CONTEMPORARY VALIDITY THEORY AND THE ASSESSMENT CONTEXT IN ENGLAND

Neil Stringer

ABSTRACT

The concept of validity developed considerably during the last half of the twentieth century, from essentially meaning that "a test measures what it says it measures", through multiple types of validity – content, criterion, and construct – to a multifaceted but essentially unitary concept of construct validity. The publishers of high stakes tests in the United States have, for the most part, embraced the modern concept of validity, whilst those responsible for general qualifications in England, such as the GCSE and GCE, appear not to have ventured far beyond evaluating content validity and reliability in ensuring the quality of these tests. These differences may be attributable to differences between the two cultures in the forms of tests and the personnel traditionally responsible for their construction. Nonetheless, the unitary concept of validity demands forms of evidence to counter threats to validity that content validity and reliability do not and, as such, the quality of English general qualifications could benefit from explicit evaluation of validity, especially during specification (syllabus) development. The validity literature contains examples of the types of evidence required to satisfy validity concerns and guidance on how to gather it, on which awarding bodies may draw. The involvement of the regulatory authorities in specification development means, however, that responsibility for validity cannot lie exclusively with the awarding bodies, and a coordinated approach to validation would be required.

INTRODUCTION

In its seventy-two pages, the Qualifications and Curriculum Authority (QCA) Code of Practice for GCSE, GCE, and AEA contains only one occurrence of the word *validity*, which is in the context of testing accommodations[1]: "the awarding body must ensure that its access arrangements…maintain the relevance, validity, reliability, comparability and integrity of the assessment" (Qualifications and Curriculum Authority, 2008a, p. 41). The Assessment and Qualifications Alliance's (AQA) *Question Paper Preparation Procedure Guidance File* (Assessment and Qualifications Alliance, 2007), which comprises two hundred and forty-two pages, contains only four occurrences of the word *validity*. Two of these occurrences are in the glossary, which defines validity as "the fitness for purpose of an assessment tool or scheme", whilst the other occurrences are passing references in specific contexts to the validity of particular tasks or response modes.

Specification development is the stage at which the content of the specification (syllabus) is determined and at which the scheme of assessment is devised and specimen question papers are written, yet AQA's generic *Specification Development Procedure Guidance File* (Assessment and Qualifications Alliance, 2005) does not contain the word *validity* at all. The

[1] Testing accommodations are changes made in the administration of a test for candidates with particular requirements to enable them to have access to fair assessment and to demonstrate attainment.

Copyright © 2011 AQA and its licensors. All rights reserved.

The Assessment and Qualifications Alliance (AQA) is a company limited by guarantee registered in England and Wales (company number 3644723). Registered address: AQA, Devas Street, Manchester M15 6EX

Specification Development Procedure Guidance File GCE 2006/7 for teaching from September 2008 (Assessment and Qualifications Alliance, 2006) contains four instances, not including those already mentioned in the question paper guidance file, the most pertinent, though unsubstantiated, being:

> *Objective Test Questions (OTQs) have some advantages… It is possible to test a wide range of abilities and subject matter in a relatively small amount of examination time and thus objective tests carry a high degree of validity in terms of the relationship between the examination and the specification.* (p. 18)

There is more frequent use of the term *valid* in these documents, but little or no reference to validity and what constitutes it. For example, "the regulators require that the awarding bodies are committed to the following principles that underpin a high-quality assessment system and promote good practice" including that "assessment instruments are fit for purpose, valid and reliable" (Qualifications and Curriculum Authority, 2008a, p. 4). In relation to the use of objective test questions, AQA states that its "policy is to adopt the most appropriate assessment method, which incorporates consideration of the most valid way to assess the subject matter and the efficiency and effectiveness of the marking" (Assessment and Qualifications Alliance, 2006, p. 18).

Although predominantly led by the United States, the last fifty or so years have produced a healthy literature on the concept of validity and test validation. Although not legally binding, the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) is a highly influential publication, through successive editions of which the developments in validity theory in educational measurement can be traced (Moss, 2007)[2]. In the context of English general qualifications, validity appears to mean 'fit for purpose', but there is little evidence of a technical definition beyond that. Are there reasons, technical or cultural, that explain the more explicit emphasis placed on validity by the publishers of tests in the United States, compared with the English awarding bodies? Are English general qualifications behind the times as far as the consideration of validity evidence is concerned, and should the regulator and the awarding bodies be more concerned with it?

THE CONCEPT OF VALIDITY

Thyne (1974) states four necessary and sufficient conditions of validity:

1) marking consistency – "…the measures within one series are *comparable* with one another, in the sense that (for instance) 64 represents a greater *magnitude* than does 63…and so on" (p. 8).

2) mark relevance – "…not only must all the marked performances be relevant ones; all the given relevant performances must be marked" (p. 13).

3) question relevance – "…all the asked questions must be relevant, and all the relevant questions must be asked" (p. 15). In practice, it is likely that a "*sample* from the 'universe' of performances comprising the list…" will have to suffice (p. 16).

[2] See also the chapter on validity in successive editions of *Educational Measurement (e.g. Cronbach, 1971; Messick, 1989b)*.

4) balance – "different systems of 'weighting' the parts of the examination will…produce sets of results which do not correspond perfectly one with another, and so not all the systems of weighting will produce results of maximum validity" (p. 17). The examination's purpose must determine the weights assigned to its various parts.

The various stages of administering an examination specification provide opportunities to satisfy Thyne's conditions. For example:

i. writing the specification may address *question relevance* (or at least the relevance of the material on which questions will be based to the domain being assessed);

ii. devising the scheme of assessment (e.g. question papers, coursework, practical tasks) may address *question relevance* and *balance*;

iii. writing question papers and mark schemes may address *question relevance* and *mark relevance*; and

iv. standardising examiners' marking may address *marking consistency*.

So, validity issues, as identified by Thyne, are addressed in our general qualifications. However, validity theory has advanced since 1974.

Traditionally, several types of validity have been described. *Content validity* concerns "how well the content of a test samples the class of situations or subject matter about which conclusions are to be drawn" (Messick, 1989b, p. 16). *Criterion-related validity* concerns how test scores compare with one or more external variables that provide a direct measure of the characteristic or behaviour being tested. It can relate to an individual's current or future level on the criterion, providing *concurrent validity* or *predictive validity*, respectively. *Construct validity* concerns "the degree to which certain explanatory concepts or constructs account for performance on the test" (Messick, 1989b, p. 16). These *types of validity* are now more commonly regarded as *types of validity evidence* that are subsumed by the unitary concept of construct validity (e.g. American Educational Research Association et al., 1999; Messick, 1989b).

