CAMBRIDGE ASSESSMENT

How hard can it be? Issues and challenges in the development of a validation method for traditional written examinations

Victoria Crisp and Stuart Shaw, Cambridge Assessment

Paper presented at the International Association for Educational Assessment Annual Conference, Bangkok, August 2010.

Abstract

There is a wealth of theoretical work on validity. However, translating this into an operational method for validating assessments has not attracted nearly as much attention, largely because validation activities are painstaking and difficult. Evidence needed for validation depends on the proposed interpretations and uses of test scores. However, providing appropriate validity evidence is a non-trivial undertaking and involves substantial research effort, requiring multiple sources of evidence collected through a range of methods to address different facets considered important to validity.

This paper will provide an overview of the issues and challenges in the development, piloting and revision of a framework for validating traditional written examinations. Recent attempts to apply the framework have uncovered a number of difficult issues. For example: what conceptualisation of validity should be used; for whom is the framework intended; how should evidence for validity be presented and used; and can one framework and set of methods be applied satisfactorily to different types of qualifications and assessments.

The paper will discuss the extent to which the full methodology is practical operationally, whether a more streamlined approach may be necessary, and how much evidence is sufficient to consider an assessment valid.

Note:
This conference paper is a summary of research and the issues involved and it is not possible to include full details for reasons of the time and space available. However, please feel free to contact us if you would like further information.

Contact details:
Cambridge Assessment, 1 Hills Road, Cambridge, CB1 2EU, England.
Victoria Crisp. Tel. +44 (0)1223 553805. Email: crisp.v@cambridgeassessment.org.uk
Stuart Shaw. Tel. +44 (0)1223 556089. Email: shaw.s@cie.org.uk

How hard can it be? Issues and challenges in the development of a validation method for traditional written examinations

Introduction

This paper reports briefly on a current strand of research which aims to develop a methodology for validating general academic qualifications such as A levels. Validity is a key principle of assessment, a central aspect of which relates to whether the interpretations and uses of test scores are appropriate and meaningful (Kane, 2006). For this to be the case, various criteria must be achieved, such as good representation of intended constructs, and avoidance of construct-irrelevant variance. Additionally, some conceptualisations of validity include consideration of the consequences that may result from the assessment, such as affects on classroom practice. The kinds of evidence needed may vary depending on the intended uses of assessment outcomes. For example, if assessment results are designed to be used to inform decisions about future study or employment, it is important to ascertain that the qualification acts as suitable preparation for this study or employment, and to some extent predicts likely success.

Validity has long been considered a crucial criterion for an assessment and there now exists a wealth of theoretical work attesting to its importance. However, practical examples of how to validate an assessment are less common largely because "validation work is unglamorous and needs to be painstaking" (Wood, 1991, p.151-2). To *validate* an assessment, evidence to support the claims made about the assessment must be provided. Providing appropriate evidence for validity is not a simple undertaking and requires multiple sources of evidence collected through a range of methods (Bachman, 1990). This allows different facets important to validity to be addressed and can thus support claims for the validity of scores on an assessment.

The current work focuses on Kane's (2006) definition which states that validity is about the extent to which the inferences made on the basis of the assessment outcomes are appropriate. Given that a key inference is usually that the scores reflect ability or attainment in relation to a particular predefined set of knowledge, understanding and skills, evaluating validity will include considering whether the assessment is measuring what it was intended to measure. Cambridge Assessment sees a vital aspect of validity as "the extent to which the inferences which are made on the basis of the outcomes of the assessment are meaningful, useful and appropriate" (2009, p.8) and argues that the concern for validation "begins with consideration of the extent to which the assessment is assessing what it is intended to assess and flows out to the uses to which the information from the assessment is being put" (2009, p.8).

