Test Architecture

Glenn Fulcher
(University of Leicester, UK)

Abstract

Test design can be compared with the design of buildings and other structures in architecture. Both activities require the development of detailed plans and blueprints that generate the actual buildings or test forms. When the blueprints are created, architects know what to what use the building is going to be put. Without knowledge of purpose they simply would not be able to design a building. Similarly, test designers need to know what inferences we intend to make from scores, and what decisions are to be made on the basis of those scores. Tests without purpose generate validity chaos. Similarly, when buildings change their use, architects must retrofit the building and follow standard procedures to ensure that health and safety regulations are being met, and that the proposed changes make the building fit for its new users. We argue that test designers must follow similar principles if the purpose of a test is to be changed or extended, or used on a group of test takers for whom it was not originally intended. We term this process *test retrofit*, and use the example of immigration testing to illustrate the argument.

Ⅰ. Introduction

Discussion of *purpose* in language testing permeates the literature. However, we have still gone little beyond the descriptive stage of trying to state whether are tests are to be used to assess proficiency, achievement or progress, to make placement decisions, or to provide diagnostic information for use by teachers (Alderson, Clapham and Wall, 1995: 11 – 12). Such gross statements of test purpose no longer provide the level of detail and delicacy that are required for test design decisions. Rather, we need detailed information about the people who are going to take the test, ideally with a great deal of background information about them and their reasons for needing an assessment of their language abilities. Of particular importance is the domain of language use and range of knowledge, skills and abilities, that are relevant to the specific decisions that will be made on the basis of test scores. The specific intended decisions that we wish to make on

the basis of test scores is closely associated with Messick's (1989) notion of consequential validity.

This notion of test purpose is driven by our definition of the intended effect that we intend the test to have when it is first designed. Test design that is oriented to the intended effects of our test is termed "effect driven testing" (Fulcher and Davidson, 2007: 144):

> "The task for the ethical language tester is to look into the test is intended to have, and to structure the test development effect. This is what we refer to as *effect-driven testing*."

Thinking of test design in terms of its intended use and effect upon stakeholders fuses test design, test use, and test ethics. For in specifying what the test is for prior to engaging in test design, we are also stating what the test is not for; it is just as important to specify what tests should not be used for as it is to specify what decisions they can be used to make.

There are very good reasons for addressing this issue at this time. One of the pressing matters facing the language testing community is the establishment of an agreed Code of Practice to complement the Code of Ethics (ILTA, 2000). One of the vexing questions that faces the International Language Testing Association at the present time is the frequent use of tests for a purpose for which the test was not designed, a practice that we have termed "retrofitting". For example, a large scale language test designed for college entrance may be appropriated by the government of a country and used as part of its immigration screening processes. Currently, we have no theoretical or ethical basis from which to adequately address this kind of practice. This paper attempts to fill this particular gap in our theoretical armoury for the analysis of test use.

Ⅱ. Purpose in Language Testing

Since the work of Messick (1989) mainstream validity theory has argued that validity is a property of test scores, expressed in terms of the inferences that we make from scores, and the justification for any decisions that might be made on the basis of those inferences. Messick shifted the focus of our interests and investigations from instruments to scores, and thus towards inferences and decisions. Messick (1989: 14 – 15) observed that "we are thus confronted with the fundamental question of whether the meaning of a

measure is context-specific or whether it generalizes across contexts." If we are to validate "an interpretation of data arising from a specified procedure" and its meaning for a particular measurement context, we must look to the relevance and utility of the inferences we draw from test scores for the specific decisions, in specific contexts, that we wish to make.

The intended inferences and decisions that we wish to make should be stated prior to test design and development and should drive all the design decisions that we make when developing a new test. The alignment of inferences and intended decisions to test design is critical for the documentation of the processes of test design, and the documentation itself forms a key element in any validity argument that justifies the subsequent use of test scores for real-world decision making. This understanding of test purpose is not entirely new, as it has been at the centre of evolving validity theory for a number of years, as this quotation from Cronbach (1984: 122) makes clear:

> "A test is selected for a particular situation and purpose. What tests are pertinent for a psychological examination of a child entering first grade? That depends on what alternative instructional plans the school is prepared to follow. What test of skill in English usage is suitable for surveying a high school class? Those teachers for whom clarify of expression is important will be discontented with a test requiring only that the student choose between grammatically correct and incorrect expressions."

