Language Testing 2009 26 (1) 123–144

Test architecture, test retrofit

Glenn Fulcher *University of Leicester, UK*
and
Fred Davidson *University of Illinois at Urbana-Champaign, USA*

Just like buildings, tests are designed and built for specific purposes, people, and uses. However, both buildings and tests grow and change over time as the needs of their users change. Sometimes, they are also both used for purposes other than those intended in the original designs. This paper explores architecture as a metaphor for language test development. Firstly, it describes test purpose and use, and how this affects test design. Secondly, it describes and illustrates the layers of test architecture and design. Thirdly, it discusses the concept of test retrofit, which is the process of altering the test after it has been put into operational use. We argue that there are two types of test retrofit: an upgrade and a change. Each type of retrofit implies changes to layers of the test architecture which must be articulated for a validity argument to be constructed and evaluated. As is true in architecture, we argue that a failure to be explicit about retrofit seriously limits validity claims and clouds issues surrounding the intended effect of the test upon users.

Keywords: language test validation, language testing, test architecture, test design, test retrofit

When architects design buildings and choose the materials they intend to use in construction, they normally know what a building is going to be used for, and therefore design it to meet the specific needs of its intended occupants. The purpose is defined at the outset of the design project.

Similarly, in language testing, a statement of test purpose is likely to include information on the target population and its ability range. Test developers normally state target domains of language use, and the range of knowledge, skills or abilities that underpin the test. This statement justifies the selection of constructs and content by articulating a

Address for correspondence: Glenn Fulcher, School of Education, University of Leicester, 21 University Road, Leicester LE1 7RF, UK; email: gf39@le.ac.uk

DOI:10.1177/0265532208097339

direct link between intended score meaning and the use to which the scores will be put in decision making. Like architects, test designers imagine the intended effect they wish the test to have, and design with the intended effect in mind (Davidson & Fulcher, 2007; Fulcher & Davidson, 2007).

Over time, however, buildings frequently undergo alterations and changes as they grow to meet the changing requirements of the occupants (Brand, 1994). These alterations are frequently termed *retrofits*. In architecture, a retrofit may be initiated to meet new design standards, introduce safety features unknown when a building was originally constructed, make equipment work more efficiently, or to make a structure fit for a new use or a new user. Test design is no different in principle, and we can identify two distinct types of retrofit. The first is an *upgrade retrofit*, the purpose of which is to make an existing test more suitable for its original stated purpose, by ensuring that it meets new or evolving standards, or uses new technologies to make the test more efficient. For example, as time passes our knowledge of the target domain grows and this may require the inclusion of new item types; a situation may arise in which it is necessary to increase test reliability; or there is a significant shift in the test taking population that requires test difficulty to be adjusted. Such events will require a test upgrade retrofit. The second is a *change retrofit*, in which the test is altered to meet a completely new purpose for which it was not originally intended, or to be used with users who were not envisaged in the original statement of test purpose.

In this paper we consider the role of test purpose, test use, and intended effect, to show how these are intricately connected with the process of test design, and the very structure or architecture of a test. We elaborate on the layers of test architecture and how they are related to validity. We describe upgrade and change retrofits in detail, and outline a process parallel to that used in architecture, adapted to initiate and carry through a test retrofit project.

I Test purpose and test use

1 Purpose, use and validity

Scores on language tests are used to make decisions, and test design needs to be closely aligned to the types of decisions that need to be made. A definition of test purpose needs to take into account the effect the test is intended to have in the real world, and so is very different from the traditional classification of tests into 'types': placement, achievement, progress, proficiency and diagnostic. Rather, test purpose in terms of effect should be related to something much more specific (Cronbach, 1984, p. 122), because unless intended score meaning is explicitly and carefully linked to test design, it is extremely difficult to demonstrate the link between the users' interpretation of the score and the decisions that they take on the basis of the score.

There are two other extremely good reasons for taking purpose seriously. First, as Cronbach (1984, p. 122) claimed: 'No test can put all desirable qualities into one test. A design feature that improves the test in one respect generally sacrifices some other quality.' If a test producer wishes to have a test that can fulfil any purpose, we have *design chaos*. Second, a test that does not have an explicitly defined purpose also creates *validity chaos* (Chalhoub-Deville & Fulcher, 2003, p. 502). It becomes impossible to decide what validity evidence should be collected in support of a particular score interpretation. And even if it is possible to list some validity evidence that might be collected using a 'checklist' approach, it is impossible to prioritize validity research. It is only through specifying purpose closely that we can create validity arguments that focus our attention on the validity questions that are relevant to a particular test, thus allowing us to make the best use of resources in validity enquiry (Haertel, 1999; Kane, 2006).

Tests that do not state purpose are as useless in decision making as are buildings that are designed without users in mind. Such buildings are hard to find, but one example is the London Dome, designed putatively for any purpose – it is basically a large empty space. After one exhibition, it has remained empty and unused for the best part of a decade. Suggestions for use include a sports stadium and a casino, both of which will need in excess of £200 million of investment to retrofit the structure for a specific purpose and particular users.

2 *Standards documents*

This interpretation of test purpose and use is supported by the most widely accepted standards for test design and use (AERA, 1999, p. 17), where it is stated that 'No test is valid for all purposes or in all situations.' It is therefore important to look at what test providers actually claim and do. The reason for this is the fundamental tension that is bound to exist within testing organizations. Pulling in one direction is the professional requirement to restrict score interpretation to those uses for which the test was designed and for which there is validity evidence (as shown in institutional documentation, such as ALTE, 1994 and ETS 2002). Pulling in the other direction is the requirement to generate revenue through increased testing volume, and the temptation to extend the use of a test without undertaking expensive retrofits.