A debated issue in validity theory is whether the social and personal consequences of assessments should be included within the conceptualisation of validity. This includes issues such as backwash onto classroom practices, and the consequences for individual students of assessment outcomes being used in particular ways. A number of key theorists, including Kane (2006) and Messick (1989) include consideration of consequences within the notion of validity. However, this is somewhat problematic in how it relates to the definition of validity, since not all types of consequences can be considered to relate to the appropriateness of interpretations and uses of test scores. For example, consequences in terms of classroom practices which prepare students for examinations do not relate directly to uses or interpretations of scores. Nonetheless, the consequences are agreed to be important, and arguably fall within a broader notion of the validity of assessment systems and associated curricula. An assessment agency cannot be held responsible for all possible uses of the outcomes of its assessments, but it can take responsibility for being very clear regarding legitimate uses and provide appropriate guidance.

The current line of research aimed to design a set of methods for validating UK qualifications such as A levels and their international counterparts. It is intended that these can later be used on a routine basis or as part of an ongoing validation programme. As the methods need to be

underpinned by theoretical understandings of validity, relevant literature was reviewed to develop a standpoint from which to work. There are significant challenges in doing this, not least because of issues around the conceptualisation of validity to be taken and the boundaries of what should be considered in a validation study.

A number of frameworks for validation have previously been proposed (e.g. Cronbach, 1988; Frederiksen and Collins, 1989; Linn, Baker and Dunbar, 1991; Messick, 1989; 1995; Crooks, Kane & Cohen, 1996; Mislevy, Steinberg and Almond, 2002; Shaw and Weir, 2007). However, these tend to involve substantial technical language, to sometimes be specific to particular assessment contexts, and often fail to suggest a set of methods to be used.

Our aim was to develop a comprehensive framework for validation that includes aspects from key theoretical models, but is more accessible and provides an associated set of methods (though the exact methods to be used may vary depending on the nature of the assessment to be validated).

Initial framework development

This research began by drawing on existing models for validation in various contexts to develop a new framework by which to structure validation exercises for general qualifications. This framework takes the form of a list of validity questions, each of which is to be answered by the collection of relevant evidence. The validity questions are structured within three areas as shown in Figure 1. The findings of validation exercises based on the framework would present '*Evidence for validity*' and any potential '*Threats to validity*'. Any identified threats to validity might provide advice for test development in future sessions, or might suggest recommendations for changes to an aspect of the qualification, its administration and procedures or associated documentation. For a full description of the development of the framework please see Shaw, Crisp and Johnson (2009).

Figure 1 – Validation framework questions

1. Assessment purpose(s) and underlying constructs
- 1.1) What is (or are) the main declared purpose(s) of the assessment and are they clearly communicated?
- 1.2) What are the constructs that we intend to assess and are the tasks appropriately designed to elicit these constructs?
- 1.3) Do the tasks elicit performances that reflect the intended constructs?

2. Adequate sampling of domain, reliability and generalisability
- 2.1) Do the tasks adequately sample the constructs that are important to the domain?
- 2.2) Are the scores dependable measures of the intended constructs?

3. Impact and inferences
- 3.1) Is guidance in place so that teachers know how to prepare students for the assessments such that negative effects on classroom practice are avoided?
- 3.2) Is guidance in place so that teachers and others know what scores/grades mean and how the outcomes should be used?
- 3.3) Does the assessment achieve the main declared purpose(s)?

The intention is that by collecting evidence relating to each of the components of validity represented by the questions in the framework, an awarding body can provide justification for the validity of its assessments. The aim is to move towards a set of methods that can be operationalised periodically for all of an Awarding Body's qualifications. Thus, an initial set of methods was devised drawing, where possible, on previous relevant research methods. By facilitating the collection of evidence relating to each question in the framework, the methods give a view of the extent to which the interpretations and uses of an assessment can be considered valid. Multiple sources of evidence are required in order to provide proof that certain inferences are justified.

Piloting with A level Geography

The provisional set of methods was piloted on the assessments involved in an A level geography syllabus which is available internationally. This A level is assessed through three written exam papers.

The piloting used a broad set of methods to explore the different validity questions in the framework. For practical reasons, it would not be possible to use all of these methods operationally for all of an awarding body's qualifications, but this pilot intentionally employed more methods than might normally be practical in order to identify which are most valuable in providing validity evidence.