However, the logic of these arguments has not necessarily been pursued in the language testing literature. The conclusion that we need to reach today is that we need to design validity into our test as an a-priori activity by establishing the closest possible link between intended test use and design decisions, and document these for use in validity arguments. Similarly, if we do not know what kinds of uses the test might be put, or if we wish the scores to carry any interpretation that a user may wish to put on them, we have no principled means of engaging in test design to start with, and certainly no means to investigate test validity. We have design-chaos at the front end of the process, and validity-chaos at the end. As Chalhoub-Deville and Fulcher (2003: 502) have pointed out:

> "A claim that a single test can serve limitless functions or purposes is a cause for serious concern. Typically, different test purposes entail different design considerations….If "limitless" applications are intended, it becomes challenging – if not impossible – to delineate and accommodate all the

knowledge, skills and processes salient in different contexts."

Ⅲ. The Metaphor of Architecture

Language testers should design tests with a specific purpose in mind, and work through the design process to the intended effect. Similarly, architects do not design buildings without a specific purpose or intended use in mind. It would be unthinkable for an architect to design a warehouse when the intended use of the building is office space, or a high rise block to function as a garage. Nor should language testers design language tests with limitless intended purposes, as architects do not design buildings that could have endless functions.

In architecture, there is one building that was constructed with no specific function in mind, the Millennium Dome in London. After the exhibition that was housed by the Dome closed at the end of the Millennium festivities, the building has remained empty. Suggestions for new uses include retrofitting the shell as a casino or a sports stadium for the London Olympics in 2012. The cost of the latter retrofit has already been estimated in excess of £200 million. At least in architecture it is clear when a building without purpose cannot be put to any reasonable use, and retrofitting the structure is seen as a challenge; in language testing it is not always so easy to detect a shift in the purpose of the test without the designers paying any attention to its structure, and whether or not it is fit for its new purpose.

Yet, this kind of practice in language testing is fundamentally as unethical as it would be to put a building to new use without a retrofit. The *Standards for Educational and Psychological Testing* (AERA 1999: 17) that we currently accept as the industry standard code of practice clearly state that "no test is valid for all purposes or in all situations." It goes on to state that:

> "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 of interpretation."

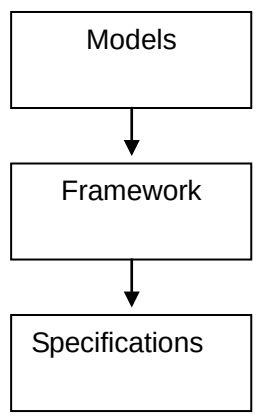
It is therefore clear that any test that has multiple or limitless potential applications must provide multiple of limitless comprehensive summaries of the evidence and theory bearing on those multiple or limitless intended uses. This is not feasible either theoretically or practically, and leads to validity chaos.

For examination boards this argument creates a tension. On the one hand they have a professional responsibility to tell test users what their test should not be used for. On the other, they wish their tests to be applicable to as many uses as possible so that testing volume remains high and test delivery remains not only financially viable, but profitable. For test users, the challenge is to evaluate the relative claims made by examination boards and test providers against the evidence that they make available to support the intended or endorsed uses of a test, which may be extremely wide.

Ⅳ. Architectural Documentation in Language Testing

Designing the architecture for a test that is fit for purpose involves three levels of documentation, as shown in figure 1.

FIGURE 1.
Levels of architectural documentation



Models are the most general level of architectural documentation. These are high level descriptions of language knowledge, skills and abilities described by Canale and Swain (1980), Bachman (1990), Bachman and Palmer (1996), or Celce-Murcia, Dörnyei and Thurrell (1995). Other documents that are called “frameworks” are in fact models, because they are general abstract definitions of what it may mean to know and use a language. However, like the Common European Framework of Reference (Council of Europe, 2001) these attempt to be encyclopaedic rather than entirely abstract. What all models have in common, however, is that they cannot be directly applied to specific language testing contexts without the mediation of other design documents

(Chalhoub-Deville, 1997).

Test Frameworks mediate between models and specifications. They select from the model – the Universe of what we may wish to test – in order to apply parts to a specific testing context. We are not able to test everything in a model, and it is the specific test purpose that provides the guidance on how we select the constructs for a particular test. We do this using two criteria: the *relevance* of the construct to test purpose, and the *usefulness* of the construct to the decisions we wish to make. The purpose of the framework document is to limit the purpose of the test.

