VALIDITY AND CAMBRIDGE ESOL EXAMINATIONS

Adapted from:

Studies in Language Testing 29

Examining Reading: Research and practice in assessing second language reading, Khalifa, H & Weir, CJ (2009) Cambridge: Cambridge University Press

Studies in Language Testing 30

Examining Speaking:Research and practice in assessing second language speaking, ed Taylor, L (2010) Cambridge: Cambridge University Press

This text discusses the Cambridge ESOL approach to validation which is represented graphically in the accompanying charts. As can be seen in the charts, the validation process is conceptualised in a *temporal frame* to identify the various types of validity evidence that need to be collected at each stage in the test development, monitoring and evaluation cycle.

The framework is described as *socio-cognitive* in that the abilities to be tested are demonstrated by the mental processing of the candidate (the cognitive dimension); equally, the use of language in performing tasks is viewed as a social rather than a purely linguistic phenomenon, resonating with the CEFR's perspective on language for social purposes which sees the learner (and presumably the test taker) as 'a social agent who needs to be able to perform certain actions in the language' (North 2009:359). The framework represents a unified approach to gathering validation evidence for a test. The accompanying charts depict how validity components fit together both temporally and conceptually (Weir 2005a:43). Conceptualising validity in terms of temporal sequencing is of value as it offers test developers a plan of what should be happening in relation to validation and when it should be happening. The model in the accompanying diagrams comprises both *a priori* (before-the-test event) validation components of *context* and *cognitive validity* and *a posteriori* (after-the-test event) components of *scoring validity*, *consequential validity* and *criterion-related validity*.

This addresses a number of critical questions in applying the socio-cognitive framework to Cambridge ESOL BULATS.

- How are the physical/physiological, psychological and experiential characteristics of candidates catered for by this test? (*test taker*) *For a detailed discussion of ESOL's approach to the test taker see Khalifa & Weir 2009: Chapter 2, Taylor 2010: Chapter 2.*
- Are the cognitive processes required to complete the test tasks appropriate? (*cognitive validity*) *For a detailed discussion of ESOL's approach to cognitive validity see Khalifa & Weir 2009: Chapter 3, Taylor 2010: Chapter 3.*
- Are the characteristics of the test tasks and their administration appropriate and fair to the candidates who are taking them? (*context validity*) *For a detailed discussion of ESOL's approach to context validity see Khalifa & Weir 2009: Chapter 4, Taylor 2010: Chapter 4.*
- How far can we depend on the scores which result from the test? (*scoring validity*) *For a detailed discussion of ESOL's approach to scoring validity see Khalifa & Weir 2009: Chapter 5, Taylor 2010: Chapter 5.*
- What effects do the the tests and test scores have on various stakeholders? (*consequential validity*) *For a detailed discussion of ESOL's approach to consequential validity see Khalifa & Weir 2009: Chapter 6, Taylor 2010: Chapter 6.*
- What external evidence is there the test is measuring the construct of interest? (*criterion-related validity*) *For a detailed discussion of ESOL's approach to criterion-related validity see Khalifa & Weir 2009: Chapter 7, Taylor 2010: Chapter 7.*

These are the types of critical questions that anyone intending to take a particular test or to use scores from that test would be advised to ask of the developers in order to be confident that the nature and quality of the test matches their requirements.

The *Test taker characteristics* in the accompanying charts connects directly to the *Cognitive* and *Context validity* boxes because these individual characteristics will directly impact on the way the individuals process the test task set up by the *Context validity* box. Obviously, the tasks will also be constructed with the overall test population and the target language use situation clearly in mind as well as with concern for their cognitive validity' (Weir 2005a:51). Individual test taker characteristics can be subdivided into three main categories:

- *physical/physiological characteristics* – eg individuals may have special needs that must be accommodated such as visual impairment or a speech impediment
- *psychological characteristics* – eg a test taker's interest or motivation may affect the way a task is managed, or other factors such as preferred learning styles or personality type may have an influence on performance
- Experiential characteristics – eg a test taker's educational and cultural background, experience in preparing and taking examinations as well as familiarity with a particular test may affect the way the task is managed.

All three types of characteristics have the potential to affect test performance.

Cognitive validity is established by *a priori* evidence on the cognitive processing activated by the test task before the live test event (eg through verbal reports from test takers), as well as through the more traditional a posteriori evidence from constructs measured involving statistical analysis of scores following test administration. Language test constructors need to be aware of the established theory relating to the cognitive processing that underpins equivalent operations in real-life use.

The term *content validity* was traditionally used to refer to the content coverage of the task. *Context validity* is preferred here as the more inclusive superordinate which signals the need to consider not just the linguistic content parameters, but also the social and cultural contexts in which the task is performed. Context validity for a speaking test, for instance, thus addresses the particular performance conditions, the setting under which it is to be performed (such as response method, time available, order of tasks as well as the linguistic demands inherent in the successful performance of the task) together with the actual examination conditions resulting from the administrative setting (Weir 2005a).

