Centre for Education Research and Policy

ASSESSMENT EXPERTISE PROJECT: VALIDITY OF ASSESSMENT

Neil Stringer

INTRODUCTION

When candidates take a test, whether it is a written examination, a practical performance, or submitting a portfolio of work, it is taken so that an examiner can assess the candidate's performance relative to a standard. There is normally a reason for doing this: to select people for further education or training; to assess suitability for a specific job; to certify what individuals can do in a particular domain; and so on. Validity is essentially the concern that the inferences drawn from a candidate's performance and any decisions based on these inferences are reasonable and fair. In some cases, demonstrating validity is relatively straightforward. For example, if you wanted to know the relative spelling abilities of a class of children, you could orally present commonly used English words and ask them to write them down using the correct spelling. Their answers could be marked as correct (scoring 1) or incorrect (scoring 0) and their scores totalled. Providing that the sample of words is sufficiently large and unbiased (e.g. they do not favour any dialects, or musical terms are not overrepresented, etc.), it should be uncontroversial to claim that the higher the child's score, the better they are at spelling. Now, suppose you claim that the higher a child's score is on this test, the better he or she is at English. The authenticity of this claim is not self-evident and, if you plan to make it, you should be prepared to demonstrate that it is fair and reasonable—which, in this case, it almost certainly is not.

In most cases where one assesses the validity of a test use, the verdict will not be as apparent as in the above example of a spelling test. There are two main threats to the validity of any test use (Messick, 1989). The first is that the test does not measure everything that it should (construct underrepresentation). In the case of the spelling test, there are clearly knowledge, skills, and abilities that constitute English, which are not encompassed by a spelling test, such as composition, reading comprehension, and speaking. Unless it is known that all individuals are equally good at all of the constituent parts, they must all be measured, not inferred from the measurement of one part. The second threat to validity is that the test measures things that it should not (construct irrelevant variance). In the case of the spelling test, were the test administrator to present the words in French so that the candidates had to translate them into English before spelling them, the test would very clearly be measuring something in addition, and irrelevant, to English spelling ability. Again, in practice, these threats are likely to be more subtle than these examples, perhaps especially in the case of construct irrelevant variance, where statistical analyses of the item scores may draw attention to a problem in the first instance.

WHO IS RESPONSIBLE FOR VALIDITY?

In the United States, the validity of high stakes assessments has, historically, received vastly more explicit attention than it has done in the United Kingdom. In the US, the *Standards for educational and psychological testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) (henceforth referred to as 'the *Standards*') serves a similar function as the Office of the Qualifications and Examinations Regulator (Ofqual) Code of Practice in England, although, unlike that document,

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

the *Standards* is not legally binding. The *Standards* is very clear that test developers should be specific about what their tests are, and are not, intended to be used for, and that they are responsible for collecting and documenting evidence to support the uses they claim. The responsibility for any specific use of a test, however, rests with the test user.

The position of English awarding bodies differs insofar as the curriculum is determined by the Qualifications and Curriculum Development Agency (QCDA)[1] whilst the subject criteria on which the exams are based are determined by the regulator, Ofqual. Ofqual is also heavily involved in the development of test specifications, for example, it may stipulate the methods of assessment used and their approximate weightings in the final test score. There is also an issue regarding what general qualifications, such as GCSE and GCE, are for. At the simplest level, they are to assess candidates' achievement at the end of a course of learning that is based on the National Curriculum. If we limit the intended use solely to this, then QCDA is responsible for making sure that the curriculum for History, for example, is appropriate for the subject. Ofqual is responsible for ensuring that the subject criteria reflect the curriculum appropriately. Finally, the awarding body—and Ofqual, to the extent that it prescribes certain details of the test specification—is then charged with assuring that the examination is a valid measure of attainment on the curriculum against the subject criteria. There are purposes to which examination grades are routinely put for which the awarding body could not reasonably hope, or be expected, to validate examinations. These include the use of examination grades for selecting candidates for any given education course, or employment, and the use of examination data for measuring the value added to individuals by schools and colleges at various stages of schooling.

There are few explicit references to validity in the current Ofqual Code of Practice and the relevant AQA procedure guidance files. In compliance with the Code of Practice, AQA has procedures and practices in place that address matters of validity – examples of which are included in the appendix to this chapter – but no concerted process for routinely providing the range of validity evidence required by the *Standards* and which is described below. Such a process is desirable and, with appropriate procedures in place, the collection, documentation, and auditing of validity evidence need not be unduly burdensome. The output of this process would inform the development, or redevelopment, of the test specification, depending on when, during the lifespan of the test specification, the auditing is conducted.

