Practical validation: organisational approaches to large-scale evaluation and continuous improvement

Phineas Hodson, Cambridge International Examinations
hodson.p@cie.org.uk

Abstract

Validity is fundamental to the development and administration of tests and a vital concern for test providers, yet validating a test's uses to the highest technical standards is an enormous and never-ending undertaking (Cronbach, 1971). The scale of the challenge is multiplied when a test provider offers hundreds of separate tests. This paper explores how providers of general qualifications can rise to this challenge, drawing on examples from international qualifications for 14-18-year-old learners.

Resources for validation will always be limited so prioritisation is a necessity. Validators must choose between depth and breadth of validation: to comprehensively validate the uses of a few tests, or cover the uses of more tests in less depth. Alternatively, one aspect of validity - such as marking reliability - can be evaluated across many tests. Using validation as a driver of qualification improvement is explored, with an emphasis on maximising impact. Strategies proposed include aligning validation with qualification development and emphasising flexibility and attention to stakeholders' needs. Finally, the outcomes for an ongoing validation programme are discussed. While validation will never be finished, since resources are finite and test uses require continual revalidation, a programme can address stakeholders' concerns and drive continuous improvement of qualifications.

Key words: validation; validity; evaluation

Introduction

The meaning of validity, and therefore the methods appropriate for validation, have evolved over time and remain contested (Newton & Shaw, 2014). The approaches discussed here are situated in the construct validity of Messick (1989) and the argumentation approach of Kane (1992), though many of the points discussed might be relevant to a validation effort drawing on a different theoretical position. Construct validity itself is a fairly broad church, which has evolved over time and accommodates several competing interpretations.

The origins of construct validity

Construct validity, as first laid out by Cronbach and Meehl (1955), requires that validation consider the internal trait or characteristic (construct) which a test measures. Defining the construct measured by a test requires a nomological network in which the construct is defined and given meaning by a set of theories. This requirement for a coherent and tightly-defined set of theories was acknowledged to be problematic:

> In practice, of course, even the most advanced physical sciences only approximate this ideal. … Psychology works with crude, half-explicit formulations.
> (Cronbach & Meehl, 1955, pp. 293–294)

This focus on the underlying construct was accompanied by an injunction to consider and make explicit the multiple interpretations and uses of a testing procedure. Cronbach and Meehl envisaged a complex, scientific theory linking constructs, measurement instruments (tests) and interpretations and uses. This theory-driven formulation of validation laid a heavy burden on validators: they had to draw together a well-defined theoretical network to define their constructs, show that their test measured those constructs, then create a scientific justification for using the test outcomes in the intended ways.

Messick's unified theory

The enormous demands of construct validation, both in terms of generating a viable nomological network and in creating a full scientific justification for the uses of a test, led to attempts to make the process more manageable. A "weak programme" emerged that largely did away with the nomological network but retained the focus on interpretations and uses justified by multiple sources of evidence. This approach was criticised by Cronbach (1988) as "sheer exploratory empiricism", with any kind of evidence accepted and the scientific emphasis of the original (strong) programme of construct validity discarded. Kane (2001) also criticised the weak programme, but acknowledged that its emergence may have been inevitable, given the "dearth of highly developed formal theories in education and the social sciences".

The difficulties of applying the strong programme and the deficiencies of the weak programme led to Messick's (1989) reformulation of the construct validity model (the unified model). The requirement for a nomological network was pushed into the background but the emphasis on the evaluation of defined interpretations and uses remained:

> Validity is an integrated evaluative judgment 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. [emphasis original]
> (Messick, 1989, p. 13)

While the removal of some of the requirements for theory made validation more practical, the need for a scientific justification of proposed interpretations and uses based on multiple sources of evidence remained. The renewed emphasis on interpretations and uses did something to combat the "dragnet empiricism" (Cronbach, 1989) of the weak programme, but validity was still characterised in such a way that "almost any information gathered in the process of developing or using a test is relevant to its validity" (Anastasi, 1986). The evidence which might be gathered extended not only across many contexts but also over time: Cronbach (1989) characterised validation as "a lengthy, even endless process". This still left validators to define the boundaries of their own studies, as few could find the resources for the intensive, wide-ranging and unending validation effort the unified model demanded.