The set of methods used involved:

- a series of tasks conducted by geography experts (four senior examiners and two external experts) such as identifying assessment constructs, rating the coverage of Assessment Objective subcomponents, and rating the demands of tasks;
- document reviews, for example, in relation to guidance on teaching practice;
- statistical analyses of item level data, including Rasch analysis;
- a multiple re-marking study, involving five markers for each paper, to explore marking reliability;
- questionnaires to teachers and to higher education institutions;
- interviews with students after they had answered example exam questions.

The various methods and analyses allowed consideration of the evidence in relation to each of the questions in the framework for A level Geography. For each, evidence for validity and any possible threats to validity could be identified. For example, a sample of scripts was obtained and the scores were analysed using various statistical methods including Item Response Theory. This provides some evidence relating to question 1.3 in the framework (see Figure 1) about whether the assessment measures the intended constructs. This offered the following insights:

- ***Evidence for validity –*** Few excessively easy, excessively difficult or misfitting questions were identified. Additionally, the difficulty measures for different optional questions were fairly similar, suggesting reasonable comparability.
- ***Possible threats to validity –*** One question part showed clear (but slight) misfit for a number of reasons.

To give another example, the questionnaire to teachers included questions about the intended meaning and uses of scores and grades and guidance provided by the examination board, thus relating to validity question 3.2 in the framework. The evidence this provided can be summarised as follows:

- ***Evidence for validity -*** Teachers reportedly knew how to use exam scores/grades to inform their teaching. Most teachers felt that the guidance available helped them advise students on their future education and/or employment.
- ***Possible threats to validity –*** Some teachers felt that more guidance could be available on the meaning and use of scores/grades.

The available evidence, from all methods and analyses, were later synthesised in order to provide an overall evaluation of the validity argument. Overall, the findings from the piloting with A level Geography suggest substantial support for the validity of the assessments. However, there were a few minor areas of concern which should be addressed to further increase the validity of the qualification's assessments. These issues have been fed back to the examining team and relevant assessment personnel.

Issues and challenges arising during piloting

The experience of piloting, feedback and discussion with colleagues and further consideration of the literature on validity left us with a number of challenging issues that we felt the framework and methods had not adequately addressed in the pilot. This section will discuss a number of these issues and how we have moved forward in dealing with them.

Issue 1: Which conceptualisation of validity should be used?

This work adopted the generally accepted theorisation of validity as about the appropriateness of inferences and uses of test results. However, there are subtly distinct conceptualisations within this view and the issue of which of these underpins the current work was raised during feedback on the piloting. These differences tend to relate to the boundaries of what is or is not considered part of validity. There are, for example, important differences between the concept of validity proposed by Cronbach (1988) and that proposed by Messick (1989). Cronbach was primarily concerned with evaluation: validity more broadly interpreted. However, Messick's principal concern was with validity, narrowly interpreted. A narrowly conceived view of validity involves identifying the kind of inferences that need to be drawn from test results, and the accuracy with which they need to be drawn given the particular use (or uses) to which they are to be put. Messick was interested in consequences but mainly in terms of the extent to which they shed light on whether inferences from results might lack validity. For example, if a school were only teaching half of the syllabus but were getting good results, one might question whether inferences from results in relation to the whole syllabus are valid. Messick argued that tests which have been used to ill effect can still be quite valid. Cronbach (1988) was also interested in whether the impacts of assessment were good or bad and considered this part of validity. So if tests have had negative consequences on teaching or on students' futures, Cronbach would consider this a threat to validity but Messick would not.

Thus, in the continuation of this work, we have attempted to clarify this, which also relates to the scope of responsibility of the exam board. For example, it may well be the exam board's responsibility to provide guidance on how scores should be used and on good teaching in relation to the assessments, but it is arguably not the exam board's responsibility if a stakeholder chooses to act in other ways not supported by this guidance.