Scoring validity is linked directly to both context and cognitive validity and is employed as a superordinate term for all aspects of reliability (see Weir 2005a, Khalifa & Weir 2009: Chapter 5, Taylor 2010: Chapter 5). Scoring validity accounts for the extent to which test scores are arrived at through the application of appropriate criteria and rating scales, as well as the extent to which they exhibit agreement, are as free as possible from measurement error, stable over time, appropriate in terms of their content sampling and engender confidence as reliable decision-making indicators.

Messick (1989) argued the case for also considering *consequential validity* in judging the validity of scores on a test. From this point of view it is necessary in validity studies to ascertain whether the social consequences of test interpretation support the intended testing purpose(s) and are consistent with other social values (Taylor 2010: Chapter 6). There is also an important concern here with the *washback* of the test on the teaching and learning that precedes it as well as the impact on institutions and society more broadly (Khalifa & Weir 2009: Chapter 6, Taylor 2010: Chapter 6).

Criterion-related validity is a predominantly quantitative and a posteriori concept, concerned with the extent to which the test scores correlate with a suitable external criterion of performance with established properties (Anastasi 1988:145, Messick 1989:16, Taylor 2010: Chapter 7). Evidence from criterion-related validity can come in three forms.

Firstly, if a relationship can be demonstrated between test scores and an external criterion which is believed to be a measure of the same ability. This type of criterion-related validity is typically subdivided into two forms: *concurrent* and *predictive*. *Concurrent validity* seeks an external indicator that has a proven track record of measuring the ability being tested (Bachman 1990:248). It involves the comparison of the test

scores with this external measure for the same candidates taken at roughly the same time as the test. The external measure may consist of scores from some other tests, or ratings of the candidate by teachers, subject specialists, or other informants (Alderson, Clapham & Wall 1995). Predictive validity entails the comparison of test scores with another measure of the ability of interest for the same candidates taken some time after the test has been given (Alderson et al 1995).

A second source of evidence is demonstration of the qualitative and quantitative equivalence of different forms of the same test, by means of validation studies involving verbal protocol analysis with test takers as they complete test tasks or generalisability analyses comparing performance across tasks.

A third source of evidence results from linking a test to an established external standard, or to an interpretative framework of reference such as the Common European Framework of Reference (CEFR) through the comprehensive and rigorous procedures of familiarisation, specification, standardisation and empirical validation (Council of Europe 2003a). Linking tests to an external standard is not straightforward, however, and the use of the CEFR in this way remains somewhat contentious. Claims about CEFR alignment for any given test needs to be considered with some caution and careful attention needs to be paid to other essential quality aspects of the test in question (Milanovic & Weir 2010).

Although for descriptive purposes the various elements in the model in the accompanying charts are presented as being separate from each other, a closer relationship undoubtedly exists between these elements, for example between context validity and cognitive validity. Decisions taken with regard to parameters in terms of task context will impact on the processing that takes place in task completion. Within the specific context of practical language testing/assessment, there exists a third dimension that cannot be ignored: the process of scoring. In other words, at the heart of any language testing activity we can conceive of a triangular relationship between three critical components:

- the test taker's cognitive abilities
- the task and content
- the scoring process

These three dimensions, which are reflected in the *Cognitive validity*, *Context validity* and *Scoring validity* boxes in the charts offer a perspective on the notion of construct validity which has both sound theoretical and direct practical relevance for test developers and producers. By maintaining a strong focus on these three elements and by undertaking a careful analysis of their tests in relation to these three tests in relation to these three dimensions, test providers should be able to provide theoretical, logical and empirical evidence to support validity claims and arguments about the quality and usefulness of their exams. In addition, the interactions between, and especially within, these aspects of validity may well eventually offer deeper insights into a closer definition of task difficulty. For the purposes of this proposal, the separability of the various aspects of validity will be maintained since they offer a helpful descriptive route through the socio-cognitive framework and the theory that underpins it.

Works referred to:

Alderson, JC, Clapham, C & Wall, D (1996) *Language Test Construction and Evaluation,* Cambridge: Cambridge University Press.

Anastasi, A (1988) *Psychological Testing* (6th Edition), New York: Macmillan.

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

Council of Europe (2003a) *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF), Manual: Preliminary Pilot Version,* DGIV/EDU/LANG 2003, 5 Strasbourg: Language Policy Division.

Messick, SA (1989) Validity, in Linn, RL (Ed), *Educational Measurement* (3rd Edition), Washington DC: The American Council of Education and National Council on Measurement in Education, 13-103.

Milanovic, M & Weir, CJ (2010) Series Editors' note, in Martyniuk, W (Ed) *Aligning Tests with the CEFR: Reflections on Using the Council of Europe's Draft Manual, Studies in Language Testing 33,* Cambridge: UCLES/Cambridge University Press, viii-xx.

North, B (2009) The educational and social impact of the CEFR, in Taylor, L and Weir CJ (Eds) *Language*

Testing Matters: Investigating the Wider Social and Educational Impact of Assessment – Proceedings of the ALTE Cambridge Conference, April 2008, Studies in Language Testing 31, Cambridge: Cambridge University Press, 21-66.

Weir, CJ (2005a) *Language Testing and Validation: An Evidence-Based Approach,* Basingstoke: Palgrave Macmillan.