At the next level of delicacy we find test specifications, which are the generative blueprint from which we can write test forms. To move from a framework to a specification, test designers ask themselves what is the most efficient and effective way of collecting the evidence that they will need to make judgments about the knowledge, skills or abilities of specific test takers on the constructs selected for inclusion in the test as articulated in the framework. This process is referred to as *operationalizing* the framework.

Ⅴ. The Architecture of Test Specifications

At the most basic level, test specifications provide a method by which we can create parallel test items, and parallel test forms. In architecture the plans for houses and housing estates allow builders to construct identical estates across the country, with houses that have identical room layouts and sizes. Test specifications fulfil this role for language tests, so that it should not matter to a test taker which form of the test, or which items, he or she is presented with in any given administration.

> "The chief tool of language test development is a test specification, which is a generative blueprint from which test items or tasks can be produced. A well-written test specification (or "spec") can generate many equivalent test tasks" (Davidson and Lynch, 2002: 3 – 4).

There are many approaches to writing test specifications (Davidson and Lynch, 2002; Fulcher and Davidson, 2007), but the one that is most clearly articulated is that of Mislevy (2003a; 2003b) in *Evidence Centred Design* (ECD). Mislevy suggests that specifications should contain six components, or *design objects*:

The *student model* draws on models of communicative language competence and use in order to define the constructs for the test. It actually fulfils two different purposes in the work of Mislevy that are not clearly distinguished. On the one hand, it is the statement of the constructs relevant for the purpose of the test, and on the other, it is an abstract model of each student as their abilities are "mapped" onto the construct model. In order to limit the concept to the first of these two uses, we prefer to use the term "construct framework" rather than "student model" for this part of the specification.

An *evidence component* defines the evidence that we need to collect in order to make inferences about the knowledge, skills or abilities of a test taker on the constructs contained in the student model. This is constructed of two parts: a statement of what we expect the student to do in the test, or the "work product", and a measurement component that explains how we move from the work product to a score.

Thirdly we have a set of *task components* that describe the items or tasks that elicit the evidence defined in the evidence component. A task component shares the definition of the work product with the evidence component, but also includes definitions of presentation material, and any task variables that describe how an item or task may change, and how allowable changes might impact upon difficulty.

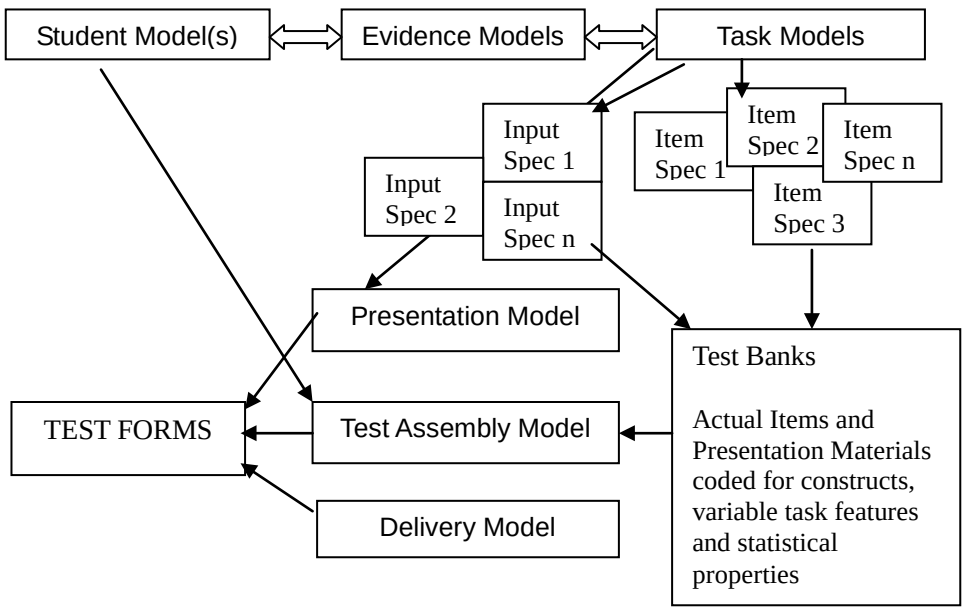
A Presentation component describes how the items and tasks are presented to the test taker; what the test taker would actually see on paper, or on a computer screen. These must be kept stable and familiar, and must not constitute construct irrelevant variance. In computer based testing the interface design is of particular importance, and needs to be researched before a test becomes operational (Chapelle and Douglas, 2006: 63 – 86; Fulcher, 2003b).