Messick defines construct validity as comprising "the evidence and rationales supporting the trustworthiness of score interpretation in terms of explanatory concepts that account for both test performance and relationships with other variables" (1989b, p. 34). For Messick, validity is "an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (1989b, p. 13). Thus, it is not the instrument that is valid; rather the inferences derived from test scores or other indicators. Messick regards validation as scientific enquiry and validity as a unitary, though faceted (see Table 1), concept, with the traditional validity types more appropriately regarded as categories of validity evidence. Similarly, Sireci (2007, p. 477) concludes that:

- *Validity is not a property of a test. Rather, it refers to the use of a test for a particular purpose.*
- *To evaluate the utility and appropriateness of a test for a particular purpose requires multiple sources of evidence.*
- *If the use of a test is to be defensible for a particular purpose, sufficient evidence must be put forward to defend the use of the test for that purpose.*

- *Evaluating test validity is not a static, one-time event; it is a continuous process.*

Table 1. Facets of Validity (adapted from Messick, 1989b)

	Test Interpretation	Test Use
Evidential Basis	Construct validity	Construct validity + Relevance to the specific applied purpose / utility in the applied setting
Consequential Basis	Value implications of the construct label, of the theory underlying test interpretation, and of the ideologies in which the theory is embedded	Social consequences (both potential and actual) of the applied testing

The *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1999) names five categories of evidence for researchers to consider in building a validity argument. The categories are for evidence based on:

a) test content – how well the content of a test samples the domain that it purports to measure;

b) response processes – the integrity of the data, i.e. the extent to which all sources of error associated with the test administration are controlled or eliminated;

c) internal structure – the statistical or psychometric characteristics of the examination questions, the scale properties, and the psychometric model used;

d) relations to other variables – confirmatory and counter-confirmatory statistical evidence (typically correlations) regarding the relationship between test scores and criterion measure scores; and

e) consequences of testing – the impact on examinees of the test scores, decisions and outcomes, and the wider impact of testing, for example on teaching and learning (Downing, 2003; Moss, 2007; Sireci, 2007).

Downing (2003) provides examples of each category of evidence, his summary of which is reproduced in Table 2. Under content, he considers such things as the coverage by the test specification of the intended subject domain, how comprehensively a test paper covers the contents of the test specification, as well as factors affecting question quality, such as the credentials of the question-writers and the removal of any linguistic, cultural, or other features not relevant to the construct that the question is intended to measure.

Downing interprets response processes as concerning technical and procedural matters, including the correct keying of objective test items, the accurate recording of marks, the rationale for, and accuracy of, aggregating component scores, and the reliability of grade classifications. The *Standards for Educational and Psychological Testing* does not address these matters under any of the five headings, its description of response processes emphasising instead the cognitive and behavioural responses to questions of candidates and also the responses of examiners to candidates' responses. Some of the factors related to

Table 2. Some sources of validity evidence for proposed score interpretations and examples of some types of evidence (adapted from Downing, 2003, p. 832)

Content	Response process	Internal structure	Relationship to other variables	Consequences
Examination blueprint	Student format familiarity	Item analysis data: 1. Item difficulty / discrimination 2. Item / test characteristic curves (ICCs / TCCs) 3. Inter-item correlations 4. Item-total correlations	Correlation with other relevant variables	Impact of test scores / results on students / society
Representativeness of test blueprint to achievement domain	Quality control of electronic scanning/scoring	Score scale reliability	Convergent correlations – internal / external: similar tests	Consequences for learners / future learning
Test specifications	Key validation of preliminary scores	Standard errors of measurement (SEM)	Divergent correlations – internal / external: dissimilar measures	Positive consequences outweigh unintended negative consequences?
Match of item content to test specifications	Accuracy in combining different formats' scores	Generalisability	Test-criterion correlations	Reasonableness of method of establishing pass-fail (cut) score
Representativeness of items to domain	Quality control / accuracy of final scores / marks / grades	Dimensionality	Generalisability of evidence	Pass-fail consequences: 1. P/F decision reliability – Classification accuracy 2. Conditional standard error of measurement at pass score (CSEM)
Logical / empirical relationship of content tested to achievement domain	Subscore / subscale analyses	Item factor analysis		False positives / negatives
Quality of test questions	Accuracy of applying pass-fail decision rules to scores	Differential Item Functioning (DIF)		Instructional / learner consequences, e.g. impact on what is taught and learned
Item writer qualifications	Quality control of score reporting to students / faculty	Psychometric model		
Sensitivity review	Understandable / accurate descriptions / interpretations of scores for students			

question quality that Downing includes under content could perhaps be included under response processes (inline with the *Standards for Educational and Psychological Testing*) and, as he acknowledges, vice versa. Ultimately, it is not terribly important which category they are included in, where there is some ambiguity, rather that they are included somewhere.

Under internal structure, Downing considers issues such as the dimensionality of the measured construct and model fit (factor analysis, item response theory), and various forms of reliability. Also included are question level analyses such as facility and discrimination indices, differential item functioning, as well as correlations between items and correlations between items and total scores.

Under relationships to other variables, Downing considers correlations, including those between the test score and: scores on other tests measuring the same or similar constructs; scores on tests measuring dissimilar constructs; and performance criteria.

Under consequences, Downing considers the implications of test results for candidates and society. He considers it important that positive consequences outweigh any unintended negative consequences (assuming that no one would *intend* to produce negative consequences) and claims that evidence of the quality of the method for setting cut scores should be included in this category. He also considers the acceptability of false positives and negatives, e.g. is it more desirable to fail some reasonably competent medical students than to pass some fairly incompetent ones? Lastly, he mentions the effect of testing on what is taught and learned.

