Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 21 Number 12, December 2016 ISSN 1531-7714

Macro- and Micro-Validation: Beyond the 'Five Sources' Framework for Classifying Validation Evidence and Analysis

Paul E. Newton, *Ofqual*

This paper argues that the dominant framework for conceptualizing validation evidence and analysis – the 'five sources' framework from the 1999 *Standards* – is seriously limited. Its limitation raises a significant barrier to understanding the nature of comprehensive validation, and this presents a significant threat to effective validation practice. Motivated by a belief that 'validity by design' ought to be substantiated through 'validation of design' this paper demonstrates the importance of adopting a broader conceptual framework. It introduces a new framework, based upon the metaphor of different validation lenses through which to scrutinize assessment procedures at differing levels of detail, with micro-validation lenses at one end of a continuum and macro-validation lenses at the other. The evolution of validation theory can be seen as a very gradual, if somewhat reluctant, acknowledgement of the importance of micro-validation. This paper recommends micro-validation as the natural foundation for any comprehensive validation program.

The 'Five Sources' Framework

This paper argues that the dominant framework for conceptualizing validation evidence and analysis – the 'five sources' framework from the 1999 *Standards for Educational and Psychological Testing* (hereafter, *Standards*) (AERA et al., 1999) – is seriously limited. Its limitation raises a significant barrier to understanding the nature of comprehensive validation, and this presents a significant threat to effective validation practice. The less standardized the assessment procedure in question, the greater the threat presented. Motivated by a belief that 'validity by design' ought to be substantiated through 'validation of design' this paper aims to demonstrate the importance of adopting a broader conceptual framework for conceptualizing validation evidence and analysis.

Evolution of the *Standards* validation framework

As North American scholars have dominated the field of validation theory, and as the *Standards* is a consensus statement of the North American measurement professions (AERA et al., 2014) and plays a key role in assessment communities worldwide (Zumbo, 2014), it seems reasonable to conclude that the 'five sources' presented in its validity chapter constitutes the dominant framework for conceptualizing validation evidence and analysis:

1. evidence based on test content
2. evidence based on response processes
3. evidence based on internal structure
4. evidence based on relations to other variables
5. evidence for validity and consequences of testing.

The evolution of this framework took place over a period that spanned half a century, having begun life in the first edition of the *Standards* (APA et al., 1954) as 'four types' of validity. Although the 'four types'

framework was modified only slightly for the second and third editions of the *Standards*, it fell into disrepute during the 1970s and 1980s because of the misleading impression that it gave to practitioners. It was taken by many to imply that if, for instance, you needed to validate an educational achievement test, then you only needed to demonstrate content validity, and you were able to do so by undertaking a single content validation study. In other words, it seemed to imply that different kinds of validation evidence and analysis were relevant to different kinds of test (or, more specifically, to different kinds of test use), and that results from a single study were sufficient to claim validity.

In the wake of seminal work by Samuel Messick and others (e.g. Guion, 1974; 1980; Messick, 1975; 1980; 1989), a new 'unified' view of validity evolved. This held that validation ought to be understood as a scientific program of research: that all sorts of evidence and analysis should be considered relevant, whatever the test or test use; and that evidence or analysis from a single study could never be considered sufficient. Since the new view was essentially an extension of the logic that already underpinned construct validation, it spawned the maxim: all validity is construct validity and all validation is construct validation. The fourth edition of the *Standards* (AERA et al., 1985) reflected a partial conversion to the unified view of validity, by reconstructing its validation framework in terms of sources of 'evidence' rather than 'types' of validity. The fifth edition (AERA et al., 1999) completed this conversion, with a new framework based upon the five sources presented above.

Impact of the *Standards* validation framework

Directly after describing its five sources, the current edition of the *Standards* explains that:

"A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses." (AERA et al., 2014, p.21)

In other words, it suggests that validation involves gathering the kind of evidence and analysis represented by its five categories, and then using those sources to construct an argument for (or conceivably against) the overarching validity claim. Messick (1989) characterized this process as integrating as much evidence and analysis as possible, from as many sources as possible, to ensure that the overall argument is as strong as possible.

Sireci (2013) has recommended using the five sources framework as a formal template for planning validation research; as has Zumbo (see Zumbo and Chan, 2014a). Following Messick, Sireci (2016) explained that all five sources are relevant to test score interpretation and use; although he acknowledged debate over the relevance of evidence from consequences to test score interpretation.

Using the five sources framework as a common reference point, contributors to Zumbo and Chan (2014b) surveyed trends in validation practices across the social, behavioral and health sciences, through a systematic search of reports – that explicitly presented themselves as validation studies – published since the 1960s. Their project concluded that evidence from both response processes and consequences was largely ignored across disciplines, despite their privileged position within the *Standards* since 1999 (Lyons-Thomas et al., 2014). Similar conclusions have been reached by Cizek et al. (2008), Cizek et al. (2010), Cook et al. (2014), Padilla and Benitez (2014).

Finally, the five sources framework not only influences the kind of evidence that is seen as relevant to validity, it also influences the kind of evidence that is not, for instance:

"Face validity is not included as a source of validity evidence by contemporary validity theorists and the *Standards* […] Considering that the *Standards* were published in 1999, it is still surprising to observe researchers report face validity as a source of validity evidence." (Ark et al., 2014, p.282)

These observations on the five sources framework illustrate its normative and prescriptive role in planning and structuring validation. They illustrate how it is used as a reference point for what ought to be included within a program of validation research and what ought not to be included. Interestingly, they also reveal clear disjunctions between validation theory and validation practice, as certain of the five sources are often overlooked and as other sources beyond the five are often included. The causes of this disjunction have been speculated upon. Some have suggested that there may be insufficient knowledge of validation frameworks amongst practitioners (e.g. Cook et al., 2014). Others have argued that validity theory itself is either too confusing (e.g. Shepard, 1997) or just plain wrong (e.g. Cizek, 2012). Cizek, for instance, argued that evidence from social consequences is largely irrelevant to the

judgement of validity, and should therefore not be privileged as one of the five sources (Cizek, 2016). If true, this might help to explain the lack of evidence from social consequences in published validation studies; although it does not explain the lack of evidence from response processes. The argument developed below is different; but it does agree that the five sources framework is problematic, and that this presents barriers to practitioner understanding and consequent threats to validation practice[1].

