standardisation and have their moderation standards checked by a senior moderator. Those judged to be unsatisfactory will no longer be allowed to undertake moderation and candidates' work in centres that they moderated will need to be re-moderated. Secondly, if a centre has its candidates' work scaled and is unhappy with the adjustments made, they can request a review of the moderation (for a fee). If it is determined that the original moderation is not acceptable then a revised moderation is implemented instead.

References

Ofqual (2011), *GCSE, GCE, Principal Learning and Project Code of Practice: May 2011*. Coventry: The Office of Qualifications and Examinations Regulation. Retrieved from http://ofqual.gov.uk/documents/gcse-gce-principal-learning-and-project-code-of-practice/

OCR (2010). *Moderation of GCE, GCSE and FSMQ centre-assessed units/components: Common principles and practices*. Cambridge: Oxford, Cambridge and RSA.

Reflections on a framework for validation – Five years on

Stuart Shaw Cambridge International Examinations and **Victoria Crisp** Research Division

Abstract

In essence, validation is simple. The basic questions which underlie any validation exercise are: what is being claimed about the test, and are the claims warranted (given all of the evidence). What could be more straightforward? Unfortunately, despite a century of theorising validity, it is still quite unclear exactly how much and what kind of evidence or analysis is required in order to establish a claim to validity. Despite Kane's attempts to simplify validation by developing a methodology to support validation practice, one which is grounded in argumentation (e.g., Kane, 1992), and the "simple, accessible direction for practitioners" (Goldstein & Behuniak, 2011, p.36) provided by the *Standards* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [AERA, APA, & NCME], 2014), good validation studies still prove surprisingly challenging to implement.

In response, a framework for evidencing assessment validity in large-scale, high-stakes examinations and a set of methods for gathering validity evidence was developed in 2008/2009. The framework includes a number of validation questions to be answered by the collection of appropriate evidence and by related analyses. Both framework and methods were piloted and refined. Systematic implementation of the validation framework followed which employs two parallel validation strategies:

1. an experimental validation strategy which entails full post-hoc validation studies undertaken solely by research staff
2. an operational validation strategy which entails the gathering and synthesis of validation evidence currently generated routinely within operational processes.

Five years on, a number of issues have emerged which prompted a review of the validation framework and several conceptual and textual changes to the language of the framework. These changes strengthen the theoretical structure underpinning the framework.

This paper presents the revised framework, and reflects on the original scope of the framework and how this has changed. We also consider the suitability and meaningfulness of the language employed by the framework.

Validation: a task too far?

Samuel Messick's extended account of validity and validation came to dominate the educational and psychological measurement and assessment landscape of the 1980s and 1990s. Instigated by Loevinger (1957), developed and articulated by Messick (1989), and endorsed through the support of significant allies including Robert Guion, Mary Tenopyr and Harold Gulliksen, the essence of validity came to be understood as being fundamentally a unitary concept. Messick's landmark treatise on validity published in the textbook *Educational Measurement* (Messick, 1989) represented the culmination and enunciation of a paradigm shift towards a unified view of validity as articulated in the description of modern construct validity. Measurement was to assume centre stage and came to be the foundation for all construct validity. Since that time, mainstream scholars have consistently affirmed the 'consensus' concerning the nature of validity (e.g., Shepard, 1993; Moss, 1995; Kane, 2001; Downing, 2003; Sireci, 2009) described in the maxim: all validity is construct validity. If validity pivots upon score meaning then by extension construct validation, that is, scientific inquiry into score meaning, is to be understood as the foundation for all validation inquiry. Hence, "… all validation is construct validation." (Cronbach, 1984, p.126).

Tests were to be evaluated holistically, on the basis of a scientific evaluation into score meaning. This approach was to have profound implications for all validation effort. Messick (1998, pp.70–71) seemed to imply that every kind of validation evidence is not only *relevant* but also *necessary* for every validation. Construct validation was to entail scientific theory-testing premised on multiple evidential sources. If the scope of modern validity theory was to be enlarged in an attempt to embrace a full evaluative treatment of consequences (as many, though not all, leading theorists of the day argued and continue to argue) then validation would require monumental effort especially if it was to include an exploration of unintended consequences.