Validation through argument

In response to the difficulties inherent in the unified model of validation, a model based on arguments was developed "mainly as a way of facilitating the process of validation" (Kane, 2013). Drawing on Toulmin's (1958) work on inferential arguments, Cronbach (1988) proposed the use of a validity argument to evaluate the theory and evidence supporting a test use. The argument-based structure "plays the role that a formal theory plays in the strong program of construct validity" (Kane, 2013), thereby avoiding the difficulties inherent in working with a nomological network but retaining logical rigour and offering guidance for validators. This change was possible partly due to an epistemological shift; validation now produced a presumptive inference rather than the more demanding standards of proof required in earlier formulations.

The use of argument was developed by Kane (1992, 2006, 2013) who proposed that validation requires two arguments: an interpretation and use argument (IUA) and a validity argument. The IUA lays out "a

rationale for whatever claims are being made by the interpretation and use" (Kane, 2013). Using Toulmin's (1958) model of argumentation, the inferences in the IUA are laid out and then supported with theoretical justifications (warrants) and empirical evidence (backing). Exceptions and uncertainties are accommodated as qualifiers, which give cases where the conclusion does not apply, or where it applies probabilistically rather than deterministically. The IUA must be constructed anew each time an interpretation or use is validated, since its structure is dependent on the claims being made.

When the IUA is complete it sets out the chain of reasoning and empirical evidence that supports the proposed interpretation and use. This structure is evaluated by the validity argument. The logical structure is examined in terms of coherence and clarity and the empirical evidence is judged on its sufficiency and quality. The amount of theoretical and empirical support required is dependent on the claim being made:

> Strong claims (e.g., causal inferences or predictions of future performance in different contexts) typically would require extensive empirical support. The most questionable assumptions should get the most attention in the validity argument.
> (Kane, 2013, p.14)

The level of certainty achieved by validity arguments does not amount to proof - Kane (2013) states that "they cannot be proven" - but instead the aim is plausibility (Kane, 1990). This lower level of certainty allows more to be assumed than would be the case under a more demanding epistemology, since "many inferences and assumptions are sufficiently plausible *a priori* to be accepted without additional evidence, unless there is some reason to doubt them in a particular case" (Kane, 2013). This circumscribes the task of validator, who must only provide empirical backing for those inferences and assumptions which seem to be questionable. This approach can also strengthen a validation study by channelling research effort towards falsification rather than building up confirmatory analyses where the argument is already strong.

The model of validity through argumentation gives a manageable structure to a validation effort and provides a point at which validation can stop. This approach has been used successfully to validate international general qualifications (Shaw, Crisp & Johnson, 2012; Shaw & Crisp, 2012), along with personality tests (Van Rooy, Viswesvaran & Pluta, 2005), clinical tests (Corrigan & Buican, 1995), selection tests (Kuncel & Sackett, 2014) and certification tests (Kane, 2004). The following section considers how this model scales up for validation of large numbers of qualifications, each of which has multiple interpretations and uses.

Approaches to large-scale validation

To validate even just one use of a test using the unified model proposed by Messick (1989) is a daunting task, but examination boards do not have just one test with one use. International general qualifications are used in many ways across many contexts and countries and a single test provider may have many hundreds of separate tests. The enormity of the task facing the validator clearly precludes a single, one-off effort to validate them all. It is not even possible to gradually validate each test in turn until the task is complete, since validating a test is a never-ending process (Sireci, 2007) and old tests are withdrawn and new tests created continually.

The model of validation through argument, as formulated by Kane (2013), allows limits to be placed on the size of each validation study but does not alter the fundamental challenge of validating a large and ever-changing set of tests, each of which has multiple uses. If the researcher accepts that validating everything is not practical under either Kane's or Messick's model, what should be done?

Be clear about the aim

Before deciding what and how to validate it is necessary to be clear about why you are validating. As validation takes place within an organisational context, the priorities and assumptions of the organisation will shape the validation programme. The more closely validation is aligned with organisational goals the more likely it is to be useful, though there is potential for conflict between the calls in the literature for "developing a scientifically sound validity argument to support the intended interpretation of test scores" (AERA/APA/NCME, 1999, p. 9) and the more diverse and complex goals of a testing organisation.