Indications of inadequacy

Interestingly, if not paradoxically, an implicit acknowledgement that the five sources framework is inadequate appears as a caveat in the validity chapter itself:

"Ultimately, the validity of an intended interpretation of test scores relies on all the available evidence relevant to the technical quality of a testing system. Different components of validity evidence are described in subsequent chapters of the *Standards*, and include evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all test takers, as appropriate to the test interpretation in question." (AERA et al., 2014, p.22)

In other words, having carefully elucidated the five sources of evidence and analysis, the chapter then almost casually proposes that a variety of additional sources need also to be investigated. Note, for instance, how 'appropriate test administration' represents a very different kind of evidence from that represented within the five sources framework. A recent validation report by the Smarter Balanced Assessment Consortium (Smarter Balanced, 2015, Chapter 2) incorporated the additional sources from the above quotation[2] within a secondary framework of nine 'essential elements' (albeit acknowledging overlap between this secondary framework and the primary five sources one):

1. careful test construction
2. adequate measurement precision (reliability)
3. appropriate test administration
4. appropriate scoring
5. accurate scaling and equating
6. appropriate standard setting
7. attention to fairness, equitable participation and access
8. validating 'on-track/readiness'
9. adequate test security.

The Smarter Balanced report described 'appropriate test administration' like this:

> "*Review of test administration procedures, including protocols for test irregularities; use of and appropriate assignment of test accommodations."* (Table 1, p.5)

Indeed, a variety of process-related sources of evidence and analysis were incorporated within the same table, to describe various of the nine elements, including: review of scoring procedures; documentation of test design; review of accommodation policies; analysis of data integrity policies; and so on.

These additional sources raise a particularly important question concerning the significance of assessment processes for validation: if assessment processes are significant, then how should they be

[1] It is important to situate the argument developed by the present paper – which is a critique of the five sources framework – within what has become known as 'the great validity debate' (see Crocker, 1997, and Newton and Baird, 2016). In 1965, Samuel Messick presaged what would become a longstanding debate amongst measurement professionals, over the role of social consequences in validity theory, when he drew a distinction between two major questions that arise in evaluating the appropriateness of a particular test administration: (i) the essentially scientific question (of technical quality) – is the test any good as a measure of the characteristic it purports to assess? and (ii) the ultimately ethical question (of societal value) – should the test be used for its present purpose? Nowadays, validity scholars can be classified in terms of the extent to which their preferred definition of validity is narrow and purely oriented towards technical quality (e.g. Borsboom et al., 2004) rather than broad and ultimately oriented towards societal value (e.g. Moss, 2016). The critique of the five sources framework does not depend upon the adoption of a particular definition of validity. However, for the sake of expositional clarity, it will assume a definition of validity that is fairly broad yet technically-oriented. In other words, it will restrict validation to the essentially scientific question of the degree to which it is possible to measure (whatever it is that needs to be measured in order to support specified purposes). Even when validity and validation are restricted in this manner, the five sources framework is still seriously limited. In other words, even ignoring the debate over the relevance of social consequences to validity, the present paper argues for a broader perspective on validation evidence and analysis. If a broader and more ethically-oriented definition of validity were to be adopted, then its validation framework would need to be correspondingly broader; in particular, a far wider range of impacts would need to be embraced. Readers may find it helpful to consult Newton and Shaw (2014) for an overview of the history of validity theory which covers the evolution of the five sources framework as well as the great validity debate.

[2] The report actually quoted the 1999 edition of the *Standards*, but the content of the quotation was essentially the same.

scrutinized, and how should their appropriateness and adequacy be established? For instance,

- does validation require evidence that certain key processes have been established?
- is evidence and analysis required to demonstrate, in principle, the appropriateness and adequacy of those processes?
- is evidence required that those processes are actually implemented, during each assessment cycle?
- is evidence and analysis required to demonstrate that those processes are implemented in the right way (i.e. to the specified operational standard) during each assessment cycle?

Questions like these begin to imagine a far broader perspective upon validation and cast doubt upon the idea that validation evidence and analysis can neatly be circumscribed using the five sources framework.

Although it is fairly obvious that assessment processes underpin assessment quality, scholars seem only recently to have discussed process scrutiny as a significant component of validation. For instance, during the 1990s, Kane (1994) introduced the idea of 'procedural validity' for standard setting; Downing and Haladyna (1997) described 'validity evidence from quality assurance procedures'; and Sireci (1998) identified 'appropriate test construction procedures' as an aspect of content validity. Similar ideas were discussed in the Downing-Haladyna *Handbook of Test Development*; particularly within section V on *Test Production and Administration* (e.g. Campion and Miller, 2006; McCallin, 2006). In relation to language assessments, process scrutiny constituted a prominent source of backing for the Assessment Use Arguments described by Bachman and Palmer (2010). Most recently of all, Cizek (2016) has suggested a revision of the 'five sources' framework to include a new category that is labelled 'evidence based on test development and administration procedures'.

If we assume that process scrutiny can (somehow) contribute evidence and analysis of importance to validation, the challenge then becomes one of how best to characterize and organize this evidence and analysis in a manner that is conceptually clear, comprehensive enough to do justice to its potential variety, and accessible to practitioners. The *Standards* has effectively ducked this challenge, by retaining the five sources framework which (by its own admission) excludes lots of important evidence and analysis.