The *Assembly component* tells test designers how the tasks and items should be combined to produce a test form. It specifies any pre-set *targets*, like how reliable any particular form must be, and *constraints* on the number of items of particular types in order to achieve adequate representation of a domain of inference, or to ensure that particular items types, perhaps testing one type of knowledge, do not begin to dominate the test.

Finally, the *delivery component* describes how the test is delivered, including administration, security and timing.

A fully developed language test architecture resembles the model in figure 2.

FIGURE 2
A language architecture (Fulcher and Davidson, forthcoming).



In figure 2 we can see that each task or item requires its own specification. Similarly, each type of input requires its own specification. Item writers or individuals employed to construct banks of parallel texts for reading tests, for example, should be able to undertake their work using the specification for guidance. These items and tasks are then collected into banks along with their task and construct codes, and their statistical properties, so that they can be selected for inclusion in a test form.

VI. Test Architecture in Practice

The following test specification (Fulcher and Davidson, forthcoming) illustrates how the approach to test architecture put forward in this paper works in practice. The example is a test specifically designed to assess whether new immigrants to an English speaking country have the reading skills necessary to aid them in the process of social and professional integration. This particular test purpose is chosen deliberately, because many tests that have not been designed for this purpose are now being used for this purpose. We wish to show that the architectural design for such purposes is quite specific, and it is likely to be inappropriate and possibly unethical to use an existing test for this

purpose without giving it an extensive retrofit.

An ECD-style Test Specification: Reading for Immigration: Integrating socially and professionally

Framework

Target Audience

This test is designed for immigrants to English speaking countries with only a high school level of education, expecting to undertake manual and unskilled work. Research evidence suggests that immigrants with a higher level of education, especially those who have already spent time in an English medium environment, are capable of undertaking work and integrating socially. There is also a positive correlation between reading ability and earnings (Chiswick and Miller, 1999; Carnevale, Fry, and Lowell, 2001), and an inverse correlation between the reading ability of immigrants and their ability to access critical services such as health care (Kefalides, 1999; Watters, 2003). The test is also inappropriate for any immigrants who have already spent up to three years in the country prior to testing, as evidence suggests that the language skill differences between those able to work and integrate socially are by this time reduced (Chiswick, Lee and Miller, 2006). Nor is this test suitable for anyone below the age of 18. However, research into the language needs of immigrants is deficient in that it has not made use of needs analysis but relies primarily on survey data that is incapable of describing the kinds of constructs that would be required for test design (Chiswick and Miller, 1998).

General Description

The primary purposes of reading for newly arrived immigrants in English Speaking countries are:

Reading to search for simple information (facts)
Comprehending consequences and reasons

Higher level reading such as synthesizing information from various sources, reading for pleasure, or reading to critique texts, are not relevant to the test purpose. We are concerned with lower-level text processing that supports survival reading in social and

work-seeking contexts.

Student Model

Background knowledge should not play a role in responding to test items.
Subject/topic knowledge is limited to everyday survival tasks for newly arrived immigrants (see Task Model below).
Cultural knowledge should not play a role in responding to test items.
Linguistic knowledge is restricted to basic reading comprehension as follows:

Word recognition/identification

Priority is given to words in the most common 2000, as these are hypothesized to be those that are needed for day-to-day reading needs of new immigrants.

Cohesion

Pronominal Reference (he, they)
Substitution (e.g., The same, one)

Reading for Factual Information

Scan a text to identify a piece of information or a word quickly
Skim prose text to identify a piece of information

Understanding logical sequence clause relations (Winter, 1977)

Cause Consequence: y is the consequence of x

e.g. “high unemployment levels are expected because of the fall in consumer demand” (explicit), “high unemployment levels are expected after a fall in consumer demand” (inexplicit).

Instrument-achievement: By doing X, Y occurs

e.g. “Register with your dentist today. In this way you will get treatment in reasonable time if you experience toothache” (explicit) and “Register with your dentist today and get treatment in reasonable time if you experience toothache” (inexplicit).