The argument-based approach to validation – as championed by Kane (e.g., 1992, 2001, 2004, 2006, 2013), was an attempt to simplify both validity theory and validation practice. Recognising the difficulties in translating construct validity theory into construct validation practice, Kane rejects the idea that all kinds of evidence are required for every validation exercise (thereby running counter to the ethos of the construct validity thesis). He introduced the idea that test score interpretation is defined as an interpretive argument[1], which serves to identify assessment inferences and their sources of evidence. The interpretive argument provides a generic version of the proposed interpretation and use of scores, which can be applied to some population of interest. Kane asserts that the structure of the interpretive argument and the inferences and assumptions it necessarily entails, depend on the type of interpretation to be validated. Different interpretive arguments necessarily entail different patterns of inference. More ambitious theory-based interpretations require more evidence than less ambitious ones (Kane, 2009, 2013). Accordingly, certain kinds of evidence are irrelevant to validation relating to certain kinds of proposed interpretations and score uses.

Part of the persuasive power of the interpretive argument is the guidance it allegedly provides would-be practitioners. Although Kane's argument-based approach is widely regarded as a positive development, there have been few examples of its implementation. Even fewer examples of validity arguments for large-scale educational assessments are available to the research community (Goldstein & Behuniak, 2011). Where examples are published, they tend to lack a strong evaluative dimension (Haertel & Lorie, 2004; Kane, 2006), fall short of providing a compelling argument (Sireci, 2009, p.33), and fail to demonstrate how a test is constructed to represent a construct independent of test use (Sijtsma, 2010, p.782).

Summarising the period over the last thirty years, the modern construct validity 'consensus' appears to have engendered a legacy of unresolved tensions between those for whom the practice of validation is "a lengthy, even endless process" (Cronbach, 1989, p.151) and those with a responsibility for test development to provide sufficient, general validity evidence (of the instrumental value) attesting to the quality of their measurement procedures.

Notwithstanding the now, near universal acceptance of the modern unified conception of validity there remains a lack of coherence between theory and practice (e.g., Jonson & Plake, 1998; Hogan & Agnello, 2004; Cizek et al., 2008; Shaw, Crisp & Johnson 2012), or, as Messick put it, a "persistent disjunction between validity conception and validation practice" (Messick, 1988, p.34). Early in the twenty-first century the practice of validation still remains somewhat "impoverished" according to Brennan (2006, p.8) though there are pockets of good practice (e.g., Sireci, et al., 2006; Shaw & Weir, 2007; Chapelle, Enright & Jamieson, 2008; Khalifa & Weir, 2009; Sireci, 2012).

Kane's (2006) theorisation of the interpretive argument for traits

In Kane's (2006) seminal chapter in *Educational Measurement* he set out the interpretive argument implicit in trait interpretations. The core of his visualisation for the interpretation of traits is represented in Figure 1 (based on Kane, 2006, p.33, Figure 2.2). This illustrates the basic notion of making inferences from student performance through to the domain (and traits) of interest.

So for any assessment, the students conduct the tasks given which results in evidence of their performance on those specific tasks in that testing situation. Usually, the intended uses and interpretations of the results from an assessment (or from several assessments making up a qualification) mean that stakeholders need to make inferences about competence beyond those specific tasks. In other words, stakeholders need to know that this tells us something about how much of the relevant trait(s) each student has, both:

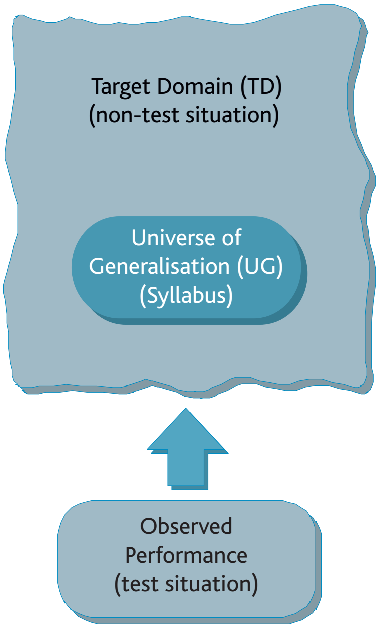
Target Domain (TD)
(non-test situation)
Universe of Generalisation (UG) (Syllabus)
Observed Performance (test situation)