Validation of Design

So, can process scrutiny contribute evidence and analysis of importance to validation? And, if so, then how?

Validity by design and validation of design

In an article entitled *Validity by design*, Mislevy (2007) traced the origins of Evidence-Centered Design to his frustration that validity theory provided an inadequate basis for developing new forms of assessment that would support valid inferences. Evidence-Centered Design was therefore proposed as an approach that could help assessment creators to practice less like craftspeople and more like engineers, by making the theory of assessment design explicit and by explaining how and why alternative design decisions might enhance or reduce validity.

"ECD, however, makes the factors that influence test design explicit and links the myriad decisions made during task creation, test assembly, and scoring into a chain of evidence-based reasoning that better supports an argument for the validity of the inferences made about test takers on the basis of their scores." (Zieky, 2014, p.85)

Consequently, when practicing validity by design, the process of creating assessments is structured in such a way that validity evidence and analysis emerges naturally.

Evidence-Centered Design can trace its ancestry to the 'rational' approach to test development (Flanagan, 1951; Travers, 1951) and to the proposition that "content validity […] is both a process and a goal" (Huddlestone, 1956, p.293). Yet, its logic actually extends way beyond content analysis to embrace each and every feature or process that is designed into an assessment procedure. This implies that there ought to be an identifiable rationale for the design of each one of those features and processes; and that this rationale ought to include its contribution to the validity of the assessment procedure overall.

Evidence-Centered Design is essentially just a systematic approach to building validity into assessment procedures. Indeed, validity by design ought to be a fundamental aspiration for any assessment creator. Or, to put it another way, 'validity by chance' would seem to

be an inappropriate aspiration. If 'validity by design' is the claim, then 'validation of design' provides an essential component of its justification. If validity emerges from the myriad decisions made during assessment design – both explicit and implicit – then it stands to reason that the 'design logic' that underlies each feature or process within an assessment procedure is a proper subject for validation research. Indeed, systematic scrutiny along these lines presents itself as the natural foundation for any comprehensive validation program. The validation of design principle is, of course, reflected within the five sources framework: content analysis provides the classic example. However, although content analysis is clearly very important, it is not uniquely important, such that it deserves its own category to the exclusion of other process-related sources. In fact, the framework excludes all sorts of evidence and analysis that might legitimately contribute to validation of design.

The centrality of the assessment procedure

Fulcher (2015) proposed that what we refer to as a 'test' is really the set of specifications from which any form of a test is generated, that make explicit the features that must not change from one form to the next. To extend this proposition: what we refer to as an 'assessment' is really the set of specifications that govern the entire activity of measuring, that make explicit the features and processes that must not change from one assessment cycle to the next. This is embodied in the idea of an assessment procedure, which is the (general) procedure through which (particular) measurements are generated, that is, the mechanism through which assessment results are delivered during each cycle.

The assessment procedure comprises all of the features and processes that are controlled, or standardized, from one assessment cycle to the next. Although assessments vary widely in the kind of features and processes that are standardized, procedures for large-scale educational assessments typically specify things like:

- the nature of the proficiency that needs to be measured
- the process for developing tasks to elicit evidence of the proficiency
- the process for administering those tasks
- the process for evaluating evidence of proficiency from task performances
- the process for transforming performance evaluations into measurement results
- the manner in which those results should (and should not) be interpreted.

The goal of assessment design is to create a valid assessment procedure; a procedure that can be relied upon to deliver accurate and useful measurement results. From this perspective, to claim that an assessment procedure has a high level of validity is to claim that the particular assessment results that it generates may be treated as though they were accurate and useful because there is a strong argument for doing so (i.e. the overall validation argument, constructed on the basis of validation evidence and analysis).

Validation lenses

It is certainly not the case that the five categories (which comprise the five sources framework) are either unhelpful or unimportant, it is simply that they do not collectively exhaust the validation space; by a long way, in fact. So, is it possible to situate the contents of these categories within a broader conceptual framework, in order to do justice to the potential variety of validation evidence and analysis, and to process-related sources in particular? One way of characterizing the nature and scope of comprehensive validation invokes the metaphor of alternative lenses through which to scrutinize assessment procedures. This suggests a fundamental distinction between micro-validation and macro-validation. Macro-validation is akin to the customer's perspective on assessment: does the assessment procedure work in the way that it ought to work? Micro-validation is akin to the engineer's perspective on assessment: is the assessment procedure built in the way that it ought to be built? The critical point is that these two perspectives represent different kinds of inquiry. Macro-validation research tends to investigate outcome-related, or product-related questions; whereas micro-validation research tends to investigate input-related, or process-related questions.

Micro-validation employs a lens that is narrow and therefore highlights detail. It focuses on the features and processes that comprise the assessment procedure, both in isolation and in interaction. It seeks 'low-level' evidence and analysis concerning the nature and operation of those features and processes and asks whether they appear to have been effectively designed. This lens embodies the idea of validation of design, which naturally complements validity by design. Micro-

validation involves scrutinizing the assessment procedure directly: judging each of its features and processes in terms of their underlying design logic and empirical evidence concerning design efficacy.

Evidence of design efficacy might be gathered in various ways, for instance:

1. routine formative analyses (e.g. item facility indices, DIF analyses, item-test correlations, linking studies, fairness reviews)
2. quality control metrics (e.g. marker-moderator consistency statistics, printing error statistics)
3. auxiliary investigations (e.g. expert judgements of item-objective congruence, 'think aloud' studies with candidates/markers/others).

Having described validity evidence required in peer reviews for compliance with NCLB requirements for state assessment systems, Schafer, Wang and Wang (2009) concluded their chapter on 'validity in action' by stressing the importance of process evidence. They identified four steps that need to be documented for each process within an assessment procedure: process, product, evaluation, and improvement. This resonates strongly with the idea of micro-validation. They described their third step in terms of evaluating the product of each process; although, from a micro-validation perspective, this could equally be framed as an evaluation of the process itself. For instance, a test form construction process might be evaluated both in terms of its design logic (e.g. the rationale underlying its approach to sampling) and in terms of its design efficacy (e.g. by asking a group of experts to judge one of its products, a particular test form, in terms of item-objective congruence).