APPLYING THE UNITARY THEORY OF VALIDITY

Cizek, Rosenberg, and Koons (2008) investigated the aspects of validity reflected in a sample of published educational and psychological tests. Their data source was the *Mental Measurements Yearbook* (Spies & Plake, 2005), which contains independent reviews of tests, the intended purposes of which include educational achievement, ability, personality, career guidance, and personnel selection, among other things. They examined the perspective on validity represented, the number and kinds of sources of validity evidence provided, the overall evaluation of the quality of the test, and whether these factors varied as a function of the test type. They found that favourable evaluations of a test tended to be associated with greater provision of validity evidence, but noted a “lingering (mis)perception of validity as adhering to a test” (p. 409) and that validity is represented as consisting of various kinds, when the unitary, inference-based view of validity has been accepted by theorists for almost twenty years. Concerning the latter point, Cizek et al. believe that the lack of an essential difference between the words *sources* and *kinds* is probably the cause of any apparent rift, rather than a genuine reluctance on the part of practitioners to accept the unitary concept. They do, however, identify a potential schism between theorists and practitioners over the matter of consequential validity, finding that very few of the tests reviewed in the *Mental Measurements Yearbook* even referred to the concept. They argue that in “test validation practice, the operationalization of so-called consequential validity and full alignment with the modern perspective on validity is so great a burden that it is simply ignored” (Cizek et al., 2008, p. 410). They note the illogicality of demanding that a test be fully validated before it is used when consequential validity evidence cannot, by definition, be obtained until the test has been employed. Whilst they do not question the importance of considering the consequences of using a test, they do suggest that this can be done outside of the concept of validity and that validity theory should be reconfigured in light

of this. This is a position shared by others (e.g. Maguire, Hattie, & Haig, 1994; Mehrens, 1997; Popham, 1997; Tenopyr, 1996), though not, as Popham notes, by a number of eminent measurement experts (e.g. Linn, 1993, 1997; Messick, 1989b, 1995; Moss, 1995; Shepard, 1993, 1997). However, it appears that Messick's position on consequential validity, and that advocated in the *Standards for Educational and Psychological Testing*, is more nuanced than this debate may suggest:

> *…it is not that adverse social consequences of test use render the use invalid but, rather, that adverse social consequences should not be attributable to any source of test invalidity such as construct irrelevant variance. If the adverse social consequences are empirically traceable to sources of test invalidity…then the validity of the test use is jeopardized… If the social consequences cannot be so traced…then the validity of the test use is not overturned.* (Messick, 1989a, p. 11)

In other words, consequences themselves do not invalidate the use of a test use but can shed light on other factors that do. This would suggest, for example, that it is not invalid *per se* to use a selection test on which females on average outperform males. Rather, validity would only be compromised were the sex differences shown to be the result of the test administration and not of real differences between the sexes on the measured construct. If the measured construct is a demonstrably adequate selection tool, in that it predicts performance in the domain for which one is selecting people, does the fact it produces politically unpalatable—but otherwise fair and valid—results make it invalid? Unlike Messick, Cronbach (1988) regards *any* adverse consequences of testing as a potential threat to validity (Moss, 1992). Taking Cronbach's position, if we chose to measure a different construct for selection purposes, on which there were no apparent sex differences, but which predicted the criterion behaviour less reliably, would that be a more valid test or a less valid test? It does seem that including an ethical or political judgment *within* the concept of validity can muddy the waters. This is not to say that the ethical and political questions are not important, and no one involved in the debate over whether consequences should be part of validity appears to be suggesting that, but rather that they are questions that may be best considered outside of validity matters. This is in fact the position taken in the *Standards for Educational and Psychological Testing*. On the other hand, as Moss (1992) argues, if consequences are considered outside of validity, there is the risk that they are considered less important than if they are considered as part of it. Curren (2004) provides an illustration of some disturbing unintended consequences that testing regimes can have:

> *Anecdotal evidence suggests that under testing regimes in which low scores have repercussions for teachers or their schools, but not for individual students, it has become common, in at least some parts of the USA, for teachers to exert influence over which students are present to take the tests. Another form of test avoidance that may be occurring is the illicit reclassification of low-scoring students as learning disabled. Under some testing regimes, the low test scores of disabled students are not damaging to their schools and districts to the extent that those of other students would be.* (pp. 235-236, 251)

Lissitz and Samuelsen, like Cizek et al., claim that there is "much…dissatisfaction with Messick's unitary concept of validity" owing to "the notion that his rather global view of the topic is impractical" (2007, p. 437). Embretson observes that:

> *Construct validity has always been the most problematic type of validity because it involves theory and the relationship of data to theory. Yet the most*

> *controversial type of validity became the sole type of validity in the revised Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999).* (2007, p. 449)

It has been argued that those responsible for validation almost always require "detailed and concrete guidance", so the unitary notion is not helpful (Brennan, 1998, p. 7) and that the concept of construct validation has been elevated to such a level that it is frequently perceived as unattainable (Fremer, 2000). The concept of validity with which theorists are concerned has, it is argued, become "strangely divorced" from that with which working researchers are concerned (Borsboom, Mellenbergh, & van Heerden, 2004, p. 1061).

Lissitz and Samuelsen (2007), in an issue of *Educational Researcher* dedicated to validity, propose an alternative concept of validity in which internal characteristics of a test such as content validity and reliability are considered as the *internal validity* of the test, whilst all other characteristics are considered as external matters. Internal validity, therefore, becomes the primary justification for the existence and acceptance for use of a test and is independent of the application of the test or its use in theoretical formulations.

Embretson (2007) claims that "Lissitz and Samulesen's framework separates the sources of test meaning for educational and psychological tests" (p. 450). Within their framework, educational tests, classified in the practical perspective, are supported by reliability and content validity evidence, whilst psychological tests are supported by evidence of latent processes (i.e. knowledge, skills, and abilities) and item interrelationships. This approach to supporting the validity of educational tests sounds rather like that currently taken for English general qualifications, though that is not to say it is appropriate. Embretson believes content validity and reliability evidence are inadequate for validating educational tests for several reasons. First, the structure of a content domain and views on complexity level may change over time. Second, there is little evidence to show that items can be reliably classified according to the categories implied by the content domain. Third, the practical imperatives of large-scale testing can lead to unrepresentative sampling of the content domain. For example, more objective item formats may be favoured at the expense of eliciting deeper levels of reasoning. Fourth, these types of evidence are not adequate to circumvent irrelevant sources of item performance (construct irrelevant variance), such as excessive wordiness. Embretson concludes that "the role of external evidence in establishing test meaning should be minimized and that internal evidence should be strongly emphasized. However, given the need for multiple sources of evidence to establish internal test meaning, including theoretical components even for educational tests…the validity system…is…best labelled as *construct validity*" (p. 454).