II Describing test architecture

1 Layers of architectural documentation

Getting to the point at which it is possible to create a test involves passing through three main layers of architectural documentation, as shown in Figure 1, each of which embodies a separate set of design issues. The higher layers of the architectural documentation are generalized and can be applied across different tests, while other layers are unique to specific test purposes and contexts of test use. The notion of 'layering' draws upon the architectural work of Duffy (1990) and Brand (1994, pp. 12–23), in which it is argued that design layers should be modular and independent (Brand, 1994, p. 20); this analogy is widely used in other industries such as software development (Bachman et al., 2000) and in testing has already been exploited by Mislevy & Riconscente (2005, p. 3), on the grounds that:

> The compelling rationale for thinking in terms of layers is that within complex processes it is often possible to identify subsystems, whose individual components are better handled at the subsystem level ... Although certain processes and constraints are in place within each layer, cross-layer communication is limited and tuned to the demands of the overall goal.

In test design, as in architecture, this makes it possible to articulate design decisions with a large degree of clarity, as we shall see below.

Models are the most general documents, providing a theoretical overview of what we understand by what it means to know and use a language, although some models (some of which are misleadingly called 'frameworks') try to be encyclopedic, by adding details of possible contexts for communication and sometimes performance conditions. The most well known of these are the Canadian Language Benchmarks (Pawlikowska-Smith, 2000) and the Common European Framework of Reference (Council of Europe, 2001).

Models do not deal with test purpose in any way, but can act as heuristics for what we might put in a test once we have a purpose. Models can be likened to the notion of 'design patterns', which derive from Alexander's (1977) work on construction. Alexander

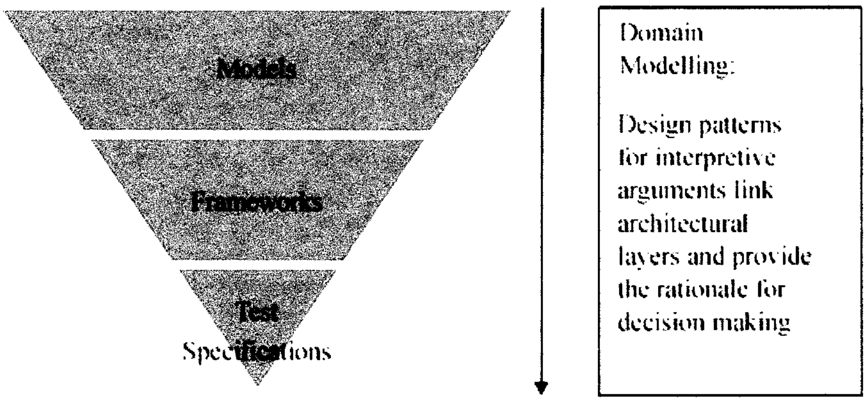
Models
Frameworks
Test Specifications
Domain Modelling:
Design patterns for interpretive arguments link architectural layers and provide the rationale for decision making

Figure 1 Three main layers of architectural documentation.

argues that there are typical design problems that are encountered in many different types of construction, and typical design patterns may be picked up and used in different contexts. In Evidence Centred Design (ECD) (Mislevy, Almond & Lukas, 2003), design patterns are most closely aligned with the layer of *domain modelling*, which attempts to define, as fully as possible, the knowledge, strategies, skills and abilities that are required for successful performance in a domain of interest.

In the next main layer of design, a test framework document is built to mediate between a model and a specification (Chalhoub-Deville, 1997; Fulcher, 2008). This document states test purpose for a particular context of score use. It lays out the constructs to be tested, selected from models, because they are shown to be *relevant* to the specific context in question, and *useful* in the decisions that need to be made. These all contribute to the intended effect of the test. The process of selection limits the purpose of the test, places boundaries upon the claims that can be associated with test scores, and removes design and validity chaos.

At this layer we move away from design patterns that can be selected and used in any test, to instantiations within a particular test. It is most closely associated with the ECD elements of the Conceptual Assessment Framework called the *student model*, the *evidence model* and the *task model*[1] that we describe below. These are sub-layers within the framework layer.

[1] The Conceptual Assessment Framework within ECD contains six 'models' or 'design objects', three of which we argue operate at the level of framework, and three at the level of overall test specification (Mislevy, 2003a, 2003b).

The ECD *Student Model* describes the construct. However, in Mislevy's work, the student model also refers to the 'picture' of an individual test taker that is built up through responses to test items; we have to imagine that each test taker is 'mapped' onto the construct framework as a clearer picture of an individual's abilities emerges. It is particularly easy to see this working in a computer adaptive test, as the measurement model and delivery model 'home in' on the ability of the test taker. This part of the notion of a *student model* is more associated with the practical activity of scoring and interpretation of test scores in use, rather than test design. As such, we will use the term *construct framework* in order to bring it in line with the second level of architectural documentation described above.

Evidence Models describe the evidence we need to collect to make inferences from performance to the constructs. The evidence is what the student actually does. The evidence model should also contain a *measurement component*, which states how the observation is turned into a score.

Task Models describe the items or tasks that elicit the evidence we need to generate scores. The components of a task model are the *presentation material*, or input, the *work products*, or what the test takers actually do, and finally the *task model variables* that describe how an item or task may change, and what alterations might make it more or less difficult. Task models for a specific test are selected from those in the domain analysis because of their relevance to score interpretation and intended test effect.

At the next layer we reach the test specifications, where we find the detail that is specific to a particular test for use in the context specified in the framework. It is not surprising that specifications are also referred to as 'blueprints', for they are literally architectural drawings for test construction. Test specifications are constructed in four sub-layers, as demonstrated in Figure 2.