Macro-validation employs a lens that is wide but lacks detail. It focuses on the assessment procedure overall. It seeks 'high-level' evidence and analysis derived from sources external to the assessment procedure – primarily measurement outcomes and systemic impacts – and asks whether this evidence and analysis is consistent with the overarching claim that it is possible to measure what needs to be measured.

The analysis of measurement outcomes includes classic sources of evidence related to:

1. external relations – based on overall results (e.g. test-criterion correlations, test-indicator correlations, multi-trait multi-method correlations, theory-based predictions)
2. internal structure – based on subtask scores (e.g. reliability statistics, factor analyses, component correlations).[3]

The analysis of systemic impacts includes more recently recognized sources, including evidence related to:

1. consequences and side-effects (e.g. progression routes, unexpected subgroup rejection rates)
2. misuse (e.g. when this indicates what really needs to be measured)
3. customer satisfaction (e.g. uptake/sales figures, general feedback)
4. public opinions (e.g. public confidence surveys).

The lens metaphor suggests that, for any particular assessment procedure, validation can be, and should be, undertaken both holistically (macro) and atomistically (micro). The distinction is specifically intended to foreground the importance of atomistic validation, validation of design, as a natural foundation for any comprehensive validation program.

The micro-macro continuum

It is possible to think of the distinction between micro- and macro-validation more in terms of a 'fuzzy' continuum than a binary division. This helps to foreground the prototypical sources within each category, whilst acknowledging that there might be an element of debate over how best to classify certain other sources. Indeed, the critical issue, here, is not so much the nature of the source, per se, but the use to which it is put. For instance, when individual item scores are correlated with the aggregate of all item scores, the intention is to evaluate particular items and, by extension, to evaluate an aspect of the item development process. So this would be a micro-level analysis. Whereas, when individual item scores are correlated with each other via Cronbach's alpha, the intention is to evaluate the overall assessment procedure. The alpha coefficient provides a (partial) thumbs-up or thumbs-

[3] From this perspective, reliability is best conceptualized as just one category of validation evidence and analysis alongside many others.

down in relation to the procedure overall. So this would be a macro-level analysis.

Figure 1 helps to illustrate the idea of a fuzzy continuum. Acknowledging that there is plenty of room for debate, the specific locations of individual boxes should not be over-interpreted; but notice how the sources towards the left of the continuum have been associated with particular features or processes, whereas the sources towards the right have not. This emphasizes that those sources towards the left concern narrow, targeted evaluations of underpinning validity claims; whereas those towards the right concern a broad, holistic evaluation of the (single) overarching validity claim. The purpose of a framework (or, perhaps, meta-framework) like this is not to restrict, by implying that the sources of evidence and analysis that appear in the boxes are the only legitimate ones. Instead, the purpose is to expand, by implying that all sorts of sources of evidence and analysis can be considered legitimate, including the five sources from the *Standards*. So the sources in the boxes are merely illustrative.