Contrary to Lissitz and Samuelsen, Gorin (2007) argues that "constructs exist across all assessment contexts, whereas content does not" (p. 457). She gives, as an example, measures of language impairment administered by speech pathologists and school psychologists:

> *The typical tasks, including sentence repetition, fast mapping, nonword repetition, and rapid naming, are content free; adequate score meaning derives from the representativeness of the skills required of the items, independent of content… A parsimonious theory of validity suitable for all possible contexts rather than limited to discipline-specific models would be preferable.* (p. 457)

Although Gorin's argument is sound, the desirability of a "parsimonious theory of validity" from the point of view of theorists does not in itself constitute an argument against the developers of

different tests subscribing to differing theories of validity. Nor does the fact that some tests are content free preclude an emphasis on content validity evidence for tests that are not content free. For instance, the use of a vocabulary test in a modern foreign language course could probably be validated quite satisfactorily with content-related evidence and evidence of marking reliability.

Gorin further claims that Lissitz and Samuelsen's proposed use of content validity tools as indicators of score meaning has been tried and discarded previously, whilst construct validity theories "advance methods…that increase the level of detail and the empirical nature of score descriptions…" (p. 457). For example, cognitive models, unlike behavioural descriptions, specify individual processes that describe item solutions, thus providing testable hypotheses regarding score meanings. Gorin states that cognitive models of test items have been used to "improve the quality of validity arguments…streamline item development procedures…and to augment typical score reports generated from tests…" (p. 458).

Messick (1989b) argues that inferences about behaviours require evidence of response or performance consistency, not just judgements of content, whilst inferences about processes require construct-related evidence; content- and construct-related inferences are thus inseparable. Within Lissitz and Samuelsen's framework, if educational tests are supported from a practical perspective with evidence of content validity and reliability, i.e. response or performance consistency, then Messick's point is not damning, provided that no inferences about processes are to be made. Perhaps more problematic, however, is his point that the concept of content validity ignores the instrument error that is introduced when one set of behaviours occurs in a test situation and the other outside the test situation. He argues that such context effects or irrelevant method variance are reasons why content-related evidence must be considered as part of a broader set of construct-related evidence.

Citing examples, Moss (2007, p. 474) describes the kinds of guidance for test developers and evaluators that have been provided under the unitary conception of validity in educational measurement, including:

> *(a) lists of categories of types of evidence, inferences, or aspects of validity, often illustrated with examples of the kinds of studies that might be undertaken (e.g. American Educational Research Association et al., 1999; Cronbach, 1988; Kane, 1992, 2006; Messick, 1989b);*
>
> *(b) principles to guide choices among the myriad kinds of evidence that are arguably relevant to a given interpretation or use (e.g. Cronbach, 1989; Kane, 2006; Shepard, 1993);*
>
> *(c) standards or guidelines about the nature of evidence that should be made available to enable professional judgment (e.g. American Educational Research Association et al., 1999; Educational Testing Service, 2002);*
>
> *(d) outlines of "interpretive arguments" (Kane, 2006) or comprehensive plans for validity research for particular types of interpretations and uses of tests (e.g. an algebra placement test), accompanied by examples of the types of evidence that might be or have been developed under the plan (Kane, 1992, 2006; Shepard, 1993);*

> *(e) descriptions and (critical) analyses of actual programs of validity research, associated with a particular test or construct (e.g. Cronbach, 1989, p. 150; Shepard, 1993, pp. 432-443); and*
>
> *(f) frameworks illustrated with extended examples (Kane, 2004, 2006; Mislevy, Steinberg, & Almond, 2003; Wilson, 2005) that take us from conceptualization through test development and implementation…*

As a proponent of validation by design, Mislevy (2007) argues that "…test creators are not carrying out validation activities but carrying out design activities structured in such a way that validity evidence emerges" (p. 467). He states that:

> *A performance arises from the interaction between a person and a situation, and any conception of capability ultimately concerns potential interactions between persons and situations of various kinds. Characterizing assessment situations and use situations through the same lens permits a test developer to distinguish essential features of assessment task and targeted use situations. Building tasks around them at once offers practical guidance and constitutes construct representation validity evidence. Furthermore, the differences between test situations and use situations, and the capabilities entailed by one but not the other, raise theoretically motivated, alternative explanations to be explored empirically in use situations. Such theory-grounded backing for task design, and hence for construct-representation validity arguments, can draw upon cognitive studies variously from the information-processing, expertise research, situative psychology, and sociocultural literatures.* (p. 466)

It would be unfair to suggest that the GCSE and GCE awarding bodies view candidates as mindlessly producing responses to questions in the way that Pavlov's dog produced saliva in response to a bell, however their distinctly atheoretical approach to validity has a hint of behaviourism about it. Norris, Leighton, and Phillips (2004) provide an excellent illustration (see Appendix 1) as to why we should be concerned about the mental processes underlying candidates' responses and not simply the response itself (see also Hamilton, Nussbaum, & Snow, 1997). Norris et al. show how, in a multiple choice test, candidates may arrive at the correct answer through entirely faulty reasoning. If we are to make inferences from test scores about candidates' reasoning abilities, this occurrence is highly unsatisfactory. This is probably more of an issue for multiple choice tests, and tests requiring short, open responses, than it is for tests using longer, constructed responses.

CONTEMPORARY VALIDITY THEORY IN THE ENGLISH CONTEXT

Perhaps the differences between the United States and the English approach to validity are historical, stemming from two different testing traditions. English public examinations in schools have their roots in the first University of London matriculation examination, which was held in 1838, albeit with only twenty-three candidates (twenty-two of whom passed). Candidates were required to pass four papers in classics, mathematics, natural philosophy and either chemistry, botany, or zoology. The question papers comprised multiple open response questions, an example of a chemistry question being: "State the composition of the atmosphere and of water, and enumerate the different compounds which result from the union of their respective elements" (Harte, 1986). Although modern GCSE and GCE papers would probably structure

this compound question as two or three part-questions, as well as indicating the marks available for answers to those parts, the question paper format is instantly recognizable as the close ancestor of the GCSE or GCE question paper. Open response questions and extended writing (either in the form of essays or coursework) remain the staple of GCSE and GCE assessments, although short answer and multiple choice questions are used increasingly.

During the twentieth century, the United States developed a quite different approach to the assessment of students, characterised by a strong theoretical base of behavioural measurement, or psychometrics, and relying firmly on objective and standardised modes of assessment, in particular multiple choice testing (Greaney, Bethell, Kellaghan, & McManus, 2001). These different traditions will have led to different mixes of professionals administering the examinations. In England, examination papers are written by practising subject teachers or lecturers and one is unlikely to hear discussions about constructs or item characteristic curves. In the United States, one would of course find subject experts writing questions (items), but the construction of tests from those items would be the responsibility of psychometricians.

