and ideas”. This is an abstract construct. One suggested task type at a very general descriptive level is:

“Translate quantitative or technical information expressed in words in a text into visual information (e.g., a table or chart) and translate information expressed visually or mathematically (e.g., in an equation) into words” (Common Core State Standards, 2010, p. 62).

The example is an operationalization of the construct. It represents the claim that if a learner is able to translate information between text and tables/charts/equations, they are providing evidence that they can “integrate knowledge and ideas”.

We can see from this example how the statement of purpose, the criterion, and the construct definition, lead directly into suggested operationalizations: the design of tasks that are directly related to the test purpose. The rationale for the use of tasks becomes part of a validation narrative that underpins why we think the first two assumptions we make about our test are warranted. More technically, Chapelle (2008, p. 320) refers to this as a “validity argument from test design.” Validity is the principal quality of a test which demonstrates that it meets industry standards for safe use.

Consider the following assessment task “School Trip” (Bowland Maths, 2010), which illustrates the first part of the test design cycle.
Mr. Richards, a teacher at Bosworth School, plans to take 30 pupils on a school trip. Here are the places they could visit.

<table>
<thead>
<tr>
<th>Growlets Zoo</th>
<th>Prison Museum</th>
<th>Space Science Show</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Lion" /></td>
<td><img src="image" alt="Prison" /></td>
<td><img src="image" alt="Planets" /></td>
</tr>
<tr>
<td>36 miles from Bosworth</td>
<td>30 miles from Bosworth</td>
<td>10 miles from Bosworth</td>
</tr>
<tr>
<td>Entrance fee £2.50 per person</td>
<td>Entrance fee £6 per person</td>
<td>Entrance fee £10 per person</td>
</tr>
</tbody>
</table>

The class vote on which place to visit. Here are the results:

<table>
<thead>
<tr>
<th>Name</th>
<th>First Choice</th>
<th>Second choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Olivia</td>
<td>Zoo</td>
<td>Space show</td>
</tr>
<tr>
<td>Grace</td>
<td>Space show</td>
<td>Prison museum</td>
</tr>
<tr>
<td>Jessica</td>
<td>Prison museum</td>
<td>Zoo</td>
</tr>
<tr>
<td>Ruby</td>
<td>Zoo</td>
<td>Space show</td>
</tr>
<tr>
<td>Emily</td>
<td>Space show</td>
<td>Prison museum</td>
</tr>
<tr>
<td>Sophie</td>
<td>Prison museum</td>
<td>Zoo</td>
</tr>
<tr>
<td>Chloe</td>
<td>Prison museum</td>
<td>Space show</td>
</tr>
<tr>
<td>Lucy</td>
<td>Prison museum</td>
<td>Space show</td>
</tr>
<tr>
<td>Lily</td>
<td>Space show</td>
<td>Prison museum</td>
</tr>
<tr>
<td>Ellie</td>
<td>Space show</td>
<td>Prison museum</td>
</tr>
<tr>
<td>Ella</td>
<td>Zoo</td>
<td>Space show</td>
</tr>
<tr>
<td>Charlotte</td>
<td>Space show</td>
<td>Prison museum</td>
</tr>
<tr>
<td>Katie</td>
<td>Space show</td>
<td>Prison museum</td>
</tr>
<tr>
<td>Mia</td>
<td>Zoo</td>
<td>Space show</td>
</tr>
<tr>
<td>Hannah</td>
<td>Zoo</td>
<td>Space show</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>First Choice</th>
<th>Second choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jack</td>
<td>Prison museum</td>
<td>Zoo</td>
</tr>
<tr>
<td>Thomas</td>
<td>Zoo</td>
<td>Prison museum</td>
</tr>
<tr>
<td>Joshua</td>
<td>Zoo</td>
<td>Prison museum</td>
</tr>
<tr>
<td>Oliver</td>
<td>Space show</td>
<td>Prison museum</td>
</tr>
<tr>
<td>Harry</td>
<td>Prison museum</td>
<td>Zoo</td>
</tr>
<tr>
<td>James</td>
<td>Zoo</td>
<td>Space show</td>
</tr>
<tr>
<td>William</td>
<td>Space show</td>
<td>Prison museum</td>
</tr>
<tr>
<td>Samuel</td>
<td>Zoo</td>
<td>Prison museum</td>
</tr>
<tr>
<td>Daniel</td>
<td>Zoo</td>
<td>Space show</td>
</tr>
<tr>
<td>Charlie</td>
<td>Prison museum</td>
<td>Prison museum</td>
</tr>
<tr>
<td>Benjamin</td>
<td>Space show</td>
<td>Zoo</td>
</tr>
<tr>
<td>Joseph</td>
<td>Zoo</td>
<td>Prison museum</td>
</tr>
<tr>
<td>Callum</td>
<td>Zoo</td>
<td>Prison museum</td>
</tr>
<tr>
<td>George</td>
<td>Prison museum</td>
<td>Space show</td>
</tr>
<tr>
<td>Jake</td>
<td>Space show</td>
<td>Prison museum</td>
</tr>
</tbody>
</table>