Evidence Model

Response attributes are restricted to the selection of a correct response from four options (multiple choice). Selecting the key is taken as an indication that the test taker has an ability to extract basic factual information from a text, understand causes of

consequences, or the reasons for actions. Measurement component: each item is dichotomously scored and items are summed as a measure of “Basic Reading for Immigration”. Scores are scaled using Rasch, and reported on a scale of 0 to 20. Cut scores are to be established by delivering the test to new immigrants judged to be “successful” and “unsuccessful” job-seekers and establishing the score that best reduces numbers of false positives and false negatives.

Task Model and Variables

Presentation Material

All texts should be concrete, not abstract
All texts should be factual, rather than theoretical

Text types and genres

Descriptions	Advertisements, job announcements, notices, signs, directories
Procedures	Public information leaflets, forms
Recounts	News items

More complex text types such as exposition, argument and narrative are excluded. Topics that can be varied may include work, leisure, health, travel, accommodation and shopping. Texts should not be more difficult than Flesch-Kincaid 40, and should not exceed 150 words.

The following is a design template for Text Types with completed data for the sample text below.

Text Title	Tractor Driver	
Text ID	#BR001	
Text Type	A Description	1 Advertisement
		2 Job announcement
		3 Notice
		4 Sign
		5 Directory
	B Procedure	1 Public information leaflet
		2 Form
	C Recount	1 News item
Topic	D Employment	1 Work
	E Social Needs	2 Leisure
		3 Health
		4 Travel
		5 Accommodation
		6 Shopping
Difficulty	Flesch-Kincaid index	43 (Grade 9)
	Word Frequency	1000 74.44% 1001 - 2000: 5.56% (apple, camp, extra, excellent, skills) 2001 – 3000: 8.89% (ASAP, tractor, strawberry, pm, am, hr, mobile, caravan) Off list words: 11.11% (availability, role, tasks, accommodation, site, ethic)

Sample Text

Work Availability:
ASAP until October (may be work after October).

Company Role:
Work is for a strawberry and apple farm as a tractor driver.

Job Description:
Experienced tractor driver needed for various work on the farm. Must have experience

driving modern tractors. Work will include tasks such as bed forming.

Working Day:
7.00 am - 3.00 pm.

Payment:
Depends on experience between £6 - £7 per hr.

Accommodation:
Accommodation on the camp site (mobile home / caravan) this is shared with others.

What to bring:
Good work ethic.

Extra:
Excellent English skills.

Work Products

Below are templates for four item types, each followed by sample items and specification supplements where required.

Item Type 1	Word Recognition			
Item ID	#WR001			
Text ID	#BR001			
Frequency	(A) 1000	(B) 2000	(C) 3000	(D) Off List
Stem	(1) Words from text		(2) Synonyms or paraphrase	
Facility Value	.86			
Discrimination	.45			

Sample Item:

How much you can earn depends on your

(a) English
(b) Experience
(c) Tasks
(d) Driving

Specification Supplement:

The stem may or may not contain words taken from the passage, or use synonyms or paraphrase for words taken from the passage, in the immediate vicinity of the target word. It is expected that stems containing words are likely to be easier, as word recognition will be achieved by matching words in the stem to those in the text, rather than identifying similar meanings.

Item Type 2	Cohesion	
Item ID	#C001	
Text ID	#BR001	
Reference	(A) Pronominal	(B) Substitution
Distance	(1) Closest to Reference	(2) Distant from Reference
Facility Value	.42	
Discrimination	.51	

Sample Item:

What will you share with others?

(a) a caravan.
(b) a mobile home
(c) accommodation
(d) the camp site

Specification Supplement:

The target word should be located within two sentences of the pronominal reference of substitution. Distracters are drawn from noun phrases in the vicinity of the target word, and may be closer to the pronominal reference of substitution.

Item Type 3	Reading for Factual Information	
Item ID	#FI001	
Text ID	#BR001	
Reference	(A) Scan	(B) Skim
Facility Value	.77	
Discrimination	.37	

Sample Item

What time will you finish work each day?

(a) October
(b) 3 o’clock
(c) 6 o’clock
(d) 7 o’clock

Specification Supplement:

Items should focus upon dates, times, numbers, facts, events, or names..