The greater (or at least vastly more explicit) emphasis on validity in the United States may stem from the nature of testing there and the professionals who have traditionally constructed the tests. However, tradition cannot justify complacency nor continuing poor practice, so should English awarding bodies be paying greater attention to validity evidence in the design and evaluation of their assessments? To a small extent, it could be a matter of making validation more explicit than it currently is. Question paper evaluation committees (QPECs) are convened to ensure content validity and to remove potential sources of construct-irrelevant variance and, as noted above, awarding bodies make accommodations for candidates with specific needs, so there is an awareness that features of an assessment can challenge candidates in ways that are not related to the purpose of the test. The application of the Disability Discrimination Act 1995 (DDA) to awarding bodies from September 2007 has further raised awareness of the importance of testing only relevant knowledge, skills, and abilities and not inadvertently testing those that are not required by the test specification.

In the case of English general qualifications, the issue of who takes responsibility for validity is an interesting one. The QCA is responsible for setting the national curriculum, which “defines the knowledge, understanding and skills to which children and young people are entitled” (Qualifications and Curriculum Authority, 2008b), and accrediting general qualifications. The QCA’s input to these qualifications is greater than simple accreditation in that they provide subject content and assessment criteria for each subject area. They stipulate broadly what is to be tested and how it will be tested, in terms of the approximate proportion of written papers, coursework assignments, and practical assessments constituting the examination. The awarding bodies are responsible for the detailed content of the specifications and also the setting of individual question papers or tasks, although accreditation takes into account specimens of the assessment materials.

The issue of who is accountable for the validity of general qualifications is broader than this. The modern concept of validity holds validity to be a property of the purposes to which assessment data are put, not of the assessment tests themselves. We could construct a fabulous test of algebra but it would not necessarily be the ideal instrument for deciding who is issued a driving license. What precisely are general qualifications such as GCSEs and GCEs valid *for*? The test specifications describe the knowledge, skills and abilities that candidates obtaining certain grades have demonstrated in the course of the assessment. The least we might want to claim about a Geography exam is that it gives a strong indication of how good an individual is at Geography; what Kane (2002) would call a *descriptive interpretation* (as opposed

to a *decision-based interpretation*). However, nobody becomes a geographer by passing a GCSE or GCE. Performance in a specific subject is likely to affect one's chances of studying that subject further, but for most of the examinations any one of us will take, it is the grade we obtain that is important, rather than the subject it is in. Examination grades in subjects are often described collectively so, with the exception of English and Mathematics, the specific GCSEs one has taken are unlikely to be important in the eyes of an employer and, in many cases, a college admissions tutor. How valid is it to use GCSEs and GCEs as a basis for selecting an applicant for a job? How valid is it to use them for measuring the performance of schools and individual classroom teachers? How valid is it to use a given GCE for selecting university applicants across a range of undergraduate courses (see, for example, Daly, 2007)? More importantly, whose responsibility is it to provide the validity evidence? In the United States context, the *Standards for Educational and Psychological Testing* states:

> *The test developer is responsible for furnishing relevant evidence and a rationale in support of the intended test use. The test user is ultimately responsible for evaluating the evidence in the particular setting in which the test is to be used.* (American Educational Research Association et al., 1999, p. 11)

Presumably, "intended test use" refers to that intended by the test developer, rather than that intended by any particular test user. In the English context, where the regulator has a substantial input to the test specification, it seems reasonable that the regulator should provide validity evidence pertaining to those aspects of the test that it has mandated. The purposes for which the tests should be validated are less obvious. Would it be disingenuous to claim that GCEs are only intended to certificate performance in specific subjects, when we know that their highest stakes use is, effectively if not explicitly, as a predictor of performance at undergraduate level?

Consequences

The proliferation of high stakes testing in the English education system has popularised terms such as 'teaching to the test' and 'curriculum backwash'. These terms essentially refer to the effect that attaching important consequences to pupils' assessment outcomes often has on the way teachers prepare future cohorts of candidates. In addition to Key Stage tests, which are designed to monitor over time the level of attainment of different cohorts at particular stages of education and the value added to individuals, the results of students' GCSEs and GCEs are used to hold schools to account. Consequently, classroom teachers are under pressure to produce excellent results and an obvious way to attempt this is to dedicate classroom time to activities that will better prepare pupils for the tests than other activities. This effect can be negative or positive, depending on the validity of the assessments.

Messick (1989b) describes two principal threats to validity. The first is construct underrepresentation, which occurs when the test is too narrow and fails to include important dimensions or facets of the construct. The second is construct irrelevant test variance, which occurs when the test contains excess reliable variance that is irrelevant to the interpreted construct. An example of construct-irrelevant *difficulty* is the intrusion of undue reading comprehension requirements in a test of subject matter knowledge (e.g. Pollitt & Ahmed, 2001). An example of construct-irrelevant *easiness* is the inclusion of extraneous clues that allow some individuals to respond correctly in ways irrelevant to the construct being assessed, e.g. multiple choice items in which the longest options are most likely to be correct. Clearly, if a high stakes test thoroughly represents the curriculum it is intended to measure, then it will encourage

thorough teaching of the curriculum – a good thing. However, if the test under-represents the curriculum, it is likely that those parts omitted from the test will be omitted from the classroom – a bad thing. Similarly, if the test inadvertently and consistently measures constructs that are irrelevant to the curriculum, i.e. test-taking skills, it will encourage teaching of these constructs rather than those on the curriculum (Harlen, 2007; Harlen & Deakin Crick, 2003).