Figure 1: Interpreting traits from performance (from Kane, 2006)

a. specifically in relation to the range of tasks that the assessment might reasonably have encompassed (based on the content and skills set out in the syllabus) and which scores are intended to represent – this is termed by Kane the 'Universe of Generalisation' (UG) and

b. more broadly to the domain of *any* possible tasks relating to the trait – this is termed the 'Target Domain' (TD) (of which the more limited Universe of Generalisation is a subset).

According to Kane (2006), "the Universe of Generalisation for the measure of a trait is often a small subset of the target domain and tends to be defined more precisely than the target domain" (p.34). The target domain can be thought of as the domain of interest in which the ability/abilities would be observed. The target domain goes beyond the scope of the testing situation to other tasks that could have been included in the assessments given the syllabus, and beyond to trait-relevant tasks in further study or employment contexts; in other words, a broader domain of non-assessment tasks and non-assessment contexts.

This underpinning notion of the interpretation of traits from performance and Kane's argument structure underpinned the validation framework development to be described in this article.

Proposing a validation strategy for large-scale, high-stakes international examinations

Following the development of an initial draft validation framework and set of methods for gathering validity evidence, the framework was piloted in 2008 with an International A level Geography qualification (Phase 1). This resulted in a number of revisions to the framework and proposed methods, involving streamlining the subset of methods used on the basis

[1] In 2013, Kane decided to abandon the label 'interpretive argument' in favour of interpretation/use arguments (IUAs) because the old formulation had given insufficient weight to uses. The new formulation also usefully allows a distinction to be made between interpretation and use arguments.

of how useful they were in providing evidence to evaluate validity and on the basis of their practicality.

In 2009 the framework (shown in Figure 2) was used to build a validity argument for an International A level Physics qualification (Phase 2). The framework provided the structure for collecting evidence to support the claim for the validity of the qualification, and to identify any potential threats to validity for this qualification such that they could be addressed. The structure of the validity argument was presented as an operationally-orientated validity portfolio which comprised details of the interpretive argument, validity evidence, and an evaluation of the validity argument.

The final phase of the developmental work attempted to ascertain how best to operationalise future validation effort. Through extensive consultation with colleagues and reflection on the experiences of the first two phases, Phase 3 aimed to provide suggestions for how to move forward with a strategy for validation of assessments.

A number of alternative validation strategies, from the stance and perspective of an international awarding body, were explored. These ranged in the degree to which they would provide sound evidence of the validity of assessments, and in the amount of resourcing that would be required. Whilst an attempt was made to develop streamlined and efficient methods, it was recognised that a robust evaluation of the validity of a qualification inevitably requires significant resource. The strategy adopted provided a practical and strategic approach to validity and validation where two approaches are undertaken in parallel:

1. an experimental strategy in which researchers conduct a full post-hoc validation of one or more syllabuses each year (or as necessary)

   plus

2. an operational strategy to be gradually introduced for all syllabuses, designed to gather and synthesise validation evidence currently generated routinely within an operational and assessment context.

Following implementation of the dual strategic approach to validation, a number of issues have emerged which have triggered not only a review of the validation methods but also the nature, scope and remit of the validation framework – in particular the questions addressed by the framework and the language employed in the framework.