The first sub-layer is the set of specifications that describe the items or tasks, and any material such as input texts, upon which they depend. Typically, a specification at this sub-level contains two key elements: samples of the tasks to be produced, and guiding language that details all information necessary to produce the task (Davidson & Lynch, 2002, p. 14). The guiding language at the level of tasks summarizes the relevant elements of the construct framework which the designers claim are being measured by a specific item or task type, and the evidence it is designed to elicit. Of particular importance is guiding language about which item features must remain 'fixed' for every item produced, and which are 'free' or allowed to vary. These fixed and free

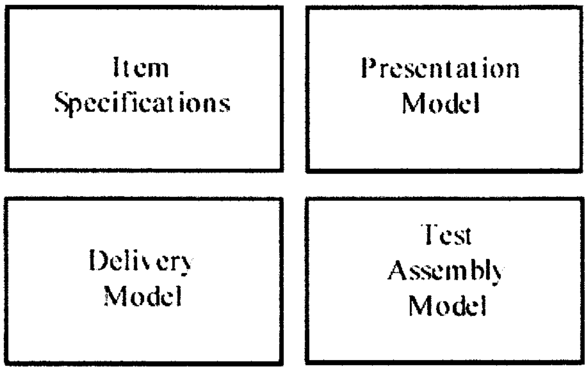

Figure 2 Four sub-layers of test specifications.

elements in specifications (Fulcher, 2003, pp. 135–136) define respectively, which item features are held steady because, while they may affect scores, they are not relevant to the intended universe of generalization, and the features that must vary in order to claim that items are representative of the universe of generalization. These latter features provide meaning to scores (Kane, 2006, p. 35) in terms of the domain, as defined in the framework layer.

We now turn our attention to the conceptual elements of test specifications (Mislevy et al., 2003; Mislevy, 2003a). A given specification may or may not label these elements, but in a well-constructed specification each is needed.

The *Presentation Model* tells us how the items and tasks are presented to the test takers. An *Assembly Model* tells the test designers how the tasks and items should be combined to produce a test form. It specifies *targets*, such as the reliability with which each construct should be measured, and the *constraints* on the mix of items that need to be included to achieve an adequate representation of the domain of inference. Finally, the *Delivery Model* explains how the actual test is delivered, including administration, security and timing.

Many specifications are needed to create a test, at least one for each of the different components (sections, tasks, items) from which the whole is constructed. On top of this are the test specifications which additionally include information on presentation, assembly and delivery.

2 Domain modelling and validity arguments

A second type of design pattern is described by Mislevy and Riconscente (2005, p. 10) as 'articulat[ing] the argument that connects observations of students' actions in various situations to inferences about what they know or can do'. We believe this refers to design patterns that can be used in constructing an interpretive argument that sets out the claims for the meaning of test scores or other outcomes, and the justification for these claims. They are therefore design patterns that link the different architectural layers and provide the rationale for the design decisions taken in a test development pro-ject. In Kane's (2006) terms, these design patterns describe the nature of the interpretative argument, which specifies the 'proposed interpretations and uses of test results by laying out the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances' (Kane, 2006, p. 23).

One design pattern that has been employed to achieve this is Toulmin's (2003, p. 97) 'layout of arguments', as presented in Figure 3. This design pattern has been used in ECD (Mislevy, 2003a, 2003b; Mislevy & Riconscente, 2005), by Bachman (2005) to articulate a language test use argument, by Kane (2006, pp. 27–29) as a template for interpretative and validity arguments, by Fulcher and Davidson (2007, pp. 162–175) to structure interpretative arguments at the level of items or tasks, and validity arguments for whole tests, and by Chapelle (2008) to create a validity argument for TOEFL iBT.

In this design pattern C represents the claim that we wish to make about score meaning, either in relation to an item type or a test, while D represents the data upon which this claim is based. Q is a modal qualifier which indicates the strength of the claim being made, and can take the form of 'always' for a necessity, or some other level of probability (Toulmin, 2003, p. 93). W represents the warrant, or the rationales upon which we assert that the data support the claim

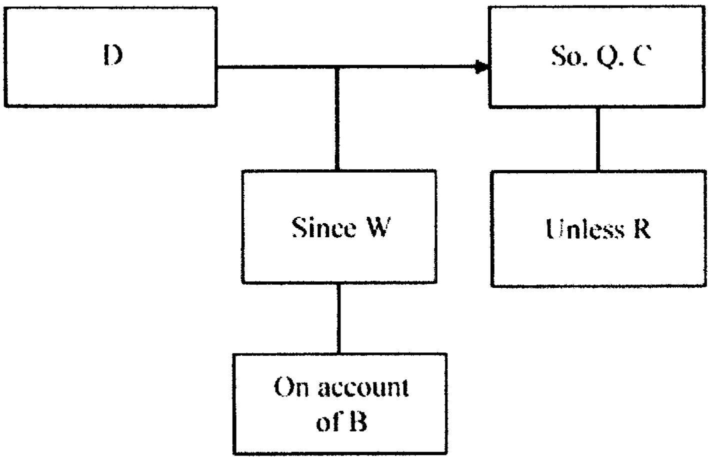

Figure 3 Toulmin's argument structure.

with the degree of probability stated. Warrants are supported by backing, which is symbolized by B. The backing provides any evidence which is needed to support warrants (Toulmin, 2003, p. 96), and may include citations of completed research that justifies the theoretical statements we accept to justify the claims we make from test data. Finally, R symbolizes a potential rebuttal or challenge to the claim, which may be interpreted as an alternative hypothesis or validity challenge (Fulcher & Davidson, 2007, pp. 169–172).