Possible reasons for carrying out a validation effort include:

- Responding to the concerns of stakeholders, such as test takers, schools and universities, about validity issues, such as reliability and fairness.
- Contributing to the improvement of tests, through the identification of good practice to disseminate and weaknesses to be rectified.
- Providing information to the users of test outcomes.
- Meeting statutory or regulatory requirements for validity evidence.

These reasons are not mutually exclusive, with continuous improvement in particular likely to be a priority for any test provider. However, by being clear about the aims of a validation programme it is easier to choose the most appropriate means to implement it.

Extending the concept of plausibility

An interpretation and use argument (IUA) is successful if it is plausible; proof is not required (Kane, 1990). Similarly, inferences and assumptions within the argument can be accepted without empirical evidence if they are sufficiently plausible *a priori* (Kane, 2013). There is no defined criterion for plausibility; it is a matter of judgement and degree. Thus, all inferences and assumptions of an IUA will have some degree of inherent plausibility. An entire IUA, and therefore an entire validation effort, could be considered to some degree supported based on *a priori* plausibility alone.

This can be extended further to validation studies not yet undertaken. If a test is produced by an accredited organisation and accepted by universities and employers is it not to some degree plausible that its use in selection has a reasonable level of validity? This is the approach to validity adopted by the government regulator of qualifications in England and Wales (Ofqual), which focuses not on individual tests but the organisations which produce them (Jones, 2011, p.18). The more validation effort that goes into investigating a proposed interpretation or use the stronger the IUA becomes, but there is no dichotomy between unvalidated and validated.

By breaking down the division between unvalidated and validated a validation programme is freed to concentrate effort where it will be most valuable, in terms of the aims of the endeavour, rather than attempting to validate all uses of all tests. The wide use of plausibility also allows for the use of weaker evidence than would otherwise be pertinent: when an inference is supported only by weak *a priori* plausibility even quite tangential evidence can tip the balance.

Embedding validity in professional practice

When many hundreds of tests require validation and only a handful of research staff are available there is an obvious mismatch between requirements and resources. One way to approach this is to transfer primary responsibility for validation to those with operational responsibility for tests. Such a system of evaluation and improvement as professional practice exists in other fields, with the practice of clinical audits providing a useful model. Clinical audit is defined as:

> a quality improvement process that seeks to improve patient care and outcomes through systematic review of care against explicit criteria and the implementation of change. (National Institute for Clinical Excellence, 2002, p. 1)

The model for evaluation in a clinical audit is that of the multi-disciplinary team supported by specialist audit staff. An analogous validation effort would involve staff from across an examination board evaluating their own work through the lens of validity, with research staff providing assistance. Users can also be included as collaborators, rather than merely as a source of data (Balogh, Simpson & Bond, 1995).This approach embeds in the validation effort the knowledge and expertise of both those who carry out the assessment processes underpinning validity and those, such as university admissions tutors, who have hands-on experience of the uses of test outcomes. A further advantage stemming from the ownership of validation by those involved in test delivery and use is that impact is much more easily achieved: those able to implement improvements are part of the validation effort, not outsiders who must be persuaded to embrace change.

The aim of an audit is to produce improvements in procedure, and validation can be informed by this focus. A structure under which a test process is evaluated, improved based on the findings, and then re-evaluated makes test improvement an integral part of validation. The appropriateness of this approach depends on the purpose of the validation effort, being especially suited to a continuous improvement focus, as it moves away from the model of validation as a detached, scientific endeavour. Under this model validation becomes a project more than a research study and the outcome is an improvement in the test as much as a report evaluating validity.

While creating a programme of this nature might seem both difficult and resource-intensive, much of the change would simply be drawing existing work into a validity framework. Any responsible test provider will monitor the quality of marking, the manageability of tests for users and the fairness of outcomes, even if such considerations are not conceptualised as validation. By drawing on insights from the literature and fitting disparate pieces of evidence together in an argument-based validation framework the researcher can strengthen and integrate these existing processes.