Item Type 4	Understanding Logical Sequence Clauses	
Item ID	#LS001	
Text ID	#BR001	
Clause Type	(A) Cause Consequence	(B) Instrument Achievement
Explicitness	(1) Marked	(2) Unmarked
Facility Value	.33	
Discrimination	.62	

Sample Item

If you take this job, you may

(a) be offered further work
(b) bring a good work ethic
(c) improve your English skills
(d) drive modern tractors

Assembly Model

Targets: This test is intended to be a necessary though not sufficient condition for immigration. That is, scores from the test will form one part of a range of evidence that will be used in decision making. As a high-stakes test it is required that test reliability reach at least .9, requiring that individual items are crafted to a high standard.

Constraints:
Form constraints
Each form contains 50 items attached to 10 texts.
Text constraints
5 description texts, 3 procedure texts and 2 recount texts.
Topic constraints
5 Employment texts and 5 Social Needs texts, one to be drawn from each sub-topic.

Item Constraints

Item Type	Code	No.
Type1	1A1	4
	1A2	3
	1B1	4
	1B2	3
	1C1	1
	1C2	1
	1D1	1
	1D2	1
Type 2	2A1	4
	2A2	2
	2B1	4
	2B2	2
Type 3	3A	5
	3B	5
Type 4	4A1	4
	4A2	1
	4B1	4
	4B2	1

Text/Item constraints
5 items attached to each text.

Delivery Model
Paper based, with one text and five questions appearing on each page. Each text and questions to appear in Times New Roman 12pt. One proctor for every 20 test takers. Delivery time: 75 minutes.

From this example we can see that text purpose and test design are so closely linked that it becomes virtually impossible to decouple them. Further, if we take a test that has been designed with no purpose in mind, or another purpose in mind, and apply it to a different specific purpose, we have no evidence of design for validity built into the test. This must be addressed by the test providers if they are to give any basis whatsoever for

researchers to address the validity of the score inferences from the test to its new intended use.

Ⅶ. Test Retrofit

In architecture there are two kinds of retrofit. The first is an *upgrade retrofit*, which involves improving a building or structure by making it stronger or adding new features. Typical of these improvements are seismic retrofits to make buildings erected before the technology existed resistant to earthquake damage. The purpose of the building or structure is not being changed; rather, the retrofit is making the building more fit for its original purpose. The second kind is *change retrofit*, in which the purpose or use of the building is changed. It is this second case with which we are concerned. In architecture a "change-of-use-process" is required in all such cases (Henehan et al., 2004). Changes need to be considered carefully to ensure that the building is suitable for its new use, and that health and safety regulations are met. Planning permission then has to be sought. A test retrofit should follow the same pattern in order to ensure that the inferences drawn from scores are relevant to the new use, and that fairness to test takers is maintained. The procedure for test retrofit can be modelled on the process undertaken in architecture, and which we summarize in seven steps.

1. Set up a team of experts which should contain external consultants. With all test specifications and other studies on test validity available, ask first: Is the retrofit essential? If the answer is yes, the primary reasons for the retrofit should be clearly and publicly laid out. One of the main reasons for any retrofit is financial. Rather than develop a new test for a new purpose, an old test may be adapted for the new purpose, just as an older building may be retrofitted as an entertainment venue. However, it is necessary to question whether it is better to build a new test. Just as it is not possible to convert an office block into an aircraft hanger, there will be some tests that are simply the wrong shape for the new use.
2. Set up a committee of experts, including external advisers, to consider the main retrofit concerns:
    a. Does the plan confirm with relevant standards documents and published guidelines?
    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?
3. Draw up detailed plans for the test retrofit and incorporate these into a new version of the test specifications.
4. Prioritise the research needed to establish the evidential basis for new inferences to be made.
5. Make a public announcement about the intended retrofit stating explicitly the new or extended test purpose, its rationale, and the research to be undertaken to support successful retrofit.
6. Consult all stakeholders.
7. Make a go-no go decision to start the retrofit process.

As this work is undertaken, the development team should continue to ask whether the retrofitted test is fit for its new purpose. In order to do this we draw on four main criteria. First and foremost, we ask whether the retrofitted test is *relevant* to the new domain of interest. Secondly, we evaluate whether the retrofitted test is *useful* in making the new decisions envisaged in the new test purpose. Thirdly, we consider whether the use of the test might have any *unintended consequences*, such as introducing bias against certain sub-groups of the new test taking population. Fourthly, we ask whether the information provided by the test is both a *necessary* and *sufficient* condition for taking these new decisions.