Wiliam (1992; 1996) uses the context of teaching to the test as a framework for assessing validity, suggesting that "a test is valid to the extent that one would be happy for teachers to teach towards the test" (1992, p. 17). He argues that if we were happy for teachers to teach to the test then we must be satisfied that:

- *the test adequately represented the whole of the domain (within-domain inferences);*
- *teaching towards the test would also increase performance in the correlates of that test, so that increased performance on the test would also increase performance on whatever the test was used to predict (beyond-domain inferences);*
- *the test adequately represented our values about what was important in the domain (within-domain consequences); and*
- *the effects of teaching towards the test on teachers and students were beneficial (beyond-domain consequences).* (1996, p. 134)

Multiple-choice questions

Multiple-choice questions have not traditionally been a large part of the GCSE assessment arsenal but, in recent times, they have become more widely used, notably in some GCSE Science specifications. Multiple-choice questions are regularly pointed at by education commentators as evidence of 'dumbing down'. Curren (2004) notes that "multiple-choice items can and have been developed to test understanding, problem solving, and complex forms of inferential, predictive and explanatory reasoning (Aiken, 1982; Balch, 1964; Haladyna, 1997; Yeh, 2001)" (p. 247). However, he acknowledges that certain aspects of complex performances, e.g. speaking a foreign language or designing and running an experiment, cannot be tested by multiple-choice questions. In addition, the ability to generate novel applications, interpretations, and solutions—what he calls divergent production—are difficult to test using multiple-choice questions (see also Martinez, 1999). Wiliam (1996) claims that generally there is little to choose between multiple-choice questions and constructed-response questions in terms of the inferences one can make from responses to them. He believes, however, that the use of multiple-choice questions can have undesirable consequences that are not associated with constructed-response questions:

> *Even where multiple-choice items have managed to assess higher-order thinking, they have done so by identifying particular higher-order skills. So, even though the full breadth of the domain is sampled, it is done in an atomistic way. The effect of this has been to cause teachers to concentrate on isolated elements of the domain, because the 'glue' never gets assessed.* (p. 139)

Everything in moderation

Messick (1989b) presents a case for using "not only multiple measures of the construct but also distinctly different methods of measurement" (p. 35). Suppose we wish to measure a complex construct that has three facets, A, B, and C, using three separate measures. Each measure under-represents some aspect of the complex construct: measure 1 taps facets A and B plus some irrelevant variance, X; measure 2 taps B and C plus Y; and measure 3 taps A and C plus Z. Owing to the overlapping components, the three measures will correlate positively, appearing to converge in the measurement of the complex construct. Using a composite of the three measures fully represents the complex construct and allows the construct-relevant variance to cumulate in the composite score while the irrelevant variance does not. Messick makes the points that if none of the measures tapped facet C, the correlations may still be strong and, furthermore, that construct-irrelevant variance may be common to the three measures, and thus cumulate in the composite score. Thus, convergent evidence, showing that measures correlate as implied by the construct theory is insufficient; further "discriminant" evidence that the tests are not related to another construct is necessary. He notes that the method of measurement is a "rich source of such rival constructs" (Messick, 1989b, p. 35) and gives as examples the effects of "recognition on multiple-choice tests of recall or knowledge retrieval" and of "verbal ability on paper-and-pencil tests of mechanical aptitude". This is a clear argument for assessing candidates using as broad a range of methods as other validity concerns permit.

GCSEs and GCEs generally do this, employing multiple-choice questions, short-answer questions, essay questions, coursework, and practical performances, as appropriate. However, as noted above, some of the new GCSE Science specifications use multiple-choice questions extensively. At the same time, coursework, which in the early days of the GCSE (1988 – 1994) constituted as much as one hundred percent of the assessment in certain specifications, has seen something of a reversal in fortune thanks to the use since 1994 of the GCSE to measure Key Stage four (Elwood, 2000) and, more recently, concerns about plagiarism, with QCA providing less scope for its use in recently redeveloped specifications. Given Messick's sound argument, we should attempt to assess candidates using as broad a range of methods as we can.

CONCLUSIONS

In the development of new GCSE and GCE specifications, there is little explicit reference to validity, in the technical sense widely embraced by validity theorists and ensconced in the *Standards for Educational and Psychological Testing*. There are procedures and practices in place that address validity, however, they would be unlikely to meet the contemporary standards of validity evidence that would be expected of high stakes tests in the United States. This disparity reflects the differences in educational testing cultures between the two countries, and the fact that English examinations have tended to be far less reliant on multiple-choice tests partly exonerates the English. Although some researchers have suggested that evidence of content validity and reliability are adequate for educational tests, the weight of opinion suggests that this is not true, as these concepts fail to capture important aspects of validity, such as construct-irrelevant variance. To illustrate the need for a more sophisticated validity framework, one can look at the current concern about possible effects of test presentation mode—paper versus onscreen—on candidates' performance, where the mode of presentation is optional. This is a validity issue because one test mode may be measuring something that is relevant, or irrelevant, to the construct of interest that the other test mode is not measuring.

There is little evidence of a public debate surrounding the validity of general qualifications, except perhaps over plagiarism in coursework and the worth of multiple-choice tests. As the latter are being used increasingly, they may be a good place for the English awarding bodies to begin developing procedures for ensuring validity through the design of the assessment and the writing and evaluation of questions (items). The *Standards for Educational and Psychological Testing* would make a sensible starting point, whilst many of the references cited above would provide more explicit and detailed advice on what can and has been done to support validity arguments for the uses of tests. As a potential starting point, Appendix B contains an outline of a possible framework for assuring—and, where necessary, defending—the validity of English general qualifications. Any detailed protocol will require a thorough review of existing procedures and the validity evidence that they can yield.

Another difference between the United States and English contexts is the level of regulation and the input the regulator has in the design of test specifications. There are important design features of GCSEs and GCEs that are beyond the control of individual awarding bodies. Any serious attempt at ensuring validity will require cooperation between the regulatory authorities and the awarding bodies over the assignment of responsibility for different aspects of the specification. Not least of these concerns is the issue of precisely what we expect general qualifications, which have myriad potential uses, to be valid for.

Neil Stringer 18 November 2008

REFERENCES

Aiken, L. R. (1982). Writing multiple-choice items to measure higher-order educational objectives. *Educational and Psychological Measurement, 42*, 803–806.

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

Assessment and Qualifications Alliance. (2005). *Specification Development Procedure Guidance File*. Retrieved 22 October 2008, from http://intranet.aqa.org.uk/pdf/spec-procedure-gf.pdf.

Assessment and Qualifications Alliance. (2006). *Specification Development Procedure Guidance File: GCE 2006/7 for teaching from September 2008*. Retrieved 22 October 2008, from http://intranet.aqa.org.uk/pdf/gce-spec-dev-pgf.pdf.

Assessment and Qualifications Alliance. (2007). *Question Paper Preparation Procedure Guidance File*. Retrieved 22 October 2008, from http://intranet.aqa.org.uk/pdf/Question-paper-procedure-guidance.pdf.

Balch, J. (1964). The influence of the evaluating instrument on students' learning. *American Educational Research Journal, 1*, 169–182.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-1071.

Brennan, R. L. (1998). Misconceptions at the intersection of measurement theory and practice. *Educational Measurement: Issues and Practice, 17*(1), 5-9.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*(3), 397-412.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, DC: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer (Ed.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: measurement, theory and public policy* (pp. 147–171). Urbana, IL: University of Illinois Press.

