Author/s:
O'Leary, TM;Hattie, JAC;Griffin, P

Title:
Actual Interpretations and Use of Scores as Aspects of Validity

Date:
2017-06-01

Citation:
O'Leary, T. M., Hattie, J. A. C. & Griffin, P. (2017). Actual Interpretations and Use of Scores as Aspects of Validity. Educational Measurement Issues and Practice, 36 (2), pp.16-23. https://doi.org/10.1111/emip.12141.

Persistent Link:
https://hdl.handle.net/11343/292623

Actual Interpretations and Use of Scores as Aspects of Validity

Timothy Mark O'Leary

John Hattie

Patrick Griffin

Abstract

Validity is the most fundamental consideration in test development. Understandably, much time, effort, and money is spent in its pursuit. Central to the modern conception of validity are the interpretations made, and uses planned, on the basis of test scores. There is, unfortunately, however, evidence that test users have difficulty understanding scores as intended. That is, whilst the proposed interpretations and use of a test scores might be theoretically valid they might never come to be because the meaning of the message is lost in translation. This necessitates pause. It is almost absurd to think that the intended interpretations and uses of test scores might fail because there is a lack of alignment with the actual interpretations made and uses enacted by the audience. Despite this, there has only recently been contributions to the literature regarding the interpretability of score reports, the mechanisms by which scores are communicated to their audience, and their relevance to validity. These contributions have focused upon linking, through evidence, the *intended* interpretation and use with the *actual* interpretations being made and actions being planned by score users. This paper reviews the current conception of validity, validation, and, validity evidence with the goal of positioning the emerging notion of validity of usage within the current paradigm.

Keywords: validity, validation, validity evidence, score reports, interpretability of score reports.

The contemporary definition of validity places central importance on both interpretations and use of scores (American Educational Research Association [AERA], American Psychological Association [APA] & National Council on Measurement in Education [NCME], 2014). Despite such clarity of definition, there is some concern that this importance has not shifted effectively beyond pure theoretical discussion into practice (Cizek, 2012; Haertel, 2013; MacIver, Robinson, Costa & Evers, 2014). To date, much of the discourse pertaining to validity and validation has been focused on theoretical interpretations and use of scores, particularly on those uses that were (or are) intended by test developers and designers. Unfortunately, given that there is evidence that score users can have difficulty interpreting scores as intended (Goodman & Hambleton, 2004; Hambleton & Slater, 1997; Jaeger, 1998), what is intended sometimes must be contrasted with the actual interpretations made and use planned once the outcomes of a test are reported to the target audience.

The differences between *intended* interpretations and use of scores and *actual* interpretation and use of scores are somewhat subtle and might seem trivial. They are, however, of the utmost importance. When there is an alignment between intended and actual interpretations and use, then the purpose of tests, the intended personal and social consequences at the core of assessment practice, have the greatest chance of being realised (Hubley & Zumbo, 2011). Consider a hypothetical diagnostic educational test as an example. The intended purpose of such a test might be to provide evidence to a teacher about their students in order to inform future instruction. The intended interpretations for such an assessment would be focused on the learning needs of students and the intended use centred on planning appropriate interventions. When a teacher receives the output of such a test and is able to correctly identify the needs of their students and makes plans to meet those needs, then there is a clear alignment between intended and actual interpretations and use. If, though, the output was misinterpreted and the teacher did not correctly identify the needs of some (or all) of their students and planned interventions which failed to meet the needs of these students, then the interpretations and use would not be in alignment with what was intended. In such circumstances, the intended impact of the test might be considered corrupt. Furthermore, depending on the circumstances, the very validity of the intended interpretation and use, no matter how well supported in theory, may well be questionable.

What should already be clear is that alignment between intended interpretations and use of scores and actual interpretations and use of scores is critical. This brings us to the focus of this paper, a discussion about validity, validation, and output from tests. The output from tests are the visible conclusion of the complex process of testing. These outputs, more often in the form of reports (and not simply numbers or scores) are fundamental in the process of communication between test developers and their audience. As such, evidence