In language testing the claim we wish to make is about the knowledge, skill or ability of a test taker on a construct of relevance, as selected from a model and articulated in a framework. The data is collected from the specific items and tasks as described in the test specifications. The rationales for the link between items and tasks, responses to these items and tasks, and the claim for score meaning, is provided in the warrants. Any evidence that we have to support the warrants is presented in the backing, and the rebuttal shows that we are aware of, and prepared to take into consideration, challenges to the validity of the claims.

We will refer to this layer in Kane's terminology as the *interpretative argument.*

III Test retrofit and design layers

1 Design layers in architecture

The argument of this paper has been that test purpose and test design are inextricably linked. The architecture of the test and the way in which its items and tasks are engineered from prototyping to field testing cannot be separated from test purpose, use, and intended effect. Only if such links exist can there be an argument that ties score meaning to the score interpretations and decisions that we wish to make.

Conceiving test architecture in this way allows us to investigate more closely the notion of test *retrofit.* Retrofits may be classified into upgrade and change retrofits, which may be explained in terms of the extent to which they require alterations to specific design layers. The difficulty of retrofitting both buildings and tests is directly related to what changes we wish to make, and why. Brand (1994, pp. 13–14) discusses the six design layers in architecture, as presented in Figure 4.

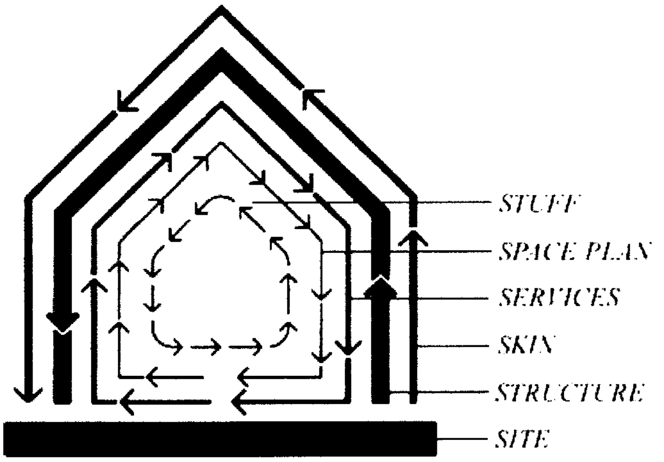

Figure 4 Layers of change (from Brand, 1994, p. 13).

The site upon which a building is placed is the most stable and difficult to change. The structure is comprised of the foundations and the 'load bearing elements', each of which hold the building up. Changing either of these can be dangerous, and once put in place, it is unlikely that either is altered. The skin is the exterior surface and is changed from time to time as fashions change, or as technology improves on insulation. Services include wiring, plumbing, heating, ventilation, lifts and so on. These are replaced more frequently in order to keep the building usable. The space plan is the interior layout, and this can change fairly regularly depending upon the changing needs of occupants. Finally, 'stuff' can change on a very regularly basis, as occupants move objects around to suit their working needs.

It is possible to plot layers of test architecture against Brand's layers, in terms of the functions described. Models tend not to change at all, and when they do, it is most frequently a reclassificatory activity that does not have a major impact on underlying theory (Fulcher & Davidson, 2007, pp. 36–51). This may be seen as the site, and it is almost never the case that a language test moves off the site, or changes its shape. When changes do begin to occur, as in the introduction of the relatively new construct of 'interactional competence', the timeline from the first introduction (Kramsch, 1986) to tentative operational descriptions (Chalhoub-Deville, 2003) is very long. The impact on tests remains negligible, and encroaches on design decisions only very slowly.

The structure is most similar to the construct framework in language testing. It is the fundamental statement of test purpose and effect that *is* the test (Fulcher & Davidson, 2007, p. 177). The structure encompasses the evidence model and the task model (that is, the definition of tasks in the target domain, not the item/task specifications for the test).

The structure also includes the name of the test, which is the brand or trade mark of its purpose, and iconic scores associated with the test, the meaning of which cannot easily be changed (2007, pp. 92–93).

It is easier to change the skin, but it is done very infrequently, and only when necessary. The parallel in language testing is the presen tation and delivery models. How the test appears to the test taker, and how it is delivered (e.g. paper based or computer based), is relatively stable. Changes at this layer, like the introduction of the TOEFL CBT in 1998, usually have high visibility and so are associated with significant unease in the test taking population, score users, and the media.

Most, if not all, retrofits, involve some alteration to the interpretive argument. This is the service layer that provides the rationale for the operation and inter-relationship of all other layers. Major changes are required when there is a change to the structure or skin, and more modest changes when there is a change to the space plan or 'stuff'. The space plan can be changed much more readily, as this is parallel to the test assembly model. It is relatively easy to introduce more item types to better represent a domain, or remove them if a domain is well represented and the number of items measure that domain reliably. Similarly, text lengths can be changed relatively easily, as long as there is evidence that this does not markedly change task difficulty. Finally, the equivalent of 'stuff' is the test items. While we would not wish to suggest that these can be moved around and changed in the same way that workplace objects are moved, they are extremely flexible. It is very common for items to be changed or altered, and for new items to be introduced to the test, in the ongoing evolution of the test specifications.

2 Architecture in practice

Incorporating the layers of architecture described above, we can now present a model of test architecture that expands upon Mislevy's (2003a, p. 5) schematic for ECD to incorporate design components at the level of tasks and items, as well as indicating which of these components are the most difficult to retrofit. Figure 5 shows how test specifications form a critical part of test architecture that relate design decisions to test purpose.