Curren, R. R. (2004). Educational measurement and knowledge of other minds. *Theory and Research in Education, 2*(3), 235-253.

Daly, A. (2007). *Law and psychology admissions tutors' perspectives of how A-levels prepare students for university study* (No. RPA_07_AD_RP_057). Guildford: Assessment and Qualifications Alliance.

Downing, S. M. (2003). Validity: on the meaningful interpretation of assessment data. *Medical Education, 37*(9), 830-837.

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

Elwood, J. (2000). Examination techniques: issues of validity and effects on pupils' performance. In D. Scott (Ed.), *Curriculum and Assessment*. Westport, CT: Ablex Publishing.

Embretson, S. E. (2007). Construct validity: a universal validity system or just another test evaluation procedure? *Educational Researcher, 36*(8), 449-455.

Fremer, J. (2000). Promoting high standards and the "problem" with construct validation. *NCME Newsletter, 8*(3), 1.

Gorin, J. S. (2007). Reconsidering issues in validity theory. *Educational Researcher, 36*(8), 456-462.

Greaney, V., Bethell, G., Kellaghan, T., & McManus, H. (2001). Public Examination System. Retrieved 1 October, 2008, from http://www1.worldbank.org/education/exams/

Haladyna, T. M. (1997). *Writing Test Items to Evaluate Higher Order Thinking*. Boston, MA: Allyn & Bacon.

Hamilton, L. S., Nussbaum, E. M., & Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10*(2), 181-200.

Harlen, W. (2007). Criteria for evaluating systems for student assessment. *Studies in Educational Evaluation, 33*, 15-28.

Harlen, W., & Deakin Crick, R. (2003). Testing and motivation for learning. *Assessment in Education, 10*(2), 169-208.

Harte, N. B. (1986). *The University of London, 1836-1986: an illustrated history*. London: The Athlone Press.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535.

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31-41.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement, 2*(3), 135–170.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education / Praeger.

Linn, R. L. (1993). Educational assessment: expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15*, 1-16.

Linn, R. L. (1997). Evaluating the validity of assessments: the consequences of use. *Educational Measurement: Issues and Practice, 16*(2), 14-16.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*(8), 437-448.

Maguire, T., Hattie, J., & Haig, B. (1994). Construct validity and achievement assessment. *Alberta Journal of Educational Research, 40*(2), 109-126.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207–218.

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16*(2), 16-18.

Messick, S. (1989a). Meaning and values in test validation: the science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Messick, S. (1989b). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8.

Mislevy, R. J. (2007). Validity by design. *Educational Researcher, 36*(8), 463-469.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–67.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: implications for performance assessment. *Review of Educational Research, 62*(3), 229-258.

Moss, P. A. (1995). Themes and variations in validity theory. *Educational Measurement: Issues and Practice, 14*(2), 5-13.

Moss, P. A. (2007). Reconstructing validity. *Educational Researcher, 36*(8), 470-476.

Norris, S. P. (1992). A demonstration of the use of verbal reports of thinking in multiple-choice critical thinking test design. *Alberta Journal of Educational Research, 38*, 155-176.

Norris, S. P., Leighton, J. P., & Phillips, L. M. (2004). What is at stake in knowing the content and capabilities of children's minds? A case for basing high stakes tests on cognitive models. *Theory and Research in Education, 2*(3), 283-308.

Pollitt, A., & Ahmed, A. (2001). *Science or reading? How students think when answering TIMSS questions*. Paper presented at the 27th International Association for Educational Assessment (IAEA) Annual Conference. Retrieved 28 October 2008, from http://www.cambridgeassessment.org.uk/ca/digitalAssets/113880_Science_or_Reading._How_Students_Think_When_Answering_TIMSS_.pdf.

Popham, W. J. (1997). Consequential validity: right concern - wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9-13.

Qualifications and Curriculum Authority. (2008a). *GCSE, GCE and AEA Code of Practice*. London: Qualifications and Curriculum Authority.

Qualifications and Curriculum Authority. (2008b). What we do. Retrieved 2 October, 2008, from http://www.qca.org.uk/qca_8710.aspx

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Educational and Psychological Measurement, 19*, 405–450.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5-8,13,24.

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher, 36*(8), 477-481.

Spies, R. A., & Plake, B. S. (Eds.). (2005). *The sixteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.

Tenopyr, M. L. (1996). *Construct-consequences confusion*. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology.

Thyne, J. M. (1974). *Principles of examining*. London: University of London Press.

Wiliam, D. (1992). Some technical issues in assessment: a user's guide. *British Journal for Curriculum and Assessment, 2*(3), 11-20.

Wiliam, D. (1996). National Curriculum assessments and programmes of study: validity and impact. *British Educational Research Journal, 22*(1), 129-141.

Wilson, M. R. (2005). *Constructing measures: an item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.

Yeh, S. S. (2001). Tests worth teaching to: constructing state-mandated tests that emphasize critical thinking. *Educational Researcher, 30*(9), 12–17.

APPENDIX A: EXAMPLE FROM A CRITICAL THINKING TEST (Norris et al., 2004, pp. 294-296)

First, consider an example taken from a study of the design and validation of a critical thinking test of observation appraisal (Norris, 1992). Here is Item 3 from Part A of the test, which is set in the context of a traffic accident:

> A policewoman has been asking Mr Wang and Ms Vernon questions. She asks Mr Wang, who was one of the people involved in the accident, whether he had used his signal.
>
> Mr Wang answers, '*Yes, I did use my signal*'.
>
> Ms Vernon had been driving a car which was not involved in the accident. She tells the officer, '*Mr Wang did not use his signal*. But this didn't cause the accident'.

Examinees were to choose which, if either, of the italicized statements is more credible. To reason through this question correctly, an examinee first needs to derive from the text the relevant information about Wang's and Vernon's involvement. The text is simple enough that most high school students should have no difficulty with this aspect of the task. Second, an examinee must retrieve from background knowledge the relevant facts that not using a turn signal can cause an accident and that being held responsible for an accident can cause trouble with the law. Again, high school students would have such common knowledge. Finally, an examinee has to infer that, because a person knows that admitting to not using a signal could be interpreted as causing an accident, Wang is in a conflict of interest that reduces his credibility with respect to Vernon.