Issue 2: Multiple purposes of educational assessment

It important to be clear about the purposes of an assessment because different uses result in different kinds of inferences being made from test results, and it is these uses/inferences that we must validate. In the pilot with A level Geography, purposes specific to the subject and qualification were identified by the examiners and experts involved as part of one of the tasks they conducted. Whilst this was reasonably successful, on reflection it should be the exam board itself that sets out the intended test score uses at the level of the qualification, rather than the level of the subject/syllabus. It is these uses or purposes that the exam board is claiming the assessment results to be appropriate for. Newton (2007) sets out three main types of purposes: to generate a particular kind of result; to allow a particular kind of decision; and to bring about a particular kind of impact. He lists a wide range of different possible purposes including, for example: **guidance** (identify most suitable courses, or vocations for students given their aptitudes); **qualification** (decide whether students are sufficiently qualified or equipped to succeed in a course or role); **selection** (predict which students will be most successful in a course or role and select between them); **formative** (identify students' learning needs and guide teaching). The purposes of different qualifications will vary and will often be multiple.

There are no explicit statements of purposes made by exam boards for International A levels, although some are implicit (e.g. that results can be used for application to university). To move forward with the issue of purpose, we chose to consult a number of key exam board personnel involved in A level qualifications using the purposes identified by Newton (2007) as a basis for discussion. A list of purposes was determined through consideration of

commonalities in views, and consideration of where the line should be drawn in terms of the purposes for which the exam board claims results are appropriate and thus is responsible for evidencing. This list can then be used as the structure against which to evaluate the appropriateness of uses and inferences related to these purposes. With further internal consultation, it might later be possible to publish an authorised list of purposes in relation to particular qualifications and thus improve transparency.

Issue 3: The interpretive argument

In the early stages of the current work and the piloting with A level Geography, the inferences intended to be made from test sores were implicit within the framework structure. However, in revisiting the literature it became apparent how important it is to be explicit regarding the inferences in order to ensure that the data collected allow each inference to be validated. According to Kane (2006) conducting validation should include setting out the *interpretive argument* for an assessment. This involves specifying the inferences that we claim it is appropriate to make from assessment outcomes and the assumptions underpinning these inferences. For example, if we intend to interpret scores as a reflection of achievement in a given area, this involves inferring that performance on test tasks reflects relevant constructs, that test scores accurately reflect relevant constructs as elicited during the test, and that test scores can be used to infer likely competence in all possible test tasks within the domain of the syllabus. Each of these inferences would have associated assumptions such as, in relation to scoring, that the scoring rules are appropriate to reflect intended constructs, and that the scoring rules are accurately and consistently applied.

Mapping out the interpretive argument in this way is not an easy task and there are few existing examples of full attempts at doing so for specific qualifications (see Chapelle, Enright and Jamieson, 2008, for one of the few examples). However, this task is necessary in order to ensure that a validation exercise can evaluate each inference, and in order to ensure that the validation exercise would be respected by theorists. Thus, to deal with this issue, we used the purposes to help consider the interpretations and inferences to be made and with these in mind revisited the validation framework designed in the early stages of this project. The interpretive argument was specified for International A levels and the framework revised and linked to this.

Issue 4: Methods – appropriate and practical

It was apparent from the piloting with A level Geography that there is an ongoing challenge relating to the need for various types of validity evidence in relation to each of the validity questions used, or in relation to each proposed inference. A substantial research effort and resource is required for this, and there are decisions to be made in relation to making such a set of methods manageable on a routine basis or as part of a long term monitoring programme considering different qualifications and subjects. The set of methods used in the pilot has been revised to give a streamlined subset of methods. Methods have been selected on the basis of how useful they were in providing evidence to evaluate validity and based on their practicality. In addition, some revisions have been made to the previously used methods in light of experience, and one alternative method has been added to reflect changes to the framework. However, even this revised set of methods requires substantial research resource.

Issue 5: Evaluating the validity arguments

Having set out the interpretive argument and used various methods and analyses to provide evidence relating to validity, the validity arguments being put forward must be evaluated. There are challenges in this respect with regard to how much positive evidence is sufficient, and are the inferences and interpretations sufficiently supported. Again, Kane (2006) provides some guidance in this area. After an initial ‘development stage’ in which the inferences and assumptions are set out and an ‘appraisal stage’ of gathering analytical and empirical evidence to review the interpretive argument, Kane proposes that there is an ‘evaluation stage’. The latter involves satisfying three questions: Does the interpretive argument address

the correct inferences and assumptions?; Are the inferences justified?; Is the validity argument as a whole plausible? Whilst this provides the researcher with a structure, and Kane's work provides greater detail on answering these questions, there is still substantial judgement involved, for example, what constitutes a 'plausible' argument? An associated challenge relates to how to make an overall judgement of validity given the need to validate each inference separately.