Figure 2: Framework for the argument of assessment validation

Interpretive argument		*Validity argument*	*Evaluation*	
Inference	*Warrant justifying the inference*	*Validation questions*	*Evidence for validity*	*Threats to validity*
Construct representation	**Tasks elicit performances that represent the intended constructs**	1. Do the tasks elicit performances that reflect the intended constructs?		
Scoring	**Scores/grades reflect the quality of performances on the assessment tasks**	2. Are the scores/ grades dependable measures of the intended constructs?		
Generalisation	**Scores/grades reflect likely performance on all possible relevant tasks**	3. Do the tasks adequately sample the constructs that are set out as important within the syllabus?		
Extrapolation	**Scores/grades reflect likely wider performance in the domain**	4. Are the constructs sampled representative of competence in the wider subject domain?		
Decision-making	**Appropriate uses of scores/grades are clear**	5. Is guidance in place so that stakeholders know what scores/ grades mean and how the outcomes should be used?		

Evaluation of claim

	Evidence for validity	*Threats to validity*
How appropriate are the intended interpretations and uses of test scores?		
Interpretation 1. Scores/grades provide a measure of relevant learning/achievement		
Interpretation 2. Scores/grades provide an indication of likely future success		

Structure for the argument of assessment validation

The framework involves a list of inferences to be justified as indicated by a number of linked validation questions, each of which is to be answered by the collection of relevant evidence. The validation framework invites the collection of a considerable body of information in relation to categories of evidence presented in the fourth and fifth editions of the *Standards* (AERA, APA, & NCME, 2014); yet it ultimately adopts Kane's argument-based approach (e.g., Kane, 2006) in order to structure and judge that information.

Drawing on Kane's chain of inferences, the framework incorporates an underpinning logic for constructing an 'interpretive argument' (statements of claimed inferences from assessment outcomes, and the warrants which justify the inferences) based on a core structure common to all interpretive arguments within educational measurement, for the purpose of establishing measurement quality: performance inference (*Construct representation*), scoring inference, generalisation inference, extrapolation inference. In addition, a decision-making inference is included which enables a decision to be taken about test takers on the basis of their score. These inferences make up an interpretive chain which flows[2]:

from the task to the test performance (*Construct representation*)

↓

from the test performance to the test score/grade (*Scoring*)

↓

from test score/grade to test competence (*Generalisation*)

↓

from test competence to domain competence (Extrapolation)

↓

from domain competence to trait competence (*Decision-making*)

[2] See Tables 1 to 5 on pages 34–35 for definitions of terms.

For each inference, an associated warrant sets out a statement that is claimed to be true. The warrant, if appropriately supported by evidence through the validity argument, justifies the intended inference to which it relates.

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. The second table within the framework facilitates making conclusions about whether the intended interpretations of assessment outcomes (as set out in test claims) are appropriate given the evidence collected. For a full description of the development of the framework see Shaw, Crisp and Johnson (2012) and Shaw and Crisp (2012).

Framework revisions: issues and challenges

Implementation of the framework – designed to be used in the context of traditional written examinations (within general, academic qualifications) – revealed the emergence of a number of issues. The issues relate to the way in which the *Generalisation, Extrapolation* and *Decision-making* inferences are conceptualised and articulated. The *Construct representation*, and *Scoring* inferences remained unchanged in meaning and terminology as no issues had arisen in relation to these. Tables 1 and 2 set out the details of these two inferences for reference along with some brief explanation. The conceptual and linguistic revisions made to the framework in the remaining three inferences will then be described and are tabulated in Tables 3–5 (revisions are shown in red highlight in column 2).

Table 1: Construct representation inference

CONSTRUCT REPRESENTATION	
from Tasks to Test performance	
Test performance =	profile of performance on test tasks
Warrant:	Tasks elicit performances that represent the intended constructs
Validation question:	Do the tasks elicit performances that reflect the intended constructs?
Infer that:	The tasks elicit the intended test constructs.

Table 2: Scoring inference

SCORING	
from Test performance to Test score/grade	
Test score/grade =	mark total across all papers within syllabus (and related grade)
Warrant:	Scores/grades reflect the quality of performances on the assessment tasks
Validation question:	Are the scores/grades dependable measures of the intended constructs?
Infer that:	Test scores/grades represent intended constructs and quality of performance.