Allocating validation effort

Whether validation is undertaken as a researcher-led or practitioner-led activity, decisions must be made on where to concentrate effort. For a test provider offering hundreds of tests, each with multiple uses and interpretations, the ground that could be covered is enormous. The interrelationship between the types of validity evidence, the tests and the test uses can be visualised with a diagram (fig.1). To make any individual validation study feasible, a subset of the possible work must be selected which accords with the aims of the endeavour and can be completed in a reasonable length of time.

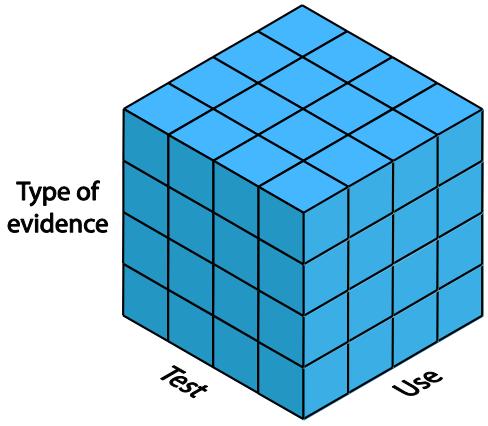

Figure 1. Possible studies based on a choice of test, type of evidence and use.

The most obvious approach is to look at the interpretations and uses of one test (fig. 2). This fits with the traditional conceptualisation of validity as being concerned with a specific test and its uses. The advantage of this method is that it allows in-depth understanding of that test and its interpretations and uses, and a full interpretation and use argument (IUA) and validity argument can be produced. This approach is particularly effective when used with tests which are about to be redeveloped, as any improvements can be implemented immediately. It is also valuable where stakeholders have concerns about a particular test and evidence is required to substantiate or allay these concerns.

The strength of this method, its deep investigation of one test, also constitutes a weakness: little is learned about other tests. If only a few studies can be conducted a focus on the uses of successive individual tests will furnish good validation evidence for those tests but very little for all the rest.

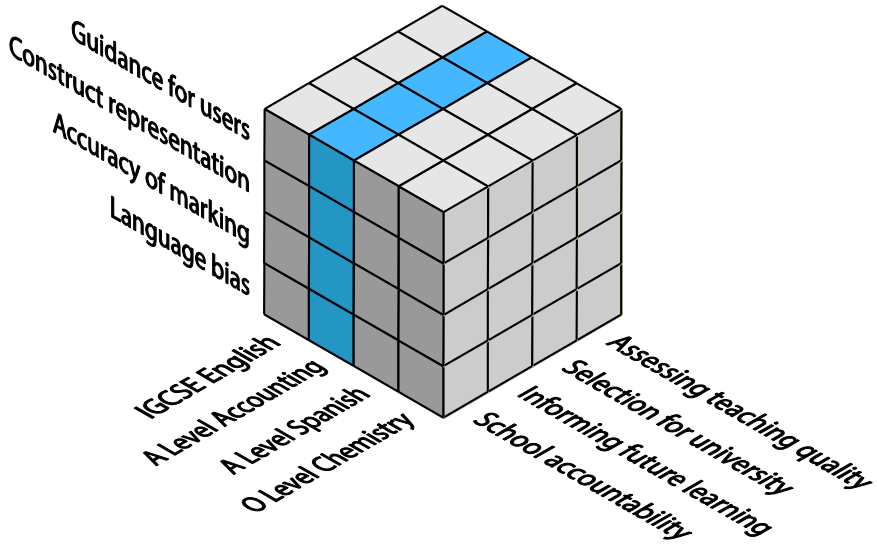

Figure 2. Focus on a specific test.

A second method is to focus on one type of evidence, cutting across multiple tests and multiple interpretations and uses (fig. 3). This approach has been used to consider quality of marking (Ahmed & Pollitt, 2011), assessment instruments (Ofqual, 2011) and the weighting of assessment objectives (Stringer, 2014). This approach allows some validation evidence to be gathered for many test uses but cannot yield a full validation study for any of them. In terms of evidence generated versus time expended this can be very efficient, since running many similar analyses in parallel is much quicker than running a series of small, discrete analyses. Stringer (2014) provides an example of this approach, with validity evidence relating to the weightings of assessment objectives produced for many tests based on a single statistical procedure.