of the effectiveness of the interpretations made by score users based upon these outputs, score reports, and their ensuing use are of the utmost significance. Arguably, such evidence is fundamental in any claims about validity. Unfortunately, current validity theory and validation practice do not incorporate any explicit references or guidance about how to deal with the actual (as opposed to the intended) interpretations made and use planned as a consequence of score users' engagement with score reports. Furthermore, there is evidence that many test score users have difficulty understanding the ways in which scores are reported as intended (Goodman & Hambleton, 2004; Hambleton & Slater, 1997; Jaeger, 1998). In concert, these two points are of concern and are shortcomings of both theory and practice which, regardless of how good any technical forms of validity evidence might be, threaten the validity of any and all intended interpretations and use. Nevertheless, it is only very recently that explicit notions of the validity of interpretation (MacIver et al., 2014) and of interpretability (of score reports) as aspects of validity (Van der Kleij, Egen & Engelen, 2014) have been articulated within the literature. These contributions represent valuable steps forward, but continued work is required in this area.

It is worth pausing to clarify some necessary terms and our position before proceeding. Firstly, validity is about both interpretations and use of scores. However, in addition to the known and anticipated interpretations and use of scores, there are many unknown interpretations and use comprising off-label, unintended/or illegitimate use and users of test scores (Zumbo, 2015). Within the current concept of validity, there is, however, a distinction between the unexpected side effects of legitimate test use and test misuse and/or illegitimate test use. Importantly, test misuse and illegitimate test use should not be considered a focus for validity and validation efforts (Messick, 1998). This is logical, as there is no doubt that it would be unreasonable to expect the measurement community to anticipate and address every imaginable use and misuse as a part of validation (Ho, 2013). Secondly, there is evidence from the existent literature that test score users have difficulty understanding the ways in which scores are reported as intended (Goodman & Hambleton, 2004; Hambleton & Slater, 1997; Jaeger, 1998). Therefore, our position is that, while it might not be possible to address and/or prevent the misuse or illegitimate use of scores, it should be possible to improve intended score users' interpretation and use of scores by enhancing the comprehensibility of score reports for their intended audience. Arguably, this would improve the validity of the intended interpretation and use.

The purpose of this article is to: (1) review and summarise the existent literature on validity, validation, and validity evidence; (2) identify limitations in the existing concept of validity evidence; and (3) build on the concept of user validity proposed by MacIver et al. (2014) and the idea of interpretability score reports as an aspect of validity discussed by Van der Kleij et al. (2014) in order to better place both within the existing validity frame.

Validity and validation: The current state of play

Validity

The concept of validity has come a long way since its initial codification in the Technical Recommendations for Psychological Tests and Diagnostic Techniques (APA, 1954). At that time, validity was considered to relate to information or evidence that indicated the degree to which a test was "capable of achieving certain aims" (APA, 1954, p. 213) and there were three distinct types of validity: (1) content; (2) predictive / concurrent criterion related; and, (3) construct. Under the technical recommendations a test was considered valid if it was able to achieve its stated aims. Furthermore, the process of validation involved the provision of evidence supporting the stated aims (Messick, 1989b). Significant work undertaken by Cronbach and Meehl (1955), Guion (1977), Tenopyr (1977), Cronbach (1980, 1988), and Messick (1989a, 1989b) led to the three types of validity being unified under the banner of construct validity. Messick's (1989b) seminal chapter 'Validity' in Educational Measurement (Linn, 1989) was instrumental in re-framing the concept of validity. As a consequence thereof, the 1999 edition of the Standards for Educational and Psychological Testing ('the Standards') (AERA, APA & NCME, 1999, p.9) defined validity as "the degree to which evidence and theory support the interpretations of test scores" and positioned validity as the "most fundamental consideration in test development and design". This concept has remained unchanged in the most recent edition of the Standards (AERA, APA & NCME, 2014). Essentially, at its very core, validity is about the interpretations and use that are based on test scores as opposed to the actual testing instrument itself (Hubley & Zumbo, 2011) and, of equal importance, it must be evaluated with respect to "the purpose of the test and how the test is used" (Sireci, 2009, p. 20).

Validation

Previously, validation was seen as concerning the provision of evidence supporting the stated aims (Messick, 1989b). Currently, however, validation is conceptualised as the process of gathering relevant and appropriate evidence in order to provide a "sound scientific basis for the proposed interpretation" (AERA et al., 2014, p. 11) with a sound validity argument integrating "various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretations of test scores for specific uses" (AERA et al., 2014, p. 21). Further, validation is considered to be a never ending process and the validity of inferences made may change over time depending on the best available evidence (AERA et al., 2014).