WHEN DO WE VALIDATE?

The nature of general qualifications is that a test specification is written, which includes specimen examination papers, with multiple versions of the test administered over the lifespan of the test specification, which is currently approximately eight years. Ideally, validation should occur during development and be completed prior to live testing, because it would be unfair to administer potentially invalid tests and because there is very little that can be changed during the lifespan of a test specification if one were to discover something wrong with it. Some forms of evidence, such as correlations between examination scores and external variables (which 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), are not easily collected in advance of live testing without extensive piloting of the qualifications. Ordinarily, qualifications are only piloted if they are a new type of qualification, such as the Extended Project (Pinot De Moira, 2007a, 2007b), or at least a significantly new form of an existing one, such as GCE Business Studies prior to the introduction of AS level (Cresswell, 2000). There

[1] Formerly the Qualifications and Curriculum Authority (QCA).

was no piloting when, for example, all the GCE specifications were revised for teaching from 2008.

Depending on the degree of similarity between a new test specification and its predecessor, it may be reasonable to use validity evidence collected during the lifespan of an outgoing test specification to inform the design of the new one. Although this would not be as satisfactory as collecting data through pilot studies, it has some practical benefits in its favour. First, whole suites of GCSEs or GCEs are, with a small number of exceptions, redeveloped concurrently, so to collect and analyse pilot data for every GCSE or GCE specification in parallel would place enormous demands on awarding body personnel, not to mention the centres and candidates who would be asked to participate in them. If the data were collected during the lifespan of the test specification, the workload could be staggered over many years. Second, the size of the datasets that could be collected would be far larger using operational data than they would be using pilot studies. Third, (ecological) validity can become a problem for the validation study itself if the data is collected under no-stakes conditions. This is not an issue when operational data are collected.

TYPES OF VALIDITY EVIDENCE

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 test 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.

Readers may find it useful to refer to the section of this volume entitled *Essentials of good question writing*.

Frequency: GCSE and GCE specifications are replaced periodically, so the content should be stable for the duration of the test specification. The credentials of item-writers will be established from the time they are engaged in the role and they should receive any necessary training prior to beginning work. Adequate sampling of the content domain is a matter for whoever compiles the question paper but this need not be an item-writer. The quality of specific questions, in terms of clarity and cultural sensitivity, is an issue for every new question or question paper.

Response processes

In the *Standards*, 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 and might include the rationale for the chosen method of aggregating component scores. We might also consider here the accuracy and interpretation of grade classifications.

Frequency: Candidates' response processes could be sampled for each paper, whilst the Principal Examiners' mark scheme and the standardisation process may serve as documentary evidence of examiners' processes. The interpretation of grade classifications should be dealt with in the first award of any test specification whilst technical aspects of grade validity, such as grade boundary widths in relation to test reliability, could be monitored over series.

Internal structure

This category of evidence deals with analyses that tell us about the statistical and psychometric properties of our test. These 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. Facility indices can inform us whether the test contains a sufficient spread of question difficulties, such that each grade or ability level is adequately targeted. Questions that do not discriminate between weak and strong candidates (as defined by their total test scores) do not provide information about candidates in relation to the knowledge, skill, or ability the test is intended to measure. They possibly measure something they were not intended to or may simply be far too easy for the candidature (in which case the facility index will be very high); either way, they are not very useful. For multiple-choice tests, *item distractor analysis* can reveal the answers most commonly chosen by candidates of different levels of ability. This can help to identify questions that are not performing well, such as 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 consistently accurate scores, or outcomes, if they took the same test again, or an alternative form of the test, under the same conditions.

Broadly speaking, there are two major steps to ensuring the reliability of non-standardised tests, such as GCSEs and GCEs. The first is to ensure that candidates would be placed in the same rank order, whichever version of the test they took. Addressing the aspects of validity discussed so far, including threats such as construct underrepresentation and construct irrelevant variance, as well as ensuring that each grade or ability level is adequately targeted, will help assure this. Furthermore, candidates should obtain the same score independent of which examiner marked his or her paper. Readers can find more on the latter in the section of this volume entitled *Marking reliability and the examination cycle*. The second major step is to ensure that whatever score a candidate receives on any version of the test, it is converted into the same grade each time. Readers can find more on this in the section of this volume entitled *Principles of standard setting*.