This type of work is not always seen as validation, partly because it is not always articulated with a larger validation programme. It can be particularly effective where one of the principal aims of the validation effort is to improve tests. Findings from such a study cover many tests and therefore feed naturally into a wide-ranging improvement programme. This type of work can even cover multiple test providers and therefore have even wider impact (Ofqual, 2014).

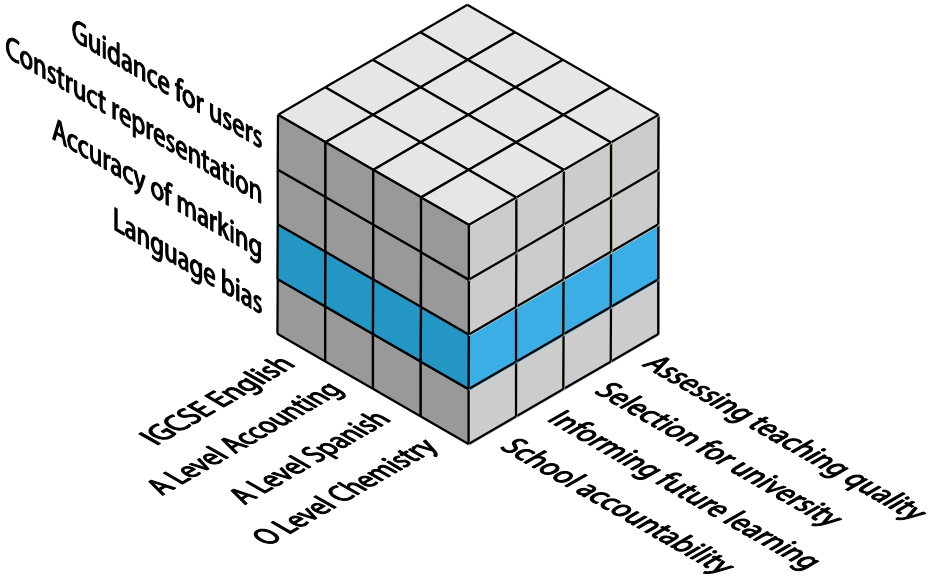

Figure 3. Focus on one type of evidence.

Finally, validation can focus on one particular type of use. This can address system-level concerns about whether a group of tests, such as A Level or the International Baccalaureate, are fit for a specific purpose. Such validation work can be carried out by test providers (Kobrin et al., 2008; Mehta, Suto & Brown, 2012; Shaw & Bailey, 2011) or by test users (Ogg, Zimdars & Heath, 2009; Partington, 2011). This approach can be valuable both for feeding into improvement in tests and for aiding the users of test outcomes in making decisions appropriately.

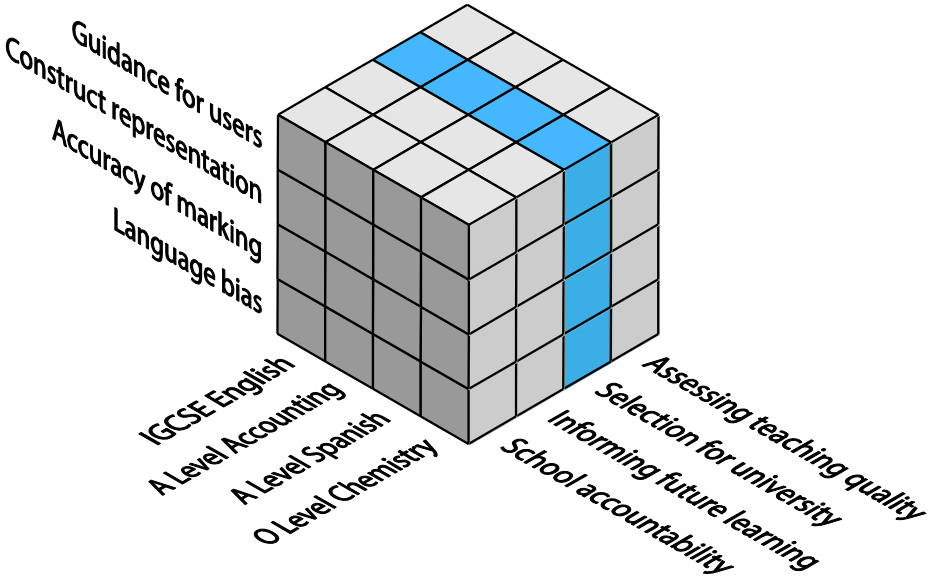

Figure 4. Focus on one type of use.