Taking first and second choices into account, where do you think Mr Richards should take them? Explain how you decided.

In this task learners are presented with data about options for a school trip in numerical and tabular format. The question requires the learners to manipulate the data in order to arrive at a judgment about which trip is most appropriate for the class. This judgment and the reasoning by which it is made has to be explained in writing. The writing constitutes the evidence upon which we infer whether (and to what degree) the respondent is able to “integrate knowledge and ideas” within this domain. The testing process is valid to the
extent that our inferences are sound and lead to correct judgments and decisions. This is the core meaning of validity in testing.

The requirement to be explicit about test purpose cannot be stressed enough.

Consider the following two statements of purpose:

1. Upon entry to the country the immediate needs of migrant workers with limited language skills have been identified as (a) searching for suitable manual or semi-skilled work, and (b) accessing local social services and healthcare. In order to assess their ability to carry out the necessary reading tasks to gain basic information a reading test is required to screen new arrivals with the purpose of planning and providing suitable language programmes.

2. Students in the college business stream are required to attend a course on negotiating contracts and agreements for international companies. All communication should be interculturally sensitive and avoid loss of face, but achieve the primary goals of the business requirements. Successful students (on this and content courses) will graduate with a certificate in intercultural business management.

These statements are short and succinct, but give a clear sense of the purpose of the assessment. The statements are also indicative of the kind of content that we expect to find in a test, because they specify the test criterion. Real world activities may include reading job postings, reading social service information leaflets (purpose 1), or taking part in business meetings and negotiating contracts/agreements (purpose 2). The language that is required for each purpose is likely to be very different, and it is unlikely that a validation narrative created to justify the tasks created for one purpose could be easily translated into another.

Activity 2: Specification Writing and Iteration

We will return to the architecture metaphor. The first thing that an architect works on is a blueprint of the building. In testing and assessment, the blueprint is called a test specification (Popham, 1978). The test specification (often shortened to “test spec”) is the design document for a test and consists of an overall plan, together with specifications for each of the task types that may appear in the test (Davidson & Lynch, 2002). We can think of this as the ‘structure’ and the ‘stuff’ that we are going to put into the space created by the structure. Writing an explicit test spec forces the deliberateness of the design and development process. Furthermore, without test and item specs it is impossible to produce other forms of a test. In fact, “the test” is actually the test spec – the blueprint. From the spec it is possible to create a number of forms. The forms are generated by the same spec, and so are supposed to be parallel in the sense that they test the same things, to the same level of difficulty, but have different content. This is often necessary when test takers are taking a test at different times (for security purposes), or when a test is being given in to assess learning in pre- and post-test research settings. A sample specification for example 1 above (migrant workers) is provided in Fulcher (2010, pp. 139 – 146) as an example of how this might be done.
Apart from the task specs, the specifications also contain information that pertains to the overall “look and feel” of the test – what we have referred to as the structure, and also the façade. Following Mislevey et al. (2003) the other parts of a specification should include:

**Evidence specification:** The way in which the test takers are expected to respond to the tasks. This may be closed response (such as multiple choice), or open response (such as writing, in the example above). It also includes information on how the response is to be scored, and what the score means. This part of the spec explicitly addresses headline assumptions 3 and 4 above. With reference to the construct “Integration of Knowledge and Ideas” for use with the task “School Trip”, the following rubric for “clarity and clearness of description” has been suggested:

- **Level 1:** Describes then decision or calculations made but this is incomplete and/or contains errors
- **Level 2:** Describes the decision made and cost of the trip but with some errors
- **Level 3:** Describes their decision making process and methods but lacks clarity
- **Level 4:** Clearly describes their decision making process and their methods

There are problems with producing evidence models. The first is deciding how many levels (scores) can be identified by human raters. The second is establishing the criteria that make one level different from another. The third is putting these criteria into verbal descriptors that make sense to judges and others who may see and use the scores. In the example above the only real distinction between levels 3 and 4 is the movement from “lacks clarity” in description to “clearly” describing. This begs the question what constitutes “clarity”. This is another construct that requires operationalization.

In practice, *benchmark samples* are frequently used to solve problems caused by vagueness and imprecision in rubrics. These are selected from responses to tasks and used to illustrate typical performances at each level. Judges are then expected to match future performances to a given level by comparing it with the benchmark sample. This common strategy, should it work, is placed into the validation narrative to support the inferences that may be drawn from the scores by linking the meaning of each number to its benchmark. The following sample response to the School Trip task above is provided as a typical level 4 response because of its “clarity”.


Assembly specification: Tests do not normally contain one task or even one task type. The assembly spec tells the test producer how many items of each type are required to construct a complete form. The purpose of the spec is to ensure that each form of the test covers the same range of constructs and domains of interest, has enough tasks to collect reliable evidence regarding performance, and to ensure that each form will take the same amount of time to complete.

Delivery specification: This spec states how the test will be delivered. This may include paper based or computer mediated. But it also includes the administrative details that are so important for test security. For example, how much space is to be left between desks, how many invigilators are to be used relative to number of candidates, instructions to test takers, or restrictions on additional help that might be available (e.g., dictionaries, internet access, and so on).

Presentation specification: The presentation spec provides the small details about precisely what a test taker will see when presented with the test tasks. This might include the font size, whether or not the font size may be altered if the test is delivered by computer, where icons appear, or what colors are used. These are actually very important because presentation decisions often impact upon test takers with disabilities (e.g., visual impairment and font/color choices). Inappropriate use of icons or navigation may slow down learners who take more time trying to find their way through the test than responding to the tasks.

The process of spec writing begins as soon as test purpose is being established. The purpose is part of the specification. So is the description of the test criterion, domains of interest, and constructs. These are all part of the general description of the test rationale. Each task type must also have a spec, together with its associated evidence spec. These specs go through iterations and are revised each time there is a change to a task type as the test designers refine their understanding of constructs and how best to get useful evidence.
through test procedures. This happens most noticeably during the process of prototyping, to which we turn next.

**Activity 3: Prototyping and Piloting**

*Prototyping* is the process of “trying out” prototype tasks to see if they work. Prototyping typically involves asking small numbers of test takers (typically around 5) to try out new tasks. As they are doing the tasks, or immediately upon completion, they are asked to verbalize how they arrived at the response they did, and describe any problems they faced in doing the task. Prototyping generates two types of data for analysis: the responses to the task, and the verbalization of how the participant responded to the task. The data is studied to discover primarily if the way they respond to the task is relevant to the construct that the designer intended to assess, and whether the item type and format is of suitable difficulty for the intended population (Nissan & Schedl, 2012, p. 283). Surprisingly, most problems with tasks are discovered with no more than two or three iterations of this process, with changes to the task (and the specification) at each iteration to remove the identified problems. In contexts where learners with disabilities are expected to undertake tasks it is essential that accommodations to mitigate the effects of a disability are designed and prototyped simultaneously (Abedi, 2012). Typical accommodations might include the provision of additional time, audio, amanuensis, braille or large print alternatives, screen magnification, or single person test administration.

The fact that prototyping can be undertaken with so few participants means that it is feasible not only for formal tests, but also as a preliminary for the use of assessment tasks in schools. It should also be stressed that prototyping is fundamental research. The test designer or teacher is asking a fundamental validity question about the use of tasks and the meaning of the responses: do these genuinely generate useful construct relevant evidence that allows successful inferencing and decision making? When prototyping is undertaken in schools, in a context where teachers collaboratively write and refine task specifications, this research enables the creation and evolution of a common understanding of learning goals and the practical meaning of constructs.