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. 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.

Frequency: With the exception of (external) reliability measures, these statistics can be produced routinely, especially in subjects where question level data are captured electronically. Marking reliability, too, is routinely monitored through standardisation procedures. Empirical studies of reliability involving candidates taking the same test paper, or two different papers, e.g. the live paper and a past paper, on two occasions could be conducted; however, they are likely to lack validity given the high stakes nature of the live paper and the low stakes nature of the past paper. This form of reliability is probably best addressed through statistical proxies taken from single test administrations.

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.

Frequency: This will depend largely on the availability of data. If measures of criteria were readily available, this evidence would ideally be collected for each exam paper. Otherwise, this should at least be done during the development of the test specification and intermittently during its lifetime.

Consequences of testing

It is difficult to be prescriptive for such a broad category of evidence. However, an example of something we might consider is a backwash effect that is often noted by coursework 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. (Time will tell whether controlled assessment in GCSE will be susceptible to the same effect.) Readers can find more on this in the section of this volume entitled *Principles of writing a scheme of assessment*.

Frequency: Again, it is difficult to be prescriptive but, for example, by their nature, backwash effects are likely to be observed over time.

CONCLUSION

Validity is simply about ensuring that test scores are fit for purpose, so that the inferences we make about candidates' scores, and any decisions we base on those inferences, are fair. Essentially this means reliably measuring enough of the thing we are interested in, without inadvertently measuring something else. There is a broad range of evidence, from qualitative judgements through to statistical analyses, which we can draw on to ensure – and demonstrate to users of our examinations – that our examinations are fit for purpose.

SUGGESTED FURTHER READING

For a more detailed introduction to validity and its place within the English examinations system, or simply an entry point to the validity literature, readers may wish to refer to a recent AQA research paper:

Stringer, N. (2008). *Contemporary validity theory and the assessment context in England*. Guildford: AQA Centre for Education Research and Policy

The section on validity in the *Standards* (referenced below) is a must-read for anyone embarking on a validation programme.

Over time, the concept of validity has become all encompassing to the extent it is difficult to point to at an assessment issue that cannot be filed under 'validity'. Accordingly, many of the other chapters in this manual relate to aspects of assessment validity.

Neil Stringer
December 2009

REFERENCES

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.

Cresswell, M. (2000). *AQA pilot AS/A Business Studies (5500p) in summer 2000: A report on the award* (No. RC99). Guildford: Assessment and Qualifications Alliance.

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

Pinot De Moira, A. (2007a). *Extended Project evaluation (cohort 1): Coordinator interviews.* Guildford: Assessment and Qualifications Alliance.

Pinot De Moira, A. (2007b). *Extended Project evaluation (cohort 1): Student information.* Guildford: Assessment and Qualifications Alliance.

Appendix: Some Examples of Existing Validity Evidence

Any validation study or programme will almost certainly require some new forms of evidence to be collected or produced; however, there are data and reports that we currently produce that might be used as evidence of validity, or of the processes and procedures in place to ensure validity. The following possible sources of evidence are examples and the list is not intended to be comprehensive.

Test Specification Development

At the time of test specification development, work is routinely carried out that addresses validity issues. The full test specification document, accompanied by specimen external assessment units and marking schemes, is submitted to the AQA Education and Training Committee (ETC) for approval before being submitted to QCDA. Similarly, members of the Subject Advisory Committee (SAC)[2] are invited formally to comment; they may also be involved at previous stages such as when the summary test specification is being circulated for comment and when decisions regarding structure and content are made prior to detailed drafting.

Question Paper Setting

Reviser's Comment Form
Scrutineer's Report
Question Paper Functioning Report (from previous series)

Sources of Guidance

The Question Paper Preparation Procedure Guidance File contains a number of sections that offer concrete guidance on ensuring validity, most notably:

Appendix 1

1.1 Preparing Question Papers and Mark Schemes. Guidance for Chief/Principal Examiners, Revisers and Scrutineers
1.2 The Use of Resources in GCSE, GCE and GNVQ examinations: General Principles – Question Setting
1.3 *The Language of Examinations* published by the British Association of Teachers of the Deaf
1.4 *Fair Access by Design* published by QCA/ACCAC/CEA
1.5 Guidelines on the Readability and Legibility of Examination Papers

Appendix 2

The Evaluation of Question Papers and Mark Schemes

[2] Note that SACs are being replaced by Subject Communities.