3 Upgrade retrofits

Tests are also frequently upgraded with new item types in order to better represent the constructs or the domain as research tells us

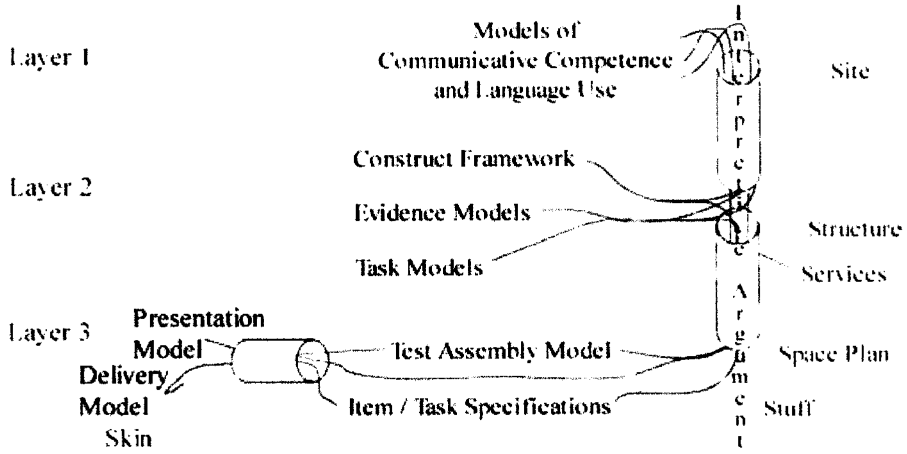

Figure 5 Test architecture.

more about how the items are working in the test. This is easily done, and requires that the designers prototype and pre-test new items to ensure that they are useful additions to the test. Adding 'stuff' may require little or no change to other layers, unless a new item-level interpretive argument is required. Changes to item features are also made when there has been a shift in the ability levels of the test taking population and test difficulty needs to be altered.

Regrettably, it has not been common for language test providers to document upgrade retrofits and place the information in the public domain. Two exceptions are Weir and Milanovic (2003) documenting changes to the Cambridge Proficiency in English (CPE), and Chapelle, Enright and Jamieson (2008) outlining recent changes to the TOEFL. Weir (2003b, p. 478) concludes the discussion of CPE with the observation:

> Stability is one of the twin pillars of public examinations that are essential if exams are to fulfil the purposes for which they are intended. However, innovation linked to improvement is just as vital if the examination is to keep up with developments and insights available from research in the field.

There could be no clearer statement of the value of the architecture metaphor; and in his account of what future research and innovation is required, Weir focuses entirely upon researching new item types to better represent constructs, and improving reliability over time. Each chapter in Weir and Milanovic (2003) considers innovation exclusively at this layer of architecture, although Weir (2003a) does recount the rare occasions upon which changes were made at other layers, including the removal of the phonetics paper in 1931 (2003a, p. 3), and the addition of the listening paper in 1966 (2003a, p. 24). It is arguable that these were rare changes in structure to the construct framework, but care was taken not to allow this to impact upon the intended effect of the test or the meaning of scores. Slightly more frequently, but still not regularly, there have been changes to the test assembly model (space plan) through changes in timing and item numbers, most significantly in 1975 (2003a, p. 33) when the length of the test was reduced by almost one quarter; however, smaller alterations to the configuration of input (text) length, number of items and timing, have continued in subsequent upgrades.

A further example of a rare documented retrofit to the structure of a test is provided by Alderson (1991; 2000, p. 284) in relation to the IELTS Revision Project. In this case it was noticed that scale descriptors for reporting what students were able to do with the language had little connection with the test specifications (2000, p. 75). This meant that the claims being made about what a test taker could actually do, bore no relationship to what the test taker was actually asked to do on the test. This was essentially a discovery of a fault in the structure of the test. The upgrade retrofit was designed to introduce methods of reporting test scores that avoided this problem through redeveloping the wording of rating scale descriptors. Yet, it was not possible to introduce a different reporting scale from the well-known nine bands, or change the meaning of those bands, because of the 'intuitive understanding' of academic admissions tutors (2000, pp. 82–83). The changes had to be as cosmetic as possible to avoid structural damage to the test.

Perhaps the largest upgrade refit to any language test was given to the TOEFL prior to its phased roll-out as the TOEFL iBT over 2005–2006. The test had hardly been altered since its introduction in the 1960s (Spolsky, 1995) apart from the addition of writing in the first computer version in 1998, and was being widely criticized by score users for not providing information meaningful to university admissions processes at the turn of the century. The starting point for the retrofit was a complete re-evaluation of the test structure, beginning with a restatement of the constructs in Jamieson, Jones, Kirsch, Mosenthal and Taylor (2000), and a series of skill-focused construct investigations. The iBT is a radically different test from its predecessor, with no layer left unaltered. It was even necessary to change the score scale, but so important was the original score meaning that it had to be preserved through the use of concordance tables, linking the new to the old architecture. In retrofits of this scale the new services, or interpretive arguments, are crucial to explaining how the new architecture will provide a better service to users. This has been extensively researched and plotted using Toulmin's design patterns (Enright, Chapelle & Jamieson, 2007; Chapelle, Enright & Jamieson, 2008).

4 Change retrofits

A change retrofit occurs when the purpose of a building or a test changes, or the purpose is extended to new users and uses. In architecture there is a 'change-of-use process' that has to be undertaken whenever a building is altered or adapted in order to serve a purpose for which it was not initially designed (Henehan, Woodson & Culbert, 2004). In the building industry any such changes need to be properly considered, and planning permission sought. The more radical the proposal, the more it is likely to be investigated. Apart from architectural and technical evaluation of the plans, change retrofit may also require a public inquiry. Significant extensions to a building may impact upon the local environment or way of life. Similarly, in high stakes tests changes or extensions will have an impact upon the test takers who are expected to take a test designed for some other purpose.