*Piloting* is only undertaken when it is necessary to generate statistical data for a test by giving it to a much larger sample of participants who are representative of the test taking population. For this an assembly model is required, and the pilot test is produced according to the delivery and presentation specification. The designers are asking the question: does this all work? (Kenyon and MacGregor, 2012). The data allows researchers to identify any task types that are too difficult or too easy, and to check that the timing of sections has been adequately estimated. Timing studies ensure that the test administration is not “speeded”, as it is well known that if learners do not have sufficient time to complete a test in the time allocated they begin to guess; and guessing introduces score variation that is not related to the constructs of interest.

**Activity 4: The Field Test and The Go/No Go Decision**

In high stakes testing, when all iterations of the design process are complete at least two full forms of the test are given to a large sample of participants drawn from the test taking
population. At this stage the norms for the test are established. The administrative procedures and security features are tested to ensure that all works well. At the end of the field test the test developer can make an informed “go/no go” decision, which determines whether the test can now be used in live assessment for its intended purpose.

The test roll-out is not the end of the process. Constant monitoring of test results is essential, because the test population may change over time. As applied linguistic research improves the definition of constructs and our understanding of communication, it becomes possible to generate new operationalizations of constructs, which leads to new task types. Over time the test undergoes upgrade retrofits as some task types are retired, and new ones introduced to better represent the construct (Fulcher and Davidson, 2009).

Trends and controversies

In recent years the most significant trend in high stakes testing has been the use of tests for secondary purposes for which they were not originally designed, sometimes referred to as “repurposing” (Wendler & Powers, 2009).

I have claimed that without statements of purpose, none of our four headline assumptions regarding score meaning make any sense at all. If I were to take a test designed to assess the readiness of well-educated high school students for University entry and use it for any of the three purposes used as examples above, all four of my assumptions cease to be meaningful. This has been a given in test design for 100 years. As Rulon (1946: 290 -291) put it, “...a test cannot be labeled as valid or not valid except for some purpose.” Further, the primary requirement of validation has been enshrined in testing standards such as AERA Standard 1.1 “A rationale should be presented for each recommended interpretation and use of test scores, together with a comprehensive summary of the evidence and theory bearing on the intended use or interpretation” (AERA et al., 1999, p. 17). Within this view, any repurposing requires the production of evidence and rationale to support the new use of the test, as made clear by Standard 1.4: “If a test is used in a way that has not been validated, it is incumbent on the user to justify the new use, collecting new evidence if necessary” (AERA et al., 1999, p. 18). To the extent that there is a mismatch between the old a new purpose, the test providers are required to undertake a change retrofit and develop a new validation narrative/argument to support the new use (Fulcher & Davidson, 2009).

However, the trend to use language testing in policy contexts has led to uncontrolled repurposing, fueled by the commercialization of education and the growth of the migration industry. The use of language tests to implement immigration policies has now become commonplace around the world (Kunnan, 2012). But with very few exceptions (such as the Occupational English Test for health professionals, see McNamara, 1996) tests for these new purposes have not been developed. Most governments have gone for “off the peg” solutions. For example, Read (2001, pp. 193 – 194) makes it clear that while the IELTS test “should have been properly re-validated to establish its appropriateness for the new purpose”, its selection for migration purposes in New Zealand was taken primarily on the basis of lack of cost to the government, compared with designing a local test for this particular purpose.
In recent years the use of the Common European Framework (CEFR) to set “standards” for policy implementation has compounded the problem. The policy makers decide which CEFR level is appropriate for a particular decision (Van Avermaet and Rocco, 2013, p. 15) and through the process of ‘mapping’ an existing test to CEFR levels, any test can claim to be relevant to policy implementation. Similarly, it is possible to state a “level” that is appropriate for entry to a health profession, and use a test of academic English to make that decision based upon a standard setting exercise (O’Neil et al., 2007); or to use a test of general business English to make judgments about the readiness of military personnel for operational deployment (Tannenbaum and Baron, 2010).