In order to determine whether students who chose the keyed answer, that Ms Vernon was more credible, reasoned well and those who chose another answer did not, we asked students to think aloud on the item. Here are the verbatim transcriptions of the verbal reports of two students:

Student A:

> …ah…ah, Mr Wang, like he probably didn't, like you know, it was just, he probably thought he used his signal, but really didn't. And Miss Vernon, she was watching, so she'd be able to tell from back if he was using it or not. Right? Being the case, so, I'd tend to believe Vernon.

Student B:

> I would say that he did use his signal because anybody who's in the car . . . coming up to an intersection or anything . . . he, he usually knows what he's doing. So I'd be more inclined to believe the first.

Student A chose the keyed response and Student B an unkeyed one. Neither student thought critically on the item, however. Student A's reasoning that Wang just 'thought he used his signal', but Vernon would 'be able to tell from back' because 'she was watching' is arbitrary. There is no information in the item to justify the claims about either Wang or Vernon. It is used to rationalize a choice of answer rather than to justify it. Given the information in the item, it is

just as reasonable to say that Wang would ‘be able to tell from inside that he used his signal’ and that Vernon just ‘thought he did not use his signal’.

Student B reasoned just this way and failed to distinguish between Wang’s knowing what he had done and his reporting accurately what he had done, and failed to allow that someone in another car can know what another driver is doing.

From these verbal reports we can conclude that Student A thought poorly but chose the correct answer nevertheless, and that Student B thought poorly and chose the incorrect answer. The evidence from Student A tells against the quality of the item; the evidence from Student B speaks to the quality of the item. Evidence such as this, accumulated across many students from the population for which the test is designed and across all the items on the test, can support general conclusions about what the test is or is not measuring.

APPENDIX B: A DRAFT FRAMEWORK FOR ASSURING THE VALIDITY OF ENGLISH GENERAL QUALIFICATIONS

This draft framework describes some of the types of analyses that AQA could perform and the procedures that could be followed to assure the validity of tests. They are collected under the five headings referred to in the *Standards for Educational and Psychological Testing*. This document is intended to provide an overview: a thorough review of AQA's processes will be required to see if they yield robust validity evidence. As the main paper has highlighted, consultation with the regulatory authorities would be essential.

CONTENT EVIDENCE

The test specification should relate the content that is tested to the subject domain, as it is described by the course learning objectives. The specification must be sufficiently detailed to describe subcategories of content and to specify precisely the proportion of test questions in each category and the level of those questions. The quality of questions is a source of content-related validity evidence and we should consider whether:

i. questions adhere to the best evidence-based principles of effective item-writing;

ii. question-writers are qualified as content experts in the disciplines;

iii. there are sufficient numbers of questions to adequately sample the large content domain;

iv. test questions have been edited for clarity, removing all ambiguities and other common item flaws;

v. test questions have been reviewed for cultural sensitivity.

RESPONSE PROCESSES

In the *Standards for Educational and Psychological Testing*, evidence based on response processes refers to the fit between the construct and the detailed nature of performance or response actually engaged in by candidates. Evidence would usually come from analyses of individual responses, e.g. questioning candidates about their performance strategies or responses to particular items. Validation may also include empirical studies of how examiners record and evaluate data along with analyses of the appropriateness of these processes to the intended interpretation.

This category of evidence has also been interpreted more broadly to refer to the integrity of the test data generally. At a purely administrative level, this could refer to simple things like the accurate totalling and entering of marks into the examination processing system. It might also include the rationale for the chosen method of aggregating component scores, e.g. the addition method (indicator 1) versus the percentile method (indicator 2). We might also consider here the accuracy and interpretation of grade classifications.

INTERNAL STRUCTURE

This category of evidence deals with analyses that tell us about the statistical and psychometric properties of our test. The Research and Policy Analysis department currently provides specification analyses on request, although papers that are marked onscreen receive them automatically. Specification analyses include question level analyses such as *facility indices*, which indicate the relative difficulty of questions on a paper, and *discrimination indices*, which measure the extent to which a question was able to differentiate between candidates with high and low total marks. For multiple-choice tests, *item distractor analysis* can reveal the answers most commonly chosen by candidates of different levels of ability. This can help identify questions that are not performing well, for example a single distractor that is never chosen, that is chosen more often than all other options, including the answer, or that correlates positively with the total score.

A crucial condition of validity is that the test produces reliable scores and outcomes (e.g. pass/fail). This means that candidates should obtain the same score, or outcome, if they took the same test again, or an alternative form of the test, under the same conditions. They should also obtain the same score independent of which examiner marked his or her paper.

There are analyses that indicate how many constructs or factors a test appears to measure by examining the relationships among candidates' scores on individual questions. A subject expert should be able to interpret these clusters of questions as measuring something common. *Item response theory* assumes unidimensionality, so any test constructed using this method must demonstrate model fit for the intended interpretations of the test score to be valid. If we claim that there are five factors underlying performance in a particular domain, *factor analysis* can be used to determine whether our test measures all five of them adequately and whether individual questions are measuring the factors that they were intended to. These techniques can also highlight sources of construct-irrelevant variance, e.g. a cluster of questions on a mathematics test that load on advanced reading comprehension.

Differential item functioning (DIF) refers to when subgroups of test takers of overall equal ability perform differently on a particular question. DIF can be perfectly legitimate, but it can also highlight the measurement of construct-irrelevant variance, e.g. non-native English speakers scoring lower on a spatial reasoning question than native English speakers of similar overall spatial reasoning ability.

RELATIONS TO OTHER VARIABLES

This category of evidence typically includes correlations between candidates' test scores and their scores on external variables. The latter may include measures of criteria that the test is expected to predict, or scores on other tests hypothesised to measure the same, related, or indeed different, constructs. Strong correlations with measures of related constructs provide what is called convergent evidence, whilst weak correlations with measures of unrelated constructs provide discriminant evidence. Measures other than test scores, such as performance criteria, may be appropriate, for example, in employment settings. Test-criterion correlations often vary considerably when a test is used to predict the same or similar criteria at different times or in different places, so statistical summaries of past validation studies in similar situations may prove useful in estimating test-criterion relationships in a new situation.

Categorical variables, such as group membership, may be relevant when group differences are expected to be present or absent.

CONSEQUENCES OF TESTING

It is difficult to be prescriptive for such a broad category of evidence. However, an example of something we should consider is a backwash effect that is often noted by Principal Moderators. That is that, where the coursework assignments are marked according to a fixed mark scheme, over time there is a tendency for coursework to become targeted at the mark scheme. The result is high scoring but unoriginal work, prescribed by teachers, which does not challenge candidates in the way that was originally intended.