The Standards (AERA et al., 2014) posit that validation begins with an explicit understanding and statement of the proposed interpretations and use of test scores in concert with a rationale for the relevance of the interpretation to the proposed use (p. 11). The argument-based approach to validation, which has evolved as a pragmatic consequence of the significant works of Cronbach (1988), House (1980), Kane (1992, 2006, 2013), and

Sheppard (1993), provides a framework for such evaluation, which "reflects the general principles of validity without requiring formal theories" (Kane, 2013, p. 9). This approach to validation, recently explained by Kane (2013), makes use of two types of arguments: a validity argument and an interpretation/uses argument (IUA).

The first step in the argument-based approach to validation is focused on building the IUA for the assessment in order to provide a framework for the validity argument. The IUA specifies the relationship between observed performances and the interpretations that one wants to support. As Kane (2013) identified, these relationships can be expressed as a series of "if-then" statements aimed at clarifying a number of assumptions underlying each proposed interpretation.

The second step is the creation of the validity argument. The validity argument should evaluate the plausibility of each of these assumptions by integrating appropriate evidence collected from a variety of sources which are relevant to each interpretation.

At its core, validation concerns the provision of sound interpretation/use and validity arguments and has the purpose of integrating and synthesising various strands of validity evidence into a coherent account of the degree to which existing theory and evidence support the intended interpretations of test scores for specific uses. The implication of this definition is that validation is not simply about one piece of evidence being used in isolation. Rather, it is the evaluation of a web or network of evidence connecting inferences and interpretations with use, in its entirety, in order to evaluate the plausibility of claims based on the scores.

Evidence of validity

If the process of validation involves gathering relevant and appropriate evidence, then it is important to consider what evidence is relevant and appropriate. While historically there were different types of validity the current conception of validity incorporates types of validity evidence. This shift from forms of validity to forms of evidence is fundamental to unified concept of validity. Importantly, Messick (2000) and others (Hubley and Zumbo, 1996; Zumbo, 2007, 2009) have contended that validity cannot and should not rely on any of these forms in isolation. Rather, they should be used with discernment and in concert as appropriate.

Within the current concept of validity, there are five broad types of evidence which can be interwoven in the evaluation of the validity of proposed interpretations and use. The five types of evidence (outlined by the Standards [AERA et al., 2014]) include:

Content: Can be obtained by an analysis of relationships between the content of a test (themes and the wording and format of items, tasks, or questions) and the construct it is intended to measure.

Response process: Relates to the response processes of test takers and might be necessary to support the fit between the construct and nature of the response or the response actually engaged by the test taker. This is particularly relevant when assumptions are made about test takers cognitive processes.

Internal structure: Analyses of the internal structure of a test can indicate the degree to which the relationship between test items conforms to the construct of interest. This is particularly relevant when the theoretical framework for a test indicates a particular dimensionality.

Relationships with other variables: Often, the intended use of a test implies a relationship with another variable and, as such, evidence relating to analyses of possible relationships can provide an important source of validity evidence in such instances.

Consequences: This is perhaps the most controversial of the types of evidence and has been the focus of many strong opinions (Cizek, 2012; Hubley & Zumbo, 2011; Lissitz & Samuelsen, 2007; Sheppard, 1997). The types of consequences that occur because of test use are broadly categorised as: (1) those that were intended by the test developer and follow directly from the interpretation of test scores; (2) those that extend beyond the interpretation or uses of scores intended by test developers; and (3) other unintended and often negative consequences. In each of these instances, evidence related to the appropriateness of consequences is required at the time of the creation of the test for intended consequences and in an ongoing evaluative manner for unintended consequences.

The evidence of my discontent

The Standards position validity as "the most fundamental consideration in developing and evaluating tests" (AERA et al., 2014, p. 11). Specifically, as Haladyna (2006) commented, "the most important concern for any test is the validity of its score interpretations or uses". Furthermore, validity should be evaluated in terms of "the purpose of the test and how the test is used" (Sireci, 2009, p. 20). As stated previously, however, there is evidence from the existent literature that test users have difficulty understanding the ways in which scores are reported as intended (Goodman & Hambleton, 2004; Hambleton & Slater, 1997; Jaeger, 1998). This suggests that something fundamental is amiss. Surely evidence of *actual* interpretations and use should also be included as a part of the process? Unfortunately, this does not appear to be the case which is disappointing. This is particularly so given the recent recognition of the "importance of providing evidence for how users make inferences and take actions", and the consequence in practice being "whether the test developer (and report developer) can provide evidence for the adequacy and appropriateness of these interpretations" (Hattie & Leeson, 2013, p. 595).