The Construct representation inference begins with the assessment tasks which (it is hoped) elicit performances representing the constructs of interest. Here, the validation question relates to whether the intended constructs are indeed reflected in the performances that are elicited. This is the first step in allowing stakeholders to make interpretations from performance (observed performance in the test situation) to the student's traits.

The Scoring inference relates to whether the scores or grades are a dependable measure of the intended constructs and reflect quality of performance in those constructs.

The Generalisation inference

The Generalisation inference advances the interpretive argument with a warrant that the test score/grade represents what would be obtained in the Universe of Generalisation (UG), that is, in all possible tasks that could fall within the scope of the syllabus. Generalisation depends on the "representativeness of the sample of observations and about the adequacy of the sample size for controlling sampling error." (Kane, 2006, p.34). If the test score/grade is an indication of expected performance over a domain of similar task performances all of which can be drawn from the content of the subject syllabus, then the syllabus itself constitutes the Universe of Generalisation. The syllabus is designed to reflect a view of the knowledge, understanding and skills that it is appropriate to develop in students at the level being assessed and is consistent with the current (or desired) curricular framework for the students for whom it is intended. Thus a claim relating to how well a test taker performs on a particular set of tasks on a particular occasion can be generalised to claims about expected performance on a larger domain of tasks drawn from the syllabus content (a universe of possible observations).

Table 3 shows details of the Generalisation inference in the validation framework before and after recent changes. In this inference, the student's overall competence in all tasks that could fall within the syllabus is inferred from the test score/grade. Changing from using the term *Test competence* in the original framework to *Syllabus competence* was intended to clarify the intended meaning of this term. The label 'Test competence' was considered too limiting in terms of the claims made about test taker performance and appeared to fail to convey its intended

Table 3: Generalisation inference – conceptual and linguistic changes

ORIGINAL	*REVISED*
from Test score/grade to Test competence	**from Test score/grade to Syllabus competence**
Test competence = overall competence in subject (all relevant subject tasks within scope of the syllabus)	**Syllabus competence** = overall competence in relation to all tasks that could be tested within the scope of the syllabus
Warrant: Scores/grades reflect likely performance on all possible relevant tasks	**Warrant**: Scores/grades reflect likely performance on all possible relevant tasks
Validation question: Do the tasks adequately sample the constructs that are set out as important within the syllabus?	**Validation question:** Do the tasks adequately sample the constructs that are set out as important within the syllabus?
Infer that: The scores on the tasks reflect scores on other tasks within the domain (expected scores).	**Infer that:** The scores/grades on the tasks reflect scores on other tasks within the syllabus.

meaning sufficiently clearly to evaluators new to the framework. For example, one evaluator expected Test competence to refer only to competence in the specific tasks in the specific assessment(s) which is not an unreasonable interpretation of the term. Thus, the term was adjusted to more clearly include the broader domain represented by the syllabus. The new term Syllabus competence was hoped to help with understanding, but does not represent a change to the underpinning meaning of the elements of the generalisation inference. In the revised framework, scores on test tasks reflect scores on other tasks within the syllabus (Table 3).

The Extrapolation inference

Extrapolation is central to educational and psychological assessment (Newton, 2013) and advances the interpretive argument further. The Extrapolation inference moves beyond reporting measures of observed performance in a relatively narrow domain to interpreting these more widely. The Extrapolation inference is an extrapolation to a broader domain of tasks (the Target Domain – TD) with the warrant that the universe score is what would be obtained in the TD and is used to predict future performance in some other, different context such as further study or employment.

In other words, extrapolation is an indication of likely wider performance beyond the local assessment context and suggests broader competence within and beyond the subject. The observed score can be interpreted as an indication of performance in the target setting (e.g., Higher Education Institution or workplace). Extrapolation translates performance in a local context (the test situation) into a prediction of performance in a future, non-test situation. How closely that future context relates to the knowledge and skills represented by the syllabus will affect how strong an indication of performance we can reasonably expect scores/grades to be. For example, a student's result for A level Physics is likely to be a stronger indicator of future performance on a Physics degree course, than on a Sociology degree course, and is likely to be a stronger indicator of likely future performance in a career as an Engineer than in a career as a Human Resources Consultant.