This has set the scene for the most profound controversy in modern language testing practice. The IELTS was “...designing specifically to measure the English language skills of candidates intending to study in academic or training contexts in English-speaking countries” (Ingram, 2004, p. 18), not to migrate and work in specific domains. TOEIC was designed as a test of general workplace English (Schmitt, 2005), not to make decisions about communicative readiness in the armed forces. For Ahern (2009), it is the commoditization of testing and education. Shohamy and McNamara (2009) see it as the politicization of test use. For Fulcher (2016) it is the subversion of validation theory, in which “validation” is reinterpreted as “recognition” by political institutions. Research funded by the test owners, on the other hand, merely recommends sensitivity to the needs of policy makers in order to maintain market share (Merrifield, 2012).

The issues at stake could not be greater for the practice of test design. The test design cycle described in this chapter rests upon the assumption that sound inferences are built into tests at the design and development stage: validity through design, told through a validation narrative that tells the story of the decisions made, illustrated with the research and theory that led to design decisions. The process is guided by a vision of the intended use and consequences of the test. Fulcher and Davidson (2007, p. 51) call this “effect-driven testing”, in which “... the ultimate test design decisions are driven by the impacts that the test will have on stakeholders.” This is the meaning of the unbroken cycle: test purpose and design provide the meaning to scores, and justification to decision making. Whereas, in a great deal of modern global practice, meaning is provided post-hoc once tests are co-opted to serve the purposes of new social policy.

### Perspectives

#### Validation Processes

The perspective adopted in this chapter is derived from the tradition that limits score meaning and inference to the primary purpose for which a testing procedure is designed. This requires a statement of purpose, and the construction of a validation narrative around the design cycle. The narrative may then be used in a validation argument (Kane, 2013) to justify the claimed meaning of the score. For some tests, like the Test of English as a Foreign Language (TOEFL iBT) this is set out explicitly (Chapelle, 2008), and contains two components.
The first component is the confirmatory research. That is, all the research focuses on investigating and presenting the elements of the validation narrative that support the intended score meaning. During the early stages of the test design cycle the function of research is to support design decisions and confirm those decisions as being useful for test purpose. The second component is non-confirmatory research. Usually carried out during piloting and field testing and later during live testing, the function is to question the weakest parts of the validity argument (Haertel, 1999). Weaknesses lie in any claim that the score is affected by any factor that is not relevant to the purpose of the test and its intended construct of assessment. This is summarized using the term construct-irrelevant variance (Messick, 1989). For example, if it is suspected that the content of the test has little relevance to the domain in which the test takers are expected to use language, a comparative study of test tasks, topics, language use, and processes with those of the criterion could be undertaken. If it is found that the test content as set out in the specifications is not sufficiently representative of the criterion, there are grounds for questioning the validity of the test.

This perspective reinforces the view that test design is a research activity, rather than a mundane task that practitioners are expected to undertake with little time or resources.

**Substantive Validation**

Test designers usually assume that test takers arrive at a correct answer, or perform in particular ways, for the reasons the designer had in mind when a task was created. We often take it for granted that we have the ability to “see into the mind” of the test taker. The rationale for prototyping is to check with real test takers whether or not this is the case. The fundamental question of substantive validation is: does the test taker respond to the task in the way that the task designer thinks they respond? The methodology of choice is concurrent or retrospective protocol analysis, or PA for short (Bowles, 2010).

The purpose of these studies during the design process is to build in non-confirmatory research that challenges the assumptions of the task designers. In the published literature one of the most informative PA studies is that of Buck (1991), in which he presented a series of listening texts and tasks to learners to discover how they responded. When analyzing the verbal protocols Buck discovered that there were radically different interpretations of questions depending upon cultural background and personal knowledge. Construct irrelevant factors could lead to “incorrect” responses for perfectly good reasons. One illustrative example will suffice.

