References

Arlett, S. (2002). *A comparability study in VCE Health and Social Care units 1, 2 and 5. A review of the examination requirements and a report on the cross moderation exercise. A study based on the summer 2002 examinations*. Organised by the Assessment and Qualification Alliance on behalf of the Joint Council for General Qualifications.

Arlett, S. (2003). *A comparability study in VCE Health and Social Care units 3, 4 and 6. A review of the examination requirements and a report on the cross moderation exercise. A study based on the summer 2002 examinations*. Organised by the Assessment and Qualification Alliance on behalf of the Joint Council for General Qualifications.

Barry, K. (1997). An analysis of the relative demands of advanced GNVQ science and A-level Chemistry. *Journal of Further and Higher Education*, **21**, 1, 45–53.

Bloom, B.S. (Ed.) (1956). *Taxonomy of Educational Objectives – Book 1 – Cognitive Domain*. Michigan: Longman.

Coles, M. & Matthews, A. (1995). *Fitness for purpose: a means of comparing qualifications. A report to Sir Ron Dearing to be considered as part of his review of 16–19 education*.

Coles, M. & Matthews, A. (1998). *Comparing qualifications – Fitness for purpose. Methodology paper*. London: Qualifications and Curriculum Authority.

Crisp, V. & Novaković, N. (2009). Are all assessments equal? The comparability of demands of college-based assessments in a vocationally-related qualification. *Research in Post Compulsory Education*, **14**, 1, 1–18.

Edwards, E. & Adams, R. (2003). *A comparability study in GCE Advanced Level Geography including the Scottish Advanced Higher Grade Examinations. A review of the examination requirements and a report on the cross moderation exercise*. A study based on the Summer 2002 Examination and organised by WJEC on behalf of the Joint Council for General Qualifications.

Gagne, R.M. (1985). *The conditions of learning and theory of instruction* (4th Ed.). New York: Holt, Rinehart and Winston.

Guthrie, K. (2003). *A comparability study in GCE business studies units 4, 5, and 6 VCE business units 4, 5, and 6. A review of the examination requirements and a report on the cross moderation exercise*. A study based on the summer 2002 examinations. Organised by EdExcel on behalf of the Joint Council for General Qualifications.

Hughes, S., Pollitt, A. & Ahmed, A. (1998). *The development of a tool for gauging the demands of GCSE and A-level examination questions*. Paper presented at the British Educational Research Association Annual Conference, The Queen's University of Belfast.

Kline, P. (1986). *A handbook of test construction: Introduction to psychometric design*. New York: Methuen.

Learning and Skills Council (2009). Jargon Buster http://www.lsc.gov.uk/Jargonbuster/Vocational+certificate+of+education.htm [Accessed September 2009]

Mitchel, L. & Bartram, D. (1994). The place of knowledge and understanding in the development of National Vocational and Scottish Vocational Qualifications. In: *Competence & assessment briefing series no. 10*.

OCR (2009). http://www.ocr.org.uk/qualifications/type/nvq/index.html [Accessed January 2010]

OCR (2009). http://www.ocr.org.uk/qualifications/type/vrq/index.html [Accessed January 2010]

Ofqual (2008). Glossary http://www.ofqual.gov.uk/501.aspx#I [Accessed January 2010]

Pollitt, A., Ahmed, A. & Crisp, V. (2007). The demand of examination syllabuses and question papers, 166–206. In: P. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.) *Techniques for monitoring the comparability of examination standards*. London: QCA.

QCA (2006a). *Comparability study of assessment practice: Personal license holder qualifications*, QCA/06/2709 http://www.ofqual.gov.uk/files/personal_licence_holder_quals_comparability_study.pdf

QCA (2006b). *Comparability study of assessment practice Door supervision qualifications* QCA/06/2710 [on line] Available at: http://www.ofqual.gov.uk/files/door_supervision_quals_comparability_report.pdf [Accessed September 2009].

QCDA (undated). Glossary http://testsandexams.qcda.gov.uk/15862.aspx#S [Accessed January 2010].

Savory, C., Hodgson, A. & Spours, K. (2003). *The Advanced Vocational Certificate of Education (AVCE): A general or vocational qualification? Broadening the Advanced Level Curriculum*. IoE/Nuffield Series Number 7, School of Lifelong Education and International Development, Institute of Education, University of London [on line] http://www.ioe.ac.uk/schools/leid/nuff/rep7.pdf [Accessed September 2009]

SCAA (1995). *Report of a comparability exercise into GCE and GNVQ business*. London: School Curriculum and Assessment Authority.

ASSURING QUALITY IN ASSESSMENT

Developing and piloting a framework for the validation of A levels

Stuart Shaw CIE Research **and Victoria Crisp** Research Division

Introduction

This article 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 and 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).

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.

Revising the framework

A further, ongoing, phase of this research aims to build on and refine the framework and methods, in order to move towards a validation model that is more manageable on a routine basis or as part of a long term monitoring programme considering different qualifications and subjects.

The experience of the piloting, feedback and discussion with colleagues and further consideration of the literature on validity has led to refinement of the framework. Changes have been made in relation to how it deals with assessment purposes and also in relation to evaluating qualifications as preparation for future study, if they are used for selection purposes.

Applying a revised set of methods to A level Physics

The set of methods used in the pilot with A level Geography 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 additional method has been added to reflect changes to the framework.

The revised set of methods is currently being used with International A level Physics, to provide evidence to support the claim for its validity, and to identify any potential threats to validity for this qualification such that they can be addressed.

Reflections on the work so far

This project so far 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. These challenges include issues relating to:

- the view of validity adopted and its boundaries;
- the scope and sufficiency of evidence;
- balancing operational manageability and comprehensiveness of evidence;
- a possible need for additional frameworks and sets of methods for different types of qualifications and assessments.

It is hoped that the continuation of this research will help resolve some of these challenges and provide a way forward. Eventually, it is proposed that validation evidence will be collected and presented in an operationally-orientated portfolio for any one particular qualification. This will show more clearly how an appropriate methodology can be used as part of regular monitoring of assessment validity.

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

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

Crooks, T.J., Kane, M.T. & 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*. 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.

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. & 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.

EXAMINATIONS RESEARCH

Statistical Reports

The Statistics Team Research Division

The ongoing 'Statistics Reports Series' provides statistical summaries of various aspects of the English examination system such as trends in pupil attainment, qualifications choice, combinations of subjects and subject provision at school. These reports, produced using national-level examination data, are available in .pdf format on the Cambridge Assessment website: http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Statistical_Reports

The following reports have been published since Issue 9 (January 2010) of *Research Matters*:

- Statistics Report Series No.14: A-level candidates attaining 3 or more 'A' grades in England, 2006–2009
- Statistics Report Series No.15: Provision of science subjects at GCSE, 2009
- Statistics Report Series No.16: A-level uptake and results by gender, 2002–2008
- Statistics Report Series No.17: GCSE uptake and results by gender, 2002–2008
- Statistics Report Series No.18: A-level uptake and results by school type, 2002–2008
- Statistics Report Series No.19: GCSE uptake and results by school type, 2002–2008