Although it has been possible to find descriptions of upgrade retrofits in the language testing literature, there is no documentation relating to change retrofits in the public domain. There has, however, been a tendency for some test providers to state on web sites and in promotional literature the range of users who 'recognize' or 'accept' scores from their tests for unintended uses, but stop short of explicitly endorsing them. Most notable are the uses of scores from tests of academic English for university entrance being applied to employment and promotion in businesses as diverse as tourism and oil exploration. Perhaps the most criticism is being levelled against their use in screening for immigration. These uses imply *minimally* that the test provider, if it is to list these uses or approve them, constructs an interpretive argument for the new purpose and goes on to evaluate the validity argument in the light of the new intended effects of the test upon test users and society. However, as we have tried to show, it is highly unlikely that any such change retrofits could be conducted without making serious alterations to the construct framework, and changes to the structure almost always require a radical redesign of all lower layers.

As in architecture, this is expensive, and so testing agencies sometimes merely imply that the test is appropriate for its new use without giving any consideration at all to the potential impact upon the test takers, or alternatively, they report the alternative uses of their tests (sometimes without comment) as part of the test adoption advertising. Commercial considerations can over-ride the ethical requirement to make both buildings and tests fit for use.

IV Retrofit procedures

A retrofit process may be activated by a regular review of how a test is functioning, often carried out in a five-year review cycle, sometimes at the request of test users, as a result of some defect that leads to perceived or actual damage (such as a legal challenge against the use of scores), or because of a planned expansion of test use.

Architects must follow established procedures in all cases of retrofits, and these processes are even more stringent if a building is listed (has a preservation order placed on it), or is the object of considerable public attention. Similarly, if a test is to change purpose it is essential that proper processes should be followed and the major stakeholders consulted.

The steps for carrying out a retrofit may be outlined as follows:

1. Set up a team of applied linguists and testing experts, including external advisers, to consider the retrofits under consideration. How this is done within Cambridge ESOL is described by Saville (2003, pp. 83–84), and this does include external representatives in a consultative capacity.
2. Assemble documentation: Collect all existing versions of test specifications and statements of test purpose, research conducted on the test, and statistical data. Highlight the studies of data that suggest the test is in need of a retrofit.
3. Address a number of key questions prior to taking action:
    a. Is the retrofit essential? The primary reasons for the retrofit should be clearly and publicly laid out.
    b. Is the retrofit an upgrade, or a change, to the test?[2]
    c. If the test is a change, is the new intended effect so different from the intended effect of the present test that it is more appropriate to design a new test?
4. Assuming that a retrofit is possible, the team then addresses the following issues, and documents decisions:
    a. Do any other tests exist, which have been designed specifically for the new intended effect? If yes, how different are they? Does this indicate the scale of the retrofit project?

[2] Limitations may be applied to a score inference when it is developed, for example, based on the norming group. Subsequent research may reveal that the original limits were too narrow, and that the test may safely be used to make similar decisions about another group of test takers. (see Davidson, 1994). Strictly speaking, this is not a chance of test purpose. Rather, it is a discovery: that the purpose of the test as originally conceived was inaccurately constrained.

   b. Have any other tests been retrofitted for the same new purpose, and if so, what changes have been made to test architecture? How successful have these changes been?
   c. Is the proposal for the change in the test architecture likely to meet the new need? What research will be needed to provide an evidential basis for the new intended inferences?
   d. Consider how you would feel if you were a test taker and had to take the retrofitted test? Would you think that this was fair?
   e. What are the risks or hazards involved with the intended retrofit, with reference to public acceptability, access for test users, and fairness in test effect?
5. Engage in extensive consultations with stakeholders regarding the nature of the intended retrofit, explaining potential benefits (e.g., improved reliability, improved domain coverage, etc.) and possible disadvantages (e.g., extended testing time, increased cost, etc.).
6. Make a 'go-no go' decision to start the retrofit process.
7. Draw up detailed plans for the test retrofit. The plans need to set out precisely what work needs to be done at different layers, what supporting research is required, and the likely timeline for the work to be completed. Resources for the work need to be identified and allocated. The extent of the research and resources required will depend upon the layers at which changes will be required. Note that this part of the process is very similar to the initial creation of the test.
8. Ensure that the retrofit plan conforms with relevant standards documents and published guidelines.
9. Prioritize the research needed to establish the evidential basis for new inferences to be made.
10. Make a public announcement about the intended retrofit stating explicitly the new or extended test purpose if there is to be one, its rationale, and the research to be undertaken to support successful retrofit.
11. At each step in the process, document all decisions, including design decisions, and which pieces of research led to those decisions. Record each version of the evolving test specifications. In testing, as in other applications of the architectural metaphor such as software engineering, this documentation is 'a vehicle for communicating the system's design to interested stakeholders at each stage of its evolution' and provides 'a basis for performing up-front analysis to validate (or uncover deficiencies in) architectural design decisions and refine or alter those decisions where necessary' (Bachman et al., 2000).

Table 1 A sample retrofit table

Category	Problem	Objective	Recommended solution	Layer
1. Reliability	Reliability has decreased due to changes in the test taking population	Increase reliability	Manipulate task features in task specifications to increase reliability (task features)	3 Stuff
2. Validity	Reporting descriptors for scores does not reflect what test takers at those levels can actually do	Make reporting language more informative and relevant to the score users and the decisions they need to make	Identify features of performance at key levels for use in reporting scales (evidence model)	2 Structure
3. Practicality	Testing takes too long and administration costs are rising because of printing and postage charges	Reduce testing time and costs	Produce a computer delivered version of the test (presentation and delivery models) Skin	3 Skin

A useful tool in designing upgrade retrofits is the use of *retrofit tables*, such as Table 1, which is an example of what might be produced by a retrofit team. There may be multiple recommended solutions for each problem, and each should carry its own research agenda to evaluate that solution before a test is retrofitted.