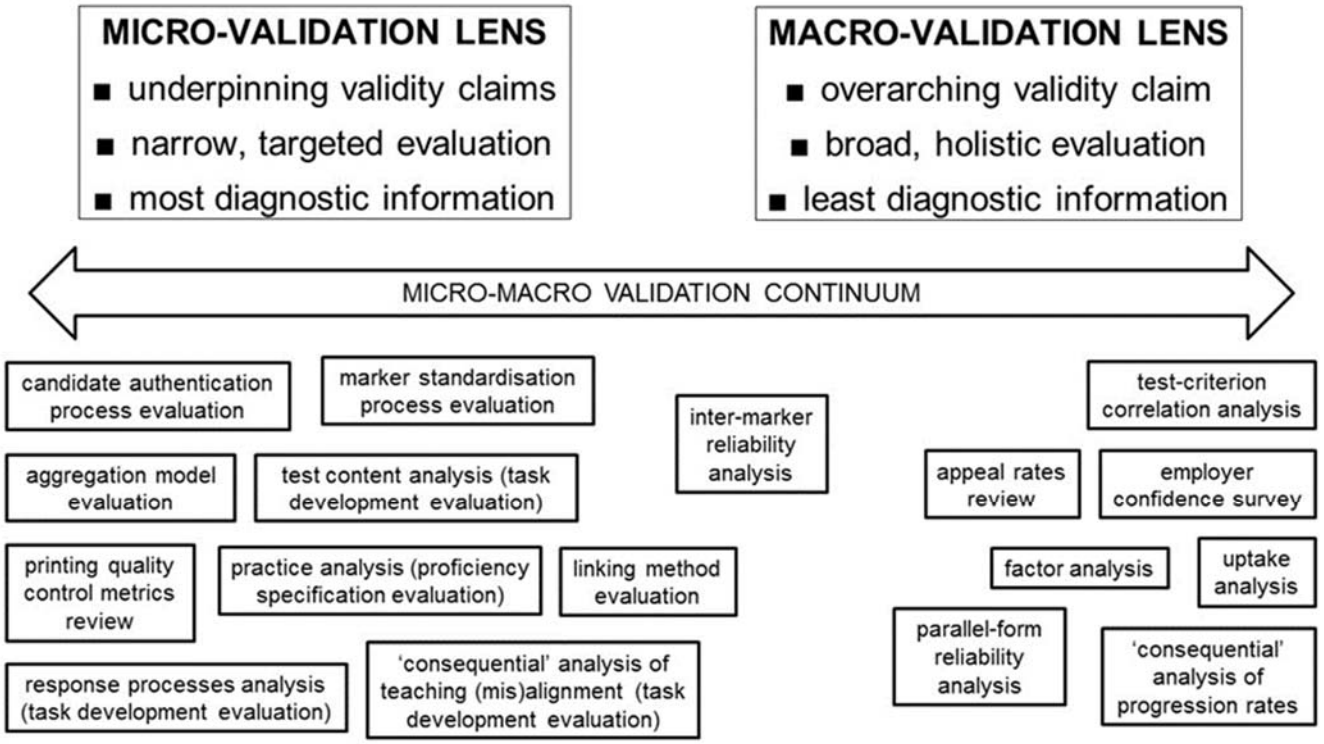

Figure 1. The Micro-Macro Validation Continuum

Notice how the first two of the five sources from the *Standards* – content and response process analysis – have been located towards the micro-validation end of the continuum. Response process analysis is the most prototypical because it is so very narrowly focused upon a specific kind of link in the overall validation argument chain, e.g. whether the cognitive processes that candidates actually engage, when answering questions of a certain kind, are the ones that they are presumed to engage. Although it investigates this kind of link in great detail, and often provides important formative insights concerning the efficacy of question types, it tends not to be very powerful, from a summative evaluation perspective, in relation to the overarching validity claim. This is because the links that it targets are relatively small in relation to the entire chain; and, as such, even favorable outcomes contribute only a small amount of information to the overall summative evaluation of an assessment procedure. Similarly, even when outcomes identify significant problems with the design of certain kinds of question, this may still make only a small impact on the overarching validity claim, if the assessment procedure specifies only a small number of questions of that kind. Having said that, unfavorable micro-validation outcomes can sometimes be very powerful, even from an overall summative evaluation perspective; for example, if it were established that an inappropriate aggregation model had been specified.

Examples that would fall under the third and fourth of the five categories – internal structure and external relations – have mainly been located towards the macro-validation end of the continuum. Test-criterion correlation analysis is the most prototypical because of its potential to support a powerful overall summative evaluation conclusion, at least in theory. This is because the analysis targets the assessment procedure overall, rather than specific features and processes, which means that its findings have the potential to contribute powerful information. Indeed, in theory, a near perfect correlation from a near perfect concurrent validation study might even be considered sufficient to claim a very high level of validity. It seems likely that this is why so much emphasis was placed upon criterion validation during the early years of the measurement movement, circa 1920s to 1940s (see Newton and Shaw, 2014). In a similar way, both favorable and unfavorable outcomes from a parallel forms reliability analysis have the potential to contribute powerful information; although, even a near perfect parallel forms reliability analysis would be far less complete, as an evaluation of the overarching validity claim, than a near perfect concurrent validation study. Of course, the idea of a 'near perfect' macro-validation study is something of a pipe-dream, particularly as far as test-criterion correlation analysis is concerned (e.g. Toops, 1944; Jenkins, 1946; Thorndike, 1949). This is why macro-validation studies – as potentially cost-effective as they might seem – cannot be relied upon exclusively when developing an overall validation argument. In addition, unfavorable outcomes from macro-validation studies provide no diagnostic information at all concerning possible causes of invalidity, so they are not useful from the perspective of re-designing the assessment procedure.

Notice how Figure 1 includes two 'consequential' analysis boxes; one towards the macro-validation end and one towards the micro-validation end. Progression routes provide evidence concerning the consequences of assessment results for learners. Evidence of widespread lack of progression into work or further learning might raise serious questions concerning whether an assessment was really measuring what it was presumed, or what it actually needed, to be measuring. Towards the other end of the continuum, teaching practices provide evidence concerning the consequences of assessment practices for learners. Evidence that a large number of school science teachers were failing to teach certain elements of the science curriculum might raise serious questions concerning whether the science examination was predictably restricted in its approach to sampling. Both of these examples illustrate how evidence from consequences can bear upon judgements of validity, even when its definition is restricted to the technically-oriented question of measurement quality.

Again, note how Figure 1 contains numerous sources of evidence and analysis that lie well beyond the five sources framework. Uptake analysis, a form of macro-validation, can raise questions concerning whether the assessment is really measuring what it is presumed, or what it actually needs, to be measuring. So too can evidence from employer confidence surveys (Cedefop, 2015). Aggregation model analysis, a form of micro-validation, can raise questions concerning the appropriateness of combining performance evaluations according to a compensatory, conjunctive or disjunctive principle. Even evidence from result appeal statistics can be – and in countries like Denmark and Austria actually is – seen as a meaningful indicator of the quality of assessment procedures (Cedefop, 2015).

The myth of incontrovertibility

The above discussion throws into relief the longstanding myth that validity evidence and analysis ought to be as incontrovertible as possible in order to qualify as a 'legitimate' or 'true' source (e.g. Downing, 2006). For example, the idea of 'face validity' has been criticized for decades on the basis that even expert judges frequently draw incorrect inferences concerning validity from scrutiny of assessment tasks alone (e.g. Guilford, 1946). Of course, *if* you believed that you only needed to demonstrate a single type of validity to demonstrate validity, and *if* you believed that validity could be demonstrated using a single study, then you *would* need that evidence or analysis to be as watertight as possible! But that way of thinking about validation is a relic from the past. The unified view of validity has recast validation as an ongoing program of scientific research, based on all sorts of evidence and analysis. Inevitably, certain sources of evidence and analysis will be weaker than others, for a host of reasons. But that does not mean that the weaker sources are either illegitimate or not useful. The critical issue is the overall integration of evidence and analysis which can straightforwardly accommodate issues of differential strength.