Essentially, what this means is that, currently, most validation is focused on technical evidence in support of inferences, interpretations, actions, and uses as theorised by test

developers as opposed to actual interpretations and use by test score users. This presents a problem and there is a pressing need for an updated conceptualisation of validity or, at the very least, validity evidence which is more inclusive of actual test user interpretations and use (Bennett, 2010; Hattie, 2009; MacIver et al., 2014).

Summary

There is a maturing concept of validity and the process by which validation takes place. There is, however, much opportunity for refinement. The Standards position validity as "the degree to which evidence and theory support the interpretations of test scores" (AERA et al., 2014, p. 11). Currently, validation is concerned with providing theory and evidence in support of intended or proposed interpretations and use. However, the importance of providing evidence for how users make inferences and take actions has recently been recognised (Hattie & Leeson, 2013). Nevertheless, within the Standards, there is no clearly articulated form of validity evidence or guidelines related to a consideration of linking how test score users make actual interpretations and subsequently plan uses based on scores. This presents a challenge.

The need for evidence regarding how test score users make actual interpretations and subsequently plan uses based on scores cannot be overstated. A lack of a consideration for such evidence means that, while test interpretations and use might theoretically and technically be considered fit for purpose, there is currently no form of evidence being considered in the context of validity theory and validation efforts which captures whether score users are actually interpreting the output from the tests in the manner that was intended. In more simple terms, despite much movement in validity theory, validity, in practice, is dominated by whether a test is capable of achieving its stated aims. This is disappointing. If validity is to be truly concerned with the appropriateness of interpretations and use, then evidence of the quality, appropriateness, and effectiveness of the actual interpretations that test score users make and the actions they plan based on how scores are reported must be central to both the validity and validation processes. Not only would this result in a more authentic realisation of the current definition, but consideration of such evidence could help to improve the overall quality of the outcomes of testing by (1) helping to identify poor interpretations and uses, un-anticipated interpretations and uses, and misuse before the fact and, (2) subsequently, informing necessary improvement with regards to how score are being reported.

Emerging focus on the interpretability of score reports

Recent contributions to the field by Van der Kleij et al. (2014) and MacIver et al. (2014) have emphasised the importance of user interpretation of score reports. MacIver et al.'s proposed concept of user validity (although we are not convinced that a new term is helpful) captures the "overall accuracy and effectiveness of interpretation resulting from test output" focusing on "the validity of the interpretations in use and the decisions that

form part of these interpretations" (2014, p. 155). Instead of focusing on *intended* interpretations and actions, MacIver at al.'s (2014) concept of user validity focuses on *actual* interpretations in use. Essentially, the base contention of MacIver et al. (2014) is that validity should "not be focused on the test scores, but on the validity of the test interpretations" (although the latter is dependent on the former). Furthermore, Van der Kleij et al. (2014) have suggested the inclusion of the interpretability of score reports as an aspect of validity.

While superficially different, both of these contributions shine a light on the importance of actual interpretations and the use of test scores. This is captured well by Van der Kleij et al. (2014, p. 25): "a correct interpretation of test results is a necessary precondition for adequate use" and "a correct interpretation of reports is especially relevant when the test results are meant to inform important or irreversible decisions". Essentially, the point these contributors are making is that validity, validation, and validity evidence must include more than technical and theoretical evidence with regards to proposed interpretations and use. What must also be included is some evidence of the interpretability of score reports. That is evidence of how well members of the intended audience are actually interpreting (and using) scores as reported. Perhaps another approach to this notion of interpretability is consideration of the level of alignment between *intended* interpretations and use and *actual* interpretations and use.

The significance of each of these contributions cannot be overstated. They provide both a synthesised view of traditional validity and score reporting intertwined with a consideration of the actual interpretations made and use planned by score users. This, while perhaps broader than previously considered, is entirely congruent with Messick's (1989b) re-definition of validity as being concerned with interpretations and use. This is, no doubt, a welcome step for those who have continued to argue that score reporting as a field is crucial in tests and testing and is in dire need of further inquiry.

User validity and interpretability of score reports *within* the unified concept of validity

bMoving forward, oth MacIver et al. (2014) and Van der Kleij et al.'s (2014