Focus on falsification

Validation should concentrate on inferences which are weakly supported, or seek out problems in other words, since this is the most productive allocation of effort. Crooks, Kane and Cohen (1996) state that "a chain of reasoning is only as strong as its weakest link"; by uncovering these weak links the real strength of an IUA can be found. This approach also maximises the potential for impact in improving tests, since dealing with significant problems does much more to strengthen the interpretations of test scores than polishing already strong aspects.

While some validators have explicitly incorporated rival hypotheses into their work (Kranzler, Brownell & Miller, 1998) this focus on falsification has sometimes been lacking in validation, since the test providers who carry out most validation can be reluctant to hunt for problems with their tests:

> Despite many statements calling for focus on rival hypotheses, most of those who undertake CV have remained confirmationist. Falsification, obviously, is something we prefer to do unto the constructions of others.
> (Cronbach, 1989, p. 153)

However, by taking a critical position and seriously entertaining a range of alternative hypotheses as part of the validity argument, the overall validation effort is strengthened, not weakened. Cronbach (1980) states that "the job of validation is not to support an interpretation, but to find out what might be wrong with it" and this approach can be extended to the allocation of effort in a large-scale validation programme.

A falsification-based approach can be applied to decisions on what to validate as well as decisions on how to validate. For a large-scale validation programme there will never be time to validate all test uses, but by choosing those which look most questionable, or open to falsification, validation effort is concentrated where it is most needed. The degree to which weaknesses are likely to be the focus of validation is also influenced by the aims and organisational context of the validation programme: where validation is used as a tool for test improvement searching for weaknesses naturally follows but where validation aims to demonstrate a test's value to stakeholders there may be a tendency towards confirmation and the documentation of strengths.

Discussion

The preceding section lays out a set of approaches, rather than a programme, for large-scale validation. No single programme can be prescribed, given that validation efforts have a number of possible aims, and the interests and concerns of stakeholders will be specific to particular test uses in particular contexts. Nonetheless, it is possible to envisage certain common features of such programmes, and what an established large-scale validation effort might achieve.

Such a programme is an ongoing commitment rather than a project to be completed, since more evidence can always be gathered for existing arguments, previous pieces of validation must be periodically revisited and new tests, or new revisions of existing tests, are continually created. That the programme must continue indefinitely is not to say it is a futile, Sisyphean task however. A test provider should be able to meet the validity concerns of stakeholders (AERA/APA/NCME, 1999, p. 17), whether they be couched in the language of validity, fairness or standards, and an ongoing, responsive validation programme is an effective way to do this.

It is also important for test providers to continually seek to improve their tests and better fit them to the needs of stakeholders. As with all instruments of public policy, tests should be continually evaluated:

> We can never afford to be satisfied with the status quo, even if we are still okay, even if our policies are still working. People say, 'if it ain't broke, don't fix it'. I say, if it ain't broke, better maintain it, lubricate it, replace it, upgrade it, try something better and make it work better than before.
> (Prime Minister of Singapore Lee Hsien Loong, National Day Speech, 22 August 2004)

A strong validation programme, embedded within the professional practice of all members of a test providing organisation, is a powerful means to meet this need for continuous improvement and renewal.

References

Ahmed, A. & Pollitt, A. (2011) Improving marking quality through a taxonomy of mark schemes. Assessment in Education: Principles, Policy & Practice, 18, 3.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985) Standards for educational and psychological testing. *Washington, DC: American Psychological Association*.

Anastasi, A. (1986) Evolving concepts of test validation. *Annual Review of Psychology*, *37*,1–15.

Balogh R, Simpson A, Bond S. (1995) Involving clients in clinical audits of mental health services. International Journal for Quality in Health Care, 7: 343–53.