From this perspective, even judgements made by novice test takers provide a legitimate and useful source of validation evidence; it may not be the strongest evidence, and it might even be contradicted by other sources, but it is legitimate evidence all the same. After all, it is quite possible to imagine test takers exerting insufficient effort on an educational achievement test which they believed, from inspection alone, to be assessing the wrong learning outcomes. And that would clearly constitute a validity threat. Similarly, social media uproar over allegedly 'unanswerable' test questions can be very helpful in identifying validity threats that might otherwise have been overlooked; even when test takers' perceptions are not entirely accurate. The same kind of reasoning can be applied to other sources of evidence and analysis mentioned above, e.g. public confidence surveys or assessment uptake figures. High uptake figures provide no guarantee of validity, obviously, but they do constitute a weak indicator. Similarly, low uptake figures prompt important validity questions, such as whether the assessment is actually measuring what its users need it to be measuring.

The Justification for a Broader Conceptual Framework

The justification for a broader conceptual framework can be argued in various ways. The most obvious argument is that the five sources framework fosters an impoverished view of validation evidence and analysis, which thereby risks practitioners designing sub-optimal validation research programs. Two other powerful arguments should be considered. First, the evolution of validation theory can be seen as a gradual rejection of a macro-validation mind-set, the logical conclusion of which is to absorb the contents of the five sources framework within a far broader one. Second, practitioners who are responsible for less standardized assessment procedures are very clearly under-resourced by the five sources framework.

Evolutionary significance

It is interesting to note that many of the earliest conceptions of technical quality were articulated exclusively at the macro level, for instance:

"Reliability has been regarded as the correlation of a given test with a parallel form. Correspondingly, the validity of a test is the correlation of the test with some criterion." (Gulliksen, 1950, p.88)

This particular formulation defined technical quality purely in terms of (relationships between) test outcomes, with no reference at all to features of the test itself. Correlation with some criterion reflected the 'empirical' approach to validation. Between the 1920s and 1940s, it was the dominant approach, and some considered it to be the only legitimate approach.

It took some time before micro-level concerns began to be recognized more widely and explicitly as fundamental to evaluating technical quality. Perhaps the most significant transition was the recognition of content validity – reflecting the 'logical' approach to validation – alongside concurrent and predictive validity in the first edition of the *Standards* (APA et al., 1954). However, it is unclear whether to interpret this first consensus statement as indicative of widespread support for a broader framework than either the dominant empirical approach or the logical approach. Indeed, many still believed in the supremacy of their own preferred approach. Guilford, for instance, preferred the empirical approach, comparing the logical approach to crystal ball gazing (Guilford, 1946). Ebel, on the other hand, promoted the logical approach, explaining that the credibility of the empirical approach was entirely dependent upon prior application of the logical approach (Ebel, 1956) which, in effect, rendered the empirical approach superfluous (Ebel, 1983). To some extent, the recognition of content validity in the first edition of the *Standards* might be seen as a concession to those who insisted that an educational achievement test that effectively represented its domain was 'obviously valid' and was therefore its own best criterion (Rulon, 1946). No doubt, the presentation of the first validity framework in terms of distinct types provided some justification for scholars and practitioners to continue promoting their own preferred type (and downplaying or ignoring other types).

It was the new, unified view of validity that really began to open the way for genuinely broader validation frameworks: the three sources framework in the 1985 *Standards*; and the five sources framework in the 1999 *Standards*. Ironically, though, it seems that Samuel Messick – champion of the new, unified view of validity and the new, expansive view of validation – may well have been responsible for imposing artificially restricted boundaries upon the concept of validation evidence and analysis, with his claim that "there are only a half dozen or so distinct sorts" of validity evidence (see Messick, 1989, p.16).

Five of the six 'aspects' that Messick identified map fairly directly onto the five sources framework; the exception being his generalizability aspect. Messick described his aspects as "general validity criteria or standards for all educational and psychological measurement" (Messick, 1995, p.744) and believed them to be important in helping practitioners to appreciate the significance of aspects that might otherwise be downplayed or overlooked, e.g. social consequences. In other words, on the one hand, he considered his own 'six aspects' framework to be importantly broader than previous frameworks, most obviously the 'three sources' framework from the 1985 *Standards*. Yet, on the other hand, the boundaries that he imposed have proved, with the passage of time, to be unduly narrow.

Fortunately, the importance of process scrutiny for validation is (gradually) beginning to be recognized. Indeed, as noted earlier, Cizek has recently proposed a new source of evidence and analysis from 'test development and administration procedures' (Cizek, 2016, p.220). Yet, although this is a step in the right direction, it is only a small step, because it excludes all sorts of evidence and analysis related to features and processes beyond test development and administration. More importantly, it is not clear how this new category is conceptually distinct from the first two sources in his revised framework – test content and response processes. In short, the problem of how to recognize the importance of process scrutiny to validation cannot be solved simply by adding an additional category to the five sources framework. Instead, the very idea that validation evidence and analysis can neatly be circumscribed by a handful of categories is in question. What is required is a far broader framework.

Practical utility

As noted earlier, the five sources framework was developed in North America, in a context that has traditionally been dominated by standardized tests constructed from multiple low-tariff items. The importance of paying due regard to micro-validation becomes even more apparent when considering assessment procedures that are far less standardized than this, which is true of the majority of qualifications in England, for instance.

Qualifications, and educational assessments more generally, come in all sorts of shapes and sizes. Certain kinds of qualification are based exclusively upon an 'external' assessment model. For instance, a qualification might be awarded on the basis of performance on a single 40-item multiple choice test, for which all candidates sit the same test each session. Almost everything is standardized under this model, with the possible exception of the particular items that feature in the test from session to session. This means that it is quite straightforward to generate many of the traditional examples of validation evidence and analysis, as derived from the five sources framework; including Cronbach's alpha, DIF statistics, factor analyses, item-objective congruence studies, 'think aloud' studies, and so on.