A validation exercise with A level Physics

Having worked through these issues, the revised framework (incorporating the interpretive argument) and refined set of methods has recently been applied to explore the validity of an International A level Physics. The purposes were defined as described above, and the interpretive argument was set out with the intended inferences and underpinning assumptions specified. This was linked to a restructured, refined version of the validation framework proposed in the early stages of this research. A streamlined set of methods were applied and the data and analyses used to provide evidence in relation to claims about the qualifications assessments.

Issues remaining

A number of unresolved issues remain. For example, there is a difficult balance to be met between practical manageability and comprehensiveness of evidence. The resourcing of validation work is not trivial and there is a potential need for an ongoing programme of research which investigates validity for several qualifications each year. Additionally, how much evidence is sufficient to be confident of validity, and can this be collected on a one off basis. Sireci (2007) argues that validation should be a continuous, ongoing exercise. A further issue relates to what extent the interpretive argument, validation framework, and set of methods would need to be adapted for different types of qualifications and assessments. It might, for example, be relatively straightforward to revise the structures for A level into some that are appropriate for GCSE which are broadly similar in some of their aims and the nature of the assessments. However, much more adaptation might be required for vocational qualifications, for example.

Conclusion

This research has made progress in developing a framework for validation that is suitable for traditional written examinations and in showing that this can be applied to assessments through use of a variety of methods and analyses. This research has also highlighted the challenges faced when validating the intended interpretation of test scores and their relevance to the proposed uses of those scores. Whilst some of these issues have been resolved to some extent, others remain and it is hoped that the continuation of this research and discussion with others will provide a way forward in this and other contexts.

References

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

Cambridge Assessment. (2009). *The Cambridge Approach. Principles for designing, administering and evaluating assessment.* Available online at: http://www.cambridgeassessment.org.uk/ca/digitalAssets/181348_cambridge_approach.pdf

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

Cronbach, L.J. (1988). Five perspectives on validity argument. In H. Wainer and H. Braun (Eds.), *Test Validity* (pp. 3-17), Hillsdale, NJ: Lawrence Erlbaum.

Crooks, T.J., Kane, M.T. and Cohen, A.S. (1996). 'Threats to the valid use of assessments'. *Assessment in Education: Principles, Policy and Practice*, 3 (3), 265-286.

Frederiksen, J.R. & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, 18 (9), 27-32.

Kane, M.T. (2006). Validation. In R.L. Brennan (Ed.), *Education Measurement* (4th ed.). Westport: Praeger.

Linn, R.L., Baker, E.L. & Dunbar, S.B. (1991). Complex, performance-based assessment: expectations and validation criteria. *Educational Researcher*, 20 (8), 15-21.

Messick, S. (1989). Validity. In R. Linn (Ed.) Educational Measurement (pp. 13-103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741-749.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2002). Design and analysis in task-based language assessment. *Language Testing*. Special Issue: Interpretation, intended uses, and designs in task-based language, 19(4), 477-496.

Newton, P.E. (2007). Evaluating assessment systems. London: QCA. Available online at: http://www.qcda.gov.uk/resources/assets/Evaluating_Assessment_Systems1.pdf.

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

Shaw, S.D., Crisp, V. & Johnson, N. (2009). *A proposed framework for evidencing assessment validity in large-scale, high-stakes international examinations.* A paper presented at the Association for Educational Assessment in Europe, 10th Annual Conference, Malta, November 2009.

Shaw, S.D. and Weir, C.J. (2007). *Examining Writing: Research and Practice in assessing second language writing*. Cambridge: Cambridge University Press.

Wood, R. (1991). *Assessment and Testing: A survey of research*. Cambridge: Cambridge University Press.