V Retrofit evaluation

Tests should be subject to the same checks as buildings if they are to be retrofitted. The proposed changes should be acceptable to the technical language testing community. And we need to ask whether the changes proposed are ecologically friendly (Messick, 1989, p. 15); that is, whether the new test is going to be welcomed in its new role by those who have used it in the past, and are likely to use it in the future.

Evaluating the success of a change retrofit is more complex than evaluating an upgrade retrofit. At least four questions need to be asked of the retrofit in order to create a validity argument, relating to its relevance, utility, potential unintended consequences, and its sufficiency for the new use (Messick, 1989).

Relevance: Is the test content relevant to the new domain of inference?

Utility: Is the test useful for making decisions, or put another way, does it remain useful? We also need to know how dependable the scores are. What is the probability of high numbers of false positives, or false negatives?

Unintended consequences: Is there bias against particular subgroups of the test taking population because scores are influenced by construct irrelevant factors?

Sufficiency: Can decisions be taken on the basis of the test alone, or should other information be taken into account? What is the desire of the test developers in that regard – do they wish to aim for sufficiency? Would the use of additional evidence reduce the likelihood of false positives and negatives? The precise weighting of the various sources of evidence and the reasons for valuing these sources should be explicitly stated.

It is only possible to evaluate a change retrofit if this information is in the public domain. Ideally, this should take the explicit form of an interpretive argument as outlined in II.2. It is thus possible for the interpretive argument to be defended or challenged through relevant studies, the evidence from which would contribute to a validity argument (Kane 2006, p. 48) that

> begins with an evaluation of the completeness and coherence of the interpretive argument and an evaluation of whether the interpretive argument provides a reasonable explication of the proposed interpretation. If the general form of the interpretive argument is satisfactory, its inferences and assumptions would be evaluated.

VI Conclusion: Reconceptualizing change in language test development

A critical component in any validity argument is the relationship between test purpose, test architecture, the claims that we wish to make about the meaning of test scores, and hence the use of the test for decision making. This means that any test retrofit requires documented changes to the test architecture – and most critically, changes to the test specifications. Some of these changes merely upgrade the test. Others introduce a new test purpose, and may require much more extensive changes to the test architecture. Regardless, the documentation of the changes would present new interpretive arguments that link the features of the test to its new purpose. Test validation *and design* are thus reconceived as an ongoing evolutionary activity (much like specs, themselves), rather than a one-time event.

Why is this important? No test can anticipate all its uses at the time it is first launched. At the same time, it is not feasible to revise a test (or tear down a building) whenever needs change. Buildings and tests must evolve *in situ* if they are going to serve the needs of their users. Brand (1994, p. 17) reasons that 'The quick processes provide originality and challenge; the slow provide continuity and constraint. Buildings steady us, which we can probably use. But if we let our buildings come to a full stop, they stop us.'

Recognizing the need for both *continuity* and *innovation* in the evolution of language tests is extremely important, as is auditing that evolution through architectural documentation or a validity narrative (Fulcher & Davidson, 2007, pp. 318–319).

Why does this not happen? There is a tendency to change or extend test purpose without articulating a new interpretive argument during the process of test retrofit. New users of a test read a validation argument for Purpose 'A' and assume it is legitimate to use the test for Purpose 'B' without opening up the original validation argument for revision. This is flawed. Our paper has appealed to common practice in architecture to ask that we do the same thing architects take for granted: never retrofit a product without an argument that the retrofit is sound.

What should be done? Potential users of tests, teachers and language testers, should look for the evidence of test retrofit and, where it is not available, they should publicly question the activities of the testing agencies who may be simply encouraging wider use of its tests 'for any purpose' in order to increase testing volume.

It is in this spirit that we reject the simplistic term 'repurpose', which is gaining usage in the USA. To repurpose a test may mean any of these forms of retrofit, and minimally, test designers are obligated to state in what kind of retrofit they are engaged. If the retrofitting triggers a change in test purpose, they must then minimally re-write the interpretive argument to defend the new use, and maximally re-design layers of the test architecture to achieve the intended effect. The test retrofitters may come to a painful realization: it is more work to properly retrofit a test than to do the architectural design for a brand new test, from the ground up.

Is all of this new? The language testing and educational measurement literature has long advised us to avoid using any test for a purpose other than it was originally designed for. Nevertheless, this kind of misuse remains common in our field. Our paper is a new metaphorical exploration of the dilemma, which can serve as a guide to test users, teachers and language testers. When retrofit does happen there should be a documentary trail showing that the new use is sound. The absence of such a trail is the most searing flaw in present practice, we believe. Paying serious attention to test architecture and the intimate relationship between test purpose, design and its intended effect, will keep us vigilant, and help us to better serve our ultimate client: the test taker.

Acknowledgements

We would like to thank the anonymous reviewers of this paper for the time they spent in giving extremely valuable feedback and directing us to literature of which we were unaware at the time of writing. We hope that their guidance has resulted in a much improved argument, although any failures remain our responsibility.

VII References

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 71–86). London: Modern English Publications and the British Council.

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Alexander, C. (1977). *A pattern language: Towns, buildings, construction*. Oxford: Oxford University Press.