Other kinds of qualification are based exclusively upon an 'internal' assessment model. Organizations that award these qualifications often devolve most, if not all, of the responsibility for critical assessment processes – materials development, performance elicitation, performance evaluation, and so on – to assessors working within schools, colleges or workplaces (known as assessment 'centers'). These assessors are often encouraged to use a variety of assessment approaches, which means that even within the same center during the same session no two candidates will necessarily be assessed in exactly the same way. In other words, for qualifications like these, many critical assessment processes are not standardized at all and are therefore not part of the overall assessment procedure that is specified by the awarding organization. For qualifications like these, it is impossible to generate the traditional sources of validation evidence mentioned above (which presume item-level analysis). Furthermore, although it might be possible to devise experiments to generate certain forms of macro-validation evidence (e.g. certain reliability coefficients, or certain outcome-criterion relationships), this might well prove to be very challenging, both technically and practically.

At this point, a validation practitioner influenced by the five sources framework might begin to give up hope of constructing a passable program of validation research. In contrast, a practitioner influenced by a broader framework might begin to search further afield for plausible evidence and analysis. Indeed, for qualifications that devolve almost all of the responsibility for critical assessment processes to individual assessors, validity is heavily dependent upon the effectiveness of higher order features or processes that are designed into the assessment procedure to ensure that all assessors have sufficient expertise, integrity and understanding of the qualification standard. Critical validation evidence, here, might include assessor credentials, documentation

of assessment strategy approval mechanisms, training and exemplification materials, documentation of moderation processes, moderation quality control metrics, and so on.

Conclusion

In recent years, the importance of process scrutiny to validation has increasingly been recognized. However, it has remained very far from clear how to accommodate evidence and analysis of this sort within the five sources framework. The present paper argues that the five sources framework is incapable of accommodating it, and that a broader framework is required. A different way of thinking about validation evidence and analysis is made possible by distinguishing between different kinds of inquiry: macro-validation research tends to investigate outcome-related, or product-related questions (akin to the customer's perspective); whereas micro-validation research tends to investigate input-related, or process-related questions (akin to the engineer's perspective).

Micro-validation is not a concession, that is, a fall-back option when macro-validation seems unduly challenging. Neither is it an added-extra, intended to bolster macro-validation. Instead, macro-validation and micro-validation are two sides of the same coin, providing complementary perspectives within a comprehensive validation program. Having said that, the less macro-validation evidence and analysis is available, the more micro-validation evidence and analysis will have to shoulder the burden. Moreover, since micro-validation is 'validation of design' this means that its evidence and analysis will arise, in part, as a natural by-product of designing and developing assessments. And this means that a large body of micro-validation evidence and analysis should be available long before macro-validation begins. As such, it can properly be understood as the natural foundation for any comprehensive validation program.

Finally, notice how macro-validation attempts to demonstrate *that* it is possible to measure, but does not attempt to demonstrate *how* or *why*. The longstanding dominance of a macro-validation mind-set helps to explain the lack of systematic attention in the validity literature to "the steps in the causal process that start with the attribute intended to measure and end with the measurement outcome" (see Bringmann and Eronen, 2016, p.34). Recognizing and emphasizing micro-validation should help to overcome this tendency, by encouraging measurement professionals and organizations to open up the logic of assessment design to the level of conceptual and empirical scrutiny that it properly deserves. This is to recommend the kind of shift in validation practice that Zumbo (e.g. 2007a; 2007b; 2009) has proposed, on the basis of his characterization of validity as contextualized and pragmatic *explanation*[4].

References

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

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

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

American Psychological Association, American Educational Research Association, and National Council on Measurements Used in Education. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, 51 (2), 1–38.

Ark, T.K., Ark, N. and Zumbo, B.D. (2014). Validation practices of the Objective Structured Clinical Examination (OSCE). In B.D. Zumbo and E.K.H. Chan (Eds.). *Validity and Validation in Social, Behavioral, and Health Sciences* (pp. 267–288). Switzerland: Springer International Publishing.

Bachman, L.F. and Palmer, A.S. (2010). *Language Assessment in Practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Borsboom, D., Mellenbergh, G.J. and van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111 (4), 1061–1071.

Bringmann, L.F. and Eronen, M.I. (2016). Heating up the measurement debate: What psychologists can learn from the history of physics. *Theory & Psychology*, 26 (1), 27–43.

[4] Zumbo (2007a) noted that descriptive analyses have traditionally dominated validation practice – including macro-validation techniques such as factor analysis, test-criterion correlation analysis, and multi-trait multi-method correlation analysis – at the expense of genuinely explanatory analyses.

Campion, D. and Miller, S. (2006). Test production effects on validity. In S.M. Downing and T.M. Haladyna (Eds.). *Handbook of Test Development* (pp.599–623). Mahwah, NJ: Lawrence Erlbaum Associates.

Cedefop (2015). Ensuring the Quality of Certification in Vocational Education and Training. Luxembourg: Publications Office. Cedefop research paper; No 51. http://dx.doi.org/10.2801/25991

Cizek, G.J. (2012). Defining and distinguishing validity: interpretations of score meaning and justification of test use. *Psychological Methods*, 17 (1), 31–43.

Cizek, G.J. (2016). Validating test score meaning and defending test score use: different aims, different methods. *Assessment in Education: Principles, Policy & Practice*, 23 (2), 212–225.

Cizek, G.J., Bowen, D. and Church, K. (2010). Sources of validity evidence for educational and psychological tests: A follow-up study. *Educational and Psychological Measurement*, 70 (5), 732–743.

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

Cook, D.A., Zendejas, B., Hamstra, S.J., Hatala, R. and Brydges, R. (2014). What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Advances in Health Sciences Education*, 19 (2), 233–250.

Crocker, L. (1997). Editorial: The great validity debate. *Educational Measurement: Issues and Practice*, 16 (2), 4–4.

Downing, S.M. (2006). Face validity of assessments: faith-based interpretations or evidence-based science? Medical Education, 40 (1), 7–8.Downing, S.M. and Haladyna, T.M. (1997): Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education*, 10 (1), 61–82.