Corrigan, P. W. & Buican, B.B.A. (1995) The Construct Validity of Subjective Quality of Life for the Severely Mentally Ill. Journal of Nervous & Mental Disease, 183, 5.

Cronbach, L.J. (1971) Test validation. In R.L. Thorndike (Ed.). Educational Measurement (2nd edition) (pp.443-507). *American Council on Education. Washington: DC.*

Cronbach, L. J. (1980) Validity on parole: How can we go straight? New Directions for Testing and Measurement: Measuring Achievement Over a Decade, 5, 99–108.

Cronbach, L. J., & Meehl, P. E. (1955) Construct validity in psychological tests. Psychological Bulletin, 52, 281–302.

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

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

Crooks, T., Kane, M., & Cohen, A. (1996) Threats to the valid use of assessments. Assessment in Education, 3, 265–285.

Jones, B. (2011) Regulation and the qualifications market. Centre for education research and policy.

Kane, M. T. (1990) An Argument-based Approach to Validation. ACT Research Report Series 90-13.

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

Kane, M. T. (2001) Current Concerns in Validity Theory. Journal of Educational Measurement, 38: 319–342.

Kane, M. T. (2004) Certification Testing as an Illustration of Argument-Based Validation Measurement: Interdisciplinary Research and Perspectives, 2:3.

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

Kane, M. T. (2013) Validating the Interpretations and Uses of Test Scores. Journal of Educational Measurement, 50: 1–73.

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008) Validity of the SAT® for Predicting First-Year College Grade Point Average. The College Board, New York.

Kranzler, J. H., Brownell, M. T. & Miller, M. D. (1998) The Construct Validity of Curriculum-Based Measurement of Reading: An Empirical Test of a Plausible Rival Hypothesis. Journal of School Psychology, 36:4, 399-415.

Kuncel, N. R. & Sackett, P. R. (2014) Resolving the assessment center construct validity problem (as we know it). Journal of Applied Psychology, 99:1, 38-47.

Mehta, S., Suto, I. & Brown, S. (2012) How effective are curricula for 16 to 19 year olds as preparation for university? A qualitative investigation of lecturers' views. Cambridge Assessment.

Messick, S. (1989) Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education and Macmillan.

National Institute for Clinical Excellence (2002) Principles for Best Practice in Clinical Audit. Radcliffe Medical Press.

Newton, P. E. & Shaw, S. D. (2014) Validity in Educational and Psychological Assessment. London: Sage.

Ofqual (2011) Inquiry into Examination Errors Summer 2011. Ofqual/11/5113.

Ofqual (2014) Review of Quality of Marking in Exams in A Levels, GCSEs and Other Academic Qualifications. Ofqual/14/5379.

Ogg, T., Zimdars, A. & Heath, A. (2009) Schooling effects on degree performance: a comparison of the predictive validity of aptitude testing and secondary school grades at Oxford University. British Educational Research Journal, 35, 5.

Partington, R. (2011) Predictive Effectiveness of Metrics in Admission to the University of Cambridge. Admissions Research Working Party Report.

Shaw, S., & Bailey, C. (2011) Success in the US: Are Cambridge International Assessments Good Preparation for University Study? Journal of College Admission, 213, 6-16.

Shaw, S. and Crisp, V. (2012) An approach to validation: Developing and applying an approach for the validation of general qualifications. Research Matters: A Cambridge Assessment Publication, Special Issue 3.

Shaw, S., Crisp, V. & Johnson, N., (2012) A framework for evidencing assessment validity in large-scale, high-stakes international examinations. Assessment in Education: Principles, Policy & Practice 19:2, pp. 159-176.

Sireci, S. G. (2007) On Validity Theory and Test Validation. Educational Researcher, 36:8, 477-481.

Stringer, N. (2014) The Achieved Weightings of Assessment Objectives as a Source of Validity Evidence. Ofqual/14/5375.

Toulmin, S. (1958) The uses of argument. *Cambridge: Cambridge UP (2003)*.

Van Rooy, D. L. Viswesvaran, C. & Pluta, P. (2005) An Evaluation of Construct Validity: What Is This Thing Called Emotional Intelligence? Human Performance, 18:4, 445-462.