Domain competence in the original framework related only to overall competence in the subject, that is, it included competencies represented within the syllabus and going beyond this to wider competence in the subject area. However, having used this term in the framework for several years it became apparent that the meaning was not entirely understood. Validators using the framework for the first time were unsure whether it should be interpreted as the domain of the syllabus or the domain of the subject. Also, through implementation and reflection it was unclear whether this inference included just extrapolation to subject competence or extrapolation to the subject and beyond. As a result of extensive consultation and further review of the literature, it was decided that the extrapolation link should relate to making inferences from competence in the syllabus to competence in the wider subject and beyond, though it is expected, of course, that scores/grades would give a weaker indication of the latter than the former. Thus, the term *Broad competence* was chosen and the warrant, validation question, and explanation adjusted to reflect this. *Broad competence* widens the concept and relates to overall competence *within and beyond* the subject (Table 4). Accordingly, the validation question is broadened in the revised framework to include related competence beyond the subject. Enlarging the concept has implications for validation practice; because the scope of the interpretation is enhanced, "new kinds of evidence for support (e.g., criterion-related studies or analyses of the commonalities between assessment performance and performance in the wider domain)" (Kane, 2011, p.8) are required.

Table 4: Extrapolation inference – conceptual and linguistic changes

ORIGINAL	*REVISED*
from Test competence to Domain competence	**from Syllabus competence to Broad competence**
Domain competence = overall competence in subject	**Broad competence** = overall competence within and beyond the subject
Warrant: Scores/grades reflect likely wider performance in the domain	**Warrant:** Scores/grades give an indication of likely wider performance
Validation question: Are the constructs sampled representative of competence in the wider subject domain?	**Validation question:** Do the constructs sampled give an indication of broader competence within and beyond the subject?
Infer that: The scores in tasks within syllabus domain reflect wider competencies in the subject.	**Infer that:** The scores/grades in tasks the within the scope of the syllabus give an indication of wider competencies in the subject and beyond.

The Decision-making inference

The Decision-making inference advances the interpretive argument still further by allowing decisions to be made on the basis of test scores/grades by inferring that these give an indication of preparedness for further study/work. Reflecting on the concepts inherent in the original framework, it was thought that adjustment of the Decision-making inference was needed to accommodate the broadening of the Extrapolation inference to beyond the subject area. The shift of emphasis from guidance to aiding appropriate decision-making appears to be a positive step (Table 5).

The link back to decisions has also been made clearer in the validation question: Do scores/grades give an indication of success in further study or employment such that they can be used to make appropriate decisions? Appropriate decisions can only be made if the meaning of test scores is clearly interpretable by a raft of relevant stakeholders. Clear guidance to

Table 5: Decision-making inference – conceptual and linguistic changes

ORIGINAL	*REVISED*
from Domain competence to Trait competence	**from Broad competence to Preparedness for future study/work**
Trait competence = readiness for studying the subject (or another subject) at a higher level (e.g., university study), and aptitude for work in a related field	**Preparedness for future study/work** = preparedness for further study in the subject (or another subject), and aptitude for work
Warrant: Appropriate uses of scores/grades are clear	**Warrant:** Scores/grades give an indication of likely success in further study or employment
Validation question: Is guidance in place so that stakeholders know what scores/grades mean and how the outcomes should be used?	**Validation question:** Do scores/grades give an indication of success in further study or employment such that they can be used to make appropriate decisions?
Infer that: A student's likely future success in education and employment in relevant fields.	**Infer that:** The scores/grades on the tasks give an indication of a student's future success in education and employment and can be used to make appropriate decisions.

university admissions staff, for example, will facilitate admissions and placement decisions, thus exam board guidance on score/grade meaning would still be one source of evidence to be used in answering this validation question.