<table>
<thead>
<tr>
<th><strong>Text Extract</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>My friend Susan was living in West Africa and while she was living there she had a problem with burglars. For a period of about two months every Sunday night someone was breaking into her house through her bedroom window and was stealing something very small from her house and she tried many things to prevent this from happening like putting bars over the windows and hiring a guard to watch the window. And still, every Sunday night, somehow someone came through the window and stole something....</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Question</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>What did Susan do about the problem?</td>
</tr>
</tbody>
</table>
One student responded: “she employed a guardman and jammed the window closed with a bar” (Buck, 1991, p. 74). The second part of this response would have been considered “incorrect”, but the protocol revealed that the Japanese participant was only familiar with sliding windows and had imagined that Susan had used something to stop them from opening. If this problem is present with young adults, it is exacerbated when attempting to assess young learners, as in the following example.

Which is the odd one out?
(a) Eggs
(b) Flowers
(c) Vegetables
(d) Trees

Pupil: Trees.
Teacher: Why do you say that?
Pupil: They’re the only ones I can’t put in a fridge.

Social Responsibility and Accountability

The statistical evidence in piloting also allows researchers to check that no tasks discriminate against an identifiable (or protected) sub-group of the population. All test designers and users should be aware of the fact that if scores are sensitive to age, gender, disability, race, religion, and a number of other protected characteristics, litigation is a potential consequence of the use of non-valid scores (Fulcher, 2014). The practical lesson is that as the stakes become higher, the amount of effort put into deliberate design, research, and validation, becomes more onerous.

Conclusions and Future Directions

Much of our current practice draws directly on innovations that took place during the First World War (Fulcher, 2012). These practices have evolved slowly over the last century, and been documented in guidelines and textbooks for each generation. These practices are reflected in the test design cycle discussed in this chapter. Two issues are of particular importance at the present time, and will occupy researchers for some time to come.

The first is the ecological sensitivity of a test or assessment process to the communities that are impacted by its use. On question we must ask is whether tests are better developed locally by educational practitioners, or provided by large testing agencies that may not be aware of the aware of local contexts within which their tests are used. Part of the question is political, and involves asking whether local or external mandates best serve the assessment needs of particular communities (Fulcher, 2012, pp. 2 – 3). Of course, stakeholder views may always be sought, and local involvement in large-scale testing studies of washback and impact may influence future revisions of a test (e.g. Green, 2007). But this is a matter of research, perhaps consultation, but not democratic involvement (Shohamy, 2001). Next is an epistemic question that requires research. To what extent might locally designed tests prove more or less useful for institutions, regions or countries, compared
with international tests, for their own particular decision making contexts? This is a key validity question (Messick, 1989, pp. 14 – 15), and must be investigated for each context of use. Nevertheless, generalized models of investigation may emerge, but these do not as yet exist.

The second under-researched area is the role of test design and development in the continuing professional development (CPD) of teachers. Davidson and Lynch (2002, pp. 98 -120) were the first to consider the group dynamics of teachers in test design contexts, looking particularly at the decision making processes involved in constructing test specifications. They showed that spec development workshops generated an environment in which individuals negotiated their understanding of constructs and learning objectives in order to converge at a common understanding of constructs of value, and assessment strategies. Within language programmes at particular institutions the professional development benefits of local test design may be immense, through agreeing upon constructs, to designing tasks that can be shared for teaching and assessment, and constructing common assessment criteria. One example of how such collaboration fosters professional development is provided by Fulcher (2010, pp. 159 – 171). The transcript of a group discussion on the value of a particular task shows how a common understanding of its suitability is arrived at. The group actually decides that the task, at first considered innovative, in fact does not appropriately assess the listening construct in which they are interested. The task is discarded. The field does not currently have sufficient studies of the dynamics of test design, how designers arrive at decisions, and how the process supports improved pedagogy.

Practical test design requires human decisions informed by empirical research and theoretical understanding. The narrative of the decisions backed up by the research and rationales constitutes a validity argument by design. The narrative makes clear the claims we wish to make for score meaning, and the use to which scores are put. The process of test design is one of discovery. It is a research enterprise, but one that requires group cohesion and interaction which may, in and of itself, be extremely valuable in professional development and team building. The end product is ideally a test that is sensitive to the constructs of interest, relevant to the learners and the domains in which they will use language, and useful for decision making.

Acknowledgements

I would like to thank Dr. Daniel Pead of the University of Nottingham and Bowland Maths for permission to reproduce the School Trip task.

References