Ebel, R.L. (1956). Obtaining and reporting evidence on content validity. *Educational and Psychological Measurement*, 16 (3), 269–282.

Ebel, R.L. (1983). The practical validation of tests of ability. *Educational Measurement: Issues and Practice*, 2 (2), 7–10.

Flanagan, J.C. (1951). The use of comprehensive rationales in test development. *Educational and Psychological Measurement*, 11 (1), 151–155.

Fulcher, G. (2015). Re-examining Language Testing: A philosophical and social inquiry. Oxford: Routledge.

Guilford, J.P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, 6 (4), 427–439.

Guion, R.M. (1974). Open a new window: Validities and values in psychological measurement. *American Psychologist*, 29 (5), 287–296.

Guion, R.M. (1980). On Trinitarian doctrines of validity. *Professional Psychology*, 11 (3), 385–398.

Gulliksen, H. (1950). Theory of Mental Tests. New York: John Wiley and Sons, Inc.

Huddlestone, E.M. (1956). Test development on the basis of content validity. *Educational and Psychological Measurement,* 16 (3), 283–293.

Jenkins J.G. (1946). Validity for what? *Journal of Consulting Psychology*, 10 (2), 93–98.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64 (3), 425-461.

Lyons-Thomas, J., Liu, Y. and Zumbo, B.D. (2014). Validation practices in the social, behavioral, and health sciences: A synthesis of syntheses. In B.D. Zumbo and E.K.H. Chan (Eds.). *Validity and Validation in Social, Behavioral, and Health Sciences.* (pp.313–319). Switzerland: Springer International Publishing.

Marion, S.F. and Pellegrino, J.W. (2006). A validity framework for evaluating the technical quality of alternate assessments. *Educational Measurement: Issues and Practice*, 25 (4), 47–57.

McCallin, R.C. (2006). Test administration. In S.M. Downing and T.M. Haladyna (Eds.). *Handbook of Test Development* (pp.625–652). Mahwah, NJ: Lawrence Erlbaum Associates.

Messick, S. (1965). Personality measurement and the ethics of assessment. *American Psychologist*, 20 (2), 136–142.

Messick, S. (1975). The standard problem: meaning and values in measurement and evaluation. *American Psychologist*, 30 (10), 955–966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, 35 (11), 1012–1027.

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

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50 (9), 741–749.

Mislevy, R.J. (2007). Validity by design. *Educational Researcher*, 36 (8), 463–469.

Moss, P.A. (2016). Shifting the focus of validity for test use. *Assessment in Education: Principles, Policy & Practice*, 23 (2), 236–251.

Newton, P.E. and Baird, J. (2016). The great validity debate. *Assessment in Education: Principles, Policy & Practice*, 23 (2), 173–177.

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

Padilla, J. and Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, 26 (1), 136–144.

Rulon, P.J. (1946). On the validity of educational tests. *Harvard Educational Review*, 16, 290–296.

Schafer, W.D., Wang, J and Wang, V. (2009). Validity in action: State assessment validity evidence for compliance with NCLB. In R.W. Lissitz (Ed.). *The Concept of Validity: Revisions, New Directions, and Applications* (pp.173–193). USA: Information Age Publishing.

Shepard, L.A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16 (2), 5–8.

Sireci, S.G. (1998). The construct of content validity. *Social Indicators Research,* 45, 83-117.

Sireci, S.G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement*, 50 (1), 99–104.

Sireci, S.G. (2016). On the validity of useless tests. *Assessment in Education: Principles, Policy & Practice*, 23 (2), 226–235.

Smarter Balanced. (2015). Smarter Balanced Technical Report. (http://www.smarterbalanced.org/technical-report/ , accessed January 2016).

Thorndike, R.L. (1949). *Personnel Selection: Test and Measurement Techniques*. New York: John Wiley & Sons, Inc.

Toops, H.A. (1944). The criterion. *Educational and Psychological Measurement*, 4 (1), 271–297.

Travers, R.M.W. (1951). Rational hypotheses in the construction of tests. *Educational and Psychological Measurement*, 11 (1), 128–137.

Zieky, M.J. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa*, 20 (2), 79–87.

Zumbo, B.D. (2007a). Validity: Foundational issues and statistical methodology. In C.R. Rao and S. Sinharay (Eds.). *Handbook of Statistics, Volume 26: Psychometrics* (pp.45–79). Amsterdam: Elsevier Science B.V.

Zumbo, B.D. (2007b). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4 (2), 223–233.

Zumbo, B.D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R.W. Lissitz (Ed.). *The Concept of Validity: Revisions, New Directions, and Applications* (pp.65–82). USA: Information Age Publishing.

Zumbo, B.D. (2014). What role does, and should, the test Standards play outside of the United States of America? *Educational Measurement: Issues and Practice*, 33 (4), 31–33.

Zumbo, B.D. and Chan, E.K.H. (2014a). Reflections on validation practices in the social, behavioral, and health sciences. In B.D. Zumbo and E.K.H. Chan (Eds.). *Validity and Validation in Social, Behavioral, and Health Sciences* (pp.321–327). Switzerland: Springer International Publishing.

Zumbo, B.D. and Chan, E.K.H. (2014b) (Eds.). *Validity and Validation in Social, Behavioral, and Health Sciences.* Switzerland: Springer International Publishing.

Citation:

Newton, Paul, E. (2016). Macro- and Micro-Validation: Beyond the 'Five Sources' Framework for Classifying Validation Evidence and Analysis. *Practical Assessment, Research & Evaluation*, 21(12). Available online: http://pareonline.net/getvn.asp?v=21&n=12

Corresponding Author

Paul E. Newton
Research Chair
Office of Qualifications and Examinations Regulation (Ofqual)
Spring Place, Herald Avenue
Coventry, CV5 6UB
England

email: Paul.Newton [at] ofqual.gov.uk