In the original version of the framework the term *Trait competence* was used in this inference to refer to readiness for further study and aptitude for work. On reflection, it was thought that the notion of 'readiness' or 'preparedness' was felt to be key and not well represented by the term 'Trait competence', hence the change to 'Preparedness for future study/work'. In the Decision-making inference the notion of competence and preparedness going beyond the specific area of study is continued from the previous inference (e.g., that a good grade in one subject can provide some level of indication of preparedness for study or work in related or less related fields).

The logical structure of an interpretive argument is valuable in the context of evaluating validity as awarding bodies are effectively making a *claim* that an assessment is valid, which needs to be backed by *evidence* (derived from theory, prior research or professional experience, or from evidence gleaned specifically as part of validation operations) via a *warrant* (justifying the inference), in order to defend the claim of validity against *rebuttals* (alternative explanations, or counter claims to the intended inference). (See Toulmin's Model of Inference, 1958/2003.)

Each inference depends on a number of assumptions which require different types of backing evidence relevant to the inference. Decision-making inferences generally rely on assumptions about the appropriateness of decisions made on the basis of test scores at the individual level. The evidence relevant to the Decision-making inference may include questionnaires to stakeholders devised in order to explore how Higher Education lecturers, undergraduate students and secondary school teachers understand and use test outcomes (e.g., scores/grades).

Revised interpretive chain

The full extent of the edits made to the original framework (specifically the validation questions and warrants) is shown in Figure 3. The revised interpretive chain now flows:

from the task to the test performance (*Construct representation*)

↓

from the test performance to the test score/grade (*Scoring*)

↓

from test score/grade to syllabus competence (*Generalisation*)

↓

from syllabus competence to broad competence (*Extrapolation*)

↓

from broad competence to preparedness for future study/work (*Decision-making*)

Figure 3: Revised framework for the argument of assessment validation

Interpretive argument		*Validity argument*	*Evaluation*	
Inference	*Warrant justifying the inference*	*Validation questions*	*Evidence for validity*	*Threats to validity*
Construct representation	**Tasks elicit performances that represent the intended constructs**	1. Do the tasks elicit performances that reflect the intended constructs?		
Scoring	**Scores/grades reflect the quality of performances on the assessment tasks**	2. Are the scores/ grades dependable measures of the intended constructs?		
Generalisation	**Scores/grades reflect likely performance on all possible relevant tasks**	3. Do the tasks adequately sample the constructs that are set out as important within the syllabus?		
Extrapolation	**Scores/grades give an indication of likely wider performance**	4. Do the constructs sampled give an indication of broader competence within and beyond the subject?		
Decision-making	**Scores/grades give an indication of likely success in further study or employment**	5. Do scores/grades give an indication of success in further study or employment such that they can be used to make appropriate decisions?		

Evaluation of claim

	Evidence for validity	*Threats to validity*
How appropriate are the intended interpretations and uses of test scores?		
Interpretation 1. Scores/grades provide a measure of relevant learning/achievement		
Interpretation 2. Scores/grades provide an indication of likely future success		

Concluding comments

This article has described a number of revisions to an established framework designed for evidencing assessment validity in large-scale, high-stakes international examinations. The original framework has recently been subject to a review resulting in a number of conceptual and textual changes. It is believed that the changes not only strengthen the theoretical structure underpinning the framework but also ensure that the framework is more transparent in terms of the clarity of its interpretive argument.

The structure for supporting validation was designed for traditional, awarding-based written examinations. These examinations can be characterised as a 'review and award' model (Section 3 of the *Cambridge Approach*, Cambridge Assessment, 2009). Other forms of established assessments (e.g., for vocational qualifications) and the emergence of other more innovative, technologically-driven forms of assessment such as twenty-first century skills (e.g., collaborative problem-solving, creativity and decision-making) and computer-based testing will only make the process of validation more complex. Indeed, the conceptual changes and textual edits described here actually make validation more of a challenge for the validator. Nevertheless, the challenge of validation – no matter how great, should not impede its continuing execution.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA, APA, & NCME). (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Brennan, R.L. (2006). Perspectives on the evolution and future of educational measurement. In R.L. Brennan (Ed). *Educational Measurement* (4th edition) (pp.3–16). Washington, DC: American Council on Education/Praeger.