Ⅷ. Conclusion

It has been argued that without a clear statement of test purpose that limits the usability of test scores, we are faced with validity chaos. One of the reasons for this situation is that there is no way to link design decisions to score use. When tests are marketed as the solution to all language assessment problems, or continually repositioned to fill new markets as they come along, we should be aware that the test providers are not acting ethically. We now know a great deal about consequential validity and test fairness, and records of test design that link purpose to structure are admissible as validity evidence. Further, any attempt to change or extend test purpose without articulating a new validity argument or engaging in the process of test retrofit merely to increase test volume is also unethical.

This paper has tried to show that design issues are closely tied to test purpose, and that score meaning cannot be easily extrapolated to domains of inference that were not envisaged at the time of test design. The concept of test architecture therefore provides the theoretical tools necessary to link fairness firmly to principles of test design.

References

Alderson, J. C. (1991). “Bands and Scores.” In Alderson, J. C. and North, B. (Eds.) Language Testing in the 1990s. London: Modern English Publications and the British Council, 71 – 86.

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

Bachman, L. F. (1990) *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press

Bachman, L. F. and Palmer, A. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Canale, M. and Swain, M. (1980) “Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing.” *Applied Linguistics*, 1, 1, 1 – 47.

Carnevale, A. P., Fry, R. A., and Lowell, L. (2001). “Understanding, Speaking, Reading, Writing, and Earnings in the Immigrant Labor Market.” *The American Economic Review*, 91, 2, 159 – 163.

Celce-Murcia, M., Dörnyei, Z. and Thurrell, S. (1995) “Communicative Competence: A Pedagogically Motivated Model with Content Specifications.” *Issues in Applied Linguistics*, 2, 5 – 35.

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

Chalhoub-Deville, M. and Fulcher, G. (2003). “The Oral Proficiency Interview: A Research Agenda.” *Foreign Language Annals* 36, 4, 498 – 506.

Chapelle, C. and Douglas, D. (2006). *Assessing Language Through Computer Technology*. Cambridge: Cambridge University Press.

Chiswick, B. R. (1991). “Speaking, Reading, and Earnings among Low-Skilled Immigrants. *Journal of Labor Economics* 9, 2, 149 – 170.

Chiswick, B. R. and Miller, P. W. (1998). “Language Skill Definition: A Study of Legalized Aliens.” *International Migration Review* 32, 4, 877 – 90.

Chiswick, B. R. and Miller, P. W. (1999). “Language skills and earnings among legalized aliens.” *Journal of Population Economics*, 12, 1, 63 – 89.

Chiswick, B. R., Lee, Y. L., and Miller, P. W. (2006). “Immigrants’ Languge Skills and Visa Category.” *International Migration Review* 40, 2, 419 – 450.

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. and Lynch, B. K. (2002) *Testcraft. A Teacher's Guide to Writing and Using Language Test Specifications*. New Haven and London: Yale University Press.

Fulcher, G. (2003a) *Testing Second Language Speaking*. London: Longman/Pearson Education.

Fulcher, G. (2003b) "Interface design in computer based language testing." *Language Testing* 20, 4, 384 - 408.

Fulcher, G. and Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book*. London and New York: Routledge.

Fulcher, G. and Davidson, F. (forthcoming). "Test Architecture, Test Retrofit".

Henehan, D. A., Woodson, R. D., and Culbert, S. (2004). *Building Change of Use: Renovating, Adapting, and Altering Commercial, Institutional, and Industrial Properties*. New York: McGraw-Hill.

International Language Testing Association. (2000). *Code of Ethics*. ILTA. Available online at: http://www.iltaonline.com/code.pdf.

Kefalides, P. T. (1999). "Illiteracy: The Silent Barrier to Health Care." *Annals of Internal Medicine* 130, 4/1, 333 – 336.

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

Mislevy, R. J. (2003a) *On the Structure of Educational Assessments*. CSE Technical Report 597. Los Angeles: 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: Centre for the Study of Evaluation, CRESST.

Watters, E. K. (2003). "Literacy for Health: An Interdisciplinary Model." *Journal of Transcultural Nursing* 14, 1, 48 – 53.

Winter, E. (1997). "A clause-relational approach to English texts: A study of some predictive lexical items in written discourse." *Instructional Science* 6, 1, 1 – 92.