Association of Language Testers in Europe. (1994). *Code of practice*. University of Cambridge: Cambridge ESOL. Also available online and retrieved 14 May 2006 from: http://www.alte.org/quality_assurance/index.php

American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Authors.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly* 2(1), 1–34.

Bachman, F., Bass, L., Carriere, J., Clements, P., Garlan, D., Ivers, J., Nord, R. & Little, R. (2000). *Software architecture documentation in practice: Documenting architectural layers*. Pittsburgh, PA: Carnegie Mellon Software Engineering Institute.

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P. & Taylor, C. (2000). *TOEFL 2000 framework: A working paper*. Princeton, NJ: Educational Testing Service.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17–64). (4th ed.). New York: National Council on Measurement in Education & Praeger Publishers.

Kramsch, C. (1986). From language proficiency to interactional competence. *Modern Language Journal 70*(4), 366–372.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). New York: Macmillan/American Council on Education,.

Mislevy, R. J. (2003a). *On the structure of educational assessments*. CSE Technical Report 597. Los Angeles, CA: Centre for the Study of Evaluation, CRESST.

Mislevy, R. J. (2003b). *Argument substance and argument structure in educational assessment*. CSE Technical Report 605. Los Angeles, CA: Centre for the Study of Evaluation, CRESST.

Mislevy, R. J., Almond, R. G. & Lukas, J. F. (2003). *A brief introduction to evidence-centred design*. Research Report RR-03–16. Princeton, NJ: Educational Testing Service.

Mislevy, R. J. & Riconscente, M. M. (2005). *Evidence-centred assessment design: Layers, structures, and terminology*. Menlo Park, CA: SRI International.

Pawlikowska-Smith, G. (2000). *Canadian language benchmarks 2000: English as a second language – for adults*, Centre for Canadian Language Benchmarks, Toronto. Available online at: http://www.language.ca/pd fs/clb_adults.pdf, retrieved 6 July 2006.

Saville, N. (2003). The process of test development and revision within UCLES EFL. In C. Weir & M. Milanovic (Eds.), *Continuity and innovation: Revising the Cambridge Proficiency in English Examination 1913–2002* (pp. 57–120). Studies in Language Testing 15. Cambridge: Cambridge University Press.

Spolsky, B. (1995). *Measured words*. Oxford: Oxford University Press.

Toulmin, S. E. (2003). *The uses of argument*. (2nd ed.). Cambridge: Cambridge University Press.

Weir, C. (2003a). A survey of the history of the Certificate of Proficiency in English (CPE) in the twentieth century. In C. Weir & M. Milanovic (Eds.), *Continuity and innovation: Revising the Cambridge Proficiency in English Examination 1913–2002* (pp. 1–56). Studies in Language Testing 15. Cambridge: Cambridge University Press.

Weir, C. (2003b). Conclusions and recommendations. In C. Weir & M. Milanovic (Eds.), *Continuity and innovation: Revising the Cambridge Proficiency in English Examination 1913–2002* (pp. 473–478). Studies in Language Testing 15. Cambridge: Cambridge University Press.

Weir, C. & Milanovic, M. (2003) (Eds.). *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002*. Studies in Language Testing 15. Cambridge: Cambridge University Press.

Brand, S. (1994). *How buildings learn: What happens after they're built.* New York: Penguin.

Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing 14*, 3–22.

Chalhoub-Deville, M. (2003). Second language interaction: current perspectives and future trends. *Language Testing 20*(4), 369–383.

Chalhoub-Deville, M. & Fulcher, G. (2003). The oral proficiency interview: A research agenda. *Foreign Language Annals 36*(4), 498–506.

Chapelle, C. (2008). The TOEFL validity argument. In C. Chapelle, M. Enright, & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319–352). London and New York: Routledge.

Chapelle, C., Enright, M. & Jamieson, J. (2008). *Building a validity argument for the Test of English as a Foreign Language.* London and New York: Routledge.

Council of Europe. (2001), *Common European framework of reference for language learning and teaching*, Cambridge University Press, Cambridge. Available online at: http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf, retrieved 3 May 2004.

Cronbach, L. J. (1984). *Essentials of psychological testing.* Fourth edition. New York: Harper and Row.

Davidson, F. (1994). Norms appropriacy of achievement tests normed on English-speaking children when applied to Spanish-speaking children. *Language Testing 11*(1), 83–95.

Davidson, F. & **Fulcher,** G. (2007). The Common European Framework of Reference (CEFR) and the design of language tests: A matter of effect. *Language Tea*ching *40*(3), 231–241.

Davidson, F. & Lynch, B. K. (2002) *Testcraft. A teacher's guide to writing and using language test specifications.* New Haven and London: Yale University Press.

Duffy, F. (1990). Measuring building performance. *Facilities 8*(5), 17–22.

Educational Testing Service. (2002). *ETS standards for quality and fairness.* Princeton, NJ: ETS.

Enright, M., Chapelle, C. & Jamieson, J. (2007). From validity research to a validity argument. Paper presented at the conference of the European Assessment and Language Testing Association, Sitges, Spain, 15 June.

Fulcher, G. (2003). *Testing second language speaking.* London: Longman/Pearson Education.

Fulcher, G. (2008). Criteria for evaluating language quality. In E. Shohamy (Ed.), *Language testing and assessment* (pp. 157–176). *Encyclopedia of language and education*, Vol 7. New York: Springer.

Fulcher, G. & Davidson, F. (2007). *Language testing and assessment.* London and New York: Routledge.

Haertel, E. H. (1999) Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice 18*(4), 5–9.

Henehan, D. A., Woodson, R. D., & Culbert, S. (2004). *Building change of use: Renovating, adapting, and altering commercial, institutional, and industrial properties.* New York: McGraw-Hill.