Cambridge Assessment. (2009). *The Cambridge Approach: Principles for designing, administering and evaluating assessment*. Cambridge: A Cambridge Assessment Publication.

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

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

Crisp, V. & Shaw, S. (2012). Applying methods to evaluate construct validity in the context of A level assessment. *Educational Studies*, *38*(2), 209–222.

Cronbach, L.J. (1984). *Essentials of psychological testing*. New York: Harper & Row.

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

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

Goldstein, J. & Behuniak, P. (2011). Assumptions in alternate assessment: An argument-based approach to validation. *Assessment for Effective Intervention*, *36*(3), 179–191.

Haertel, E.H. & Lorie, W.A. (2004). Validating standards based test score interpretations. *Measurement: Interdisciplinary Research and Perspectives*, *2*, 61–104.

Hogan, T.P. & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement*, *64*(4), 802–812.

Jonson, J.L. & Plake, B.S. (1998). A historical comparison of validity standards and validity practices. *Educational and Psychological Measurement*, *58*(5), 736–753.

Kane, M.T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535.

Kane, M.T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342.

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

Kane, M.T. (2006). Validation. In R.L. Brennan (Ed). *Educational Measurement* (4th edition) (pp.17–64). Washington, DC: American Council on Education/Praeger.

Kane, M.T. (2009). Validating the interpretations and uses of test scores. In R.W. Lissitz (Ed.). *The Concept of Validity: Revisions, new directions, and applications* (pp.39–64). Charlotte, NC: Information Age Publishing, Inc.

Kane, M.T. (2011). Validating score interpretations and uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing*, *29*(1), 3–17.

Kane, M.T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Khalifa, H. & Weir, C.J. (2009). *Examining Reading: Research and practice in assessing second language reading*. Cambridge: Cambridge University Press.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*, 635–694.

Messick, S. (1988). The once and future issues of validity: assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp.33–45). Hillsdale, NJ: Lawrence Erlbaum Associates.

Messick, S. (1989). Validity. In R. Linn (Ed.). *Educational Measurement* (3rd edition) (pp.13–100). Washington, DC: American Council on Education.

Messick, S. (1998). Alternative modes of assessment, uniform standards of validity. In M. Hakel (Ed.). Beyond multiple choice: *Evaluating alternatives to traditional testing for selection* (pp.59–74). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Newton, P.E. (2013). Two kinds of argument. *Journal of Educational Measurement*, *50*(1), 105–109.

Shaw, S. & 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*, 1–44.

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

Shaw, S. & Weir, C.J. (2007). *Examining Writing: Research and Practice in Assessing Second Language Writing*. Cambridge: Cambridge University Press.

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

Sijtsma, K. (2010). Book review. In R.W. Lissitz (Ed.). (2009) The concept of validity: Revisions, new directions, and applications. Charlotte, NC: Information Age Publishing, Inc. *Psychometrika*, *75*(4), 780–782.

Sireci, S.G. (2009). Packing and unpacking sources of validity evidence: history repeats itself again. In R.W. Lissitz (Ed.). *The concept of validity: revisions, new directions, and applications*, (pp.19–37). Charlotte, NC: Information Age Publishing, Inc.

Sireci, S.G. (2012). De-"Constructing" Test Validation. *Center for Educational Assessment Research Report No. 814*. Amherst, MA: Center for Educational Assessment. Paper presented at the annual conference of the National Council on Measurement in Education, Vancouver, April 2012.

Sireci, S.G., Baldwin, P., Martone, A., Zenisky, A.L., Hambleton, R.K., & Han, K. (2006). *Massachusetts Adult Proficiency Tests Technical Manual*. Amherst, MA: Center for Educational Assessment. Retrieved from: http://www.umass.edu/remp/CEA_TechMan.html.

Toulmin, S. (1958/2003). *The uses of argument*. Cambridge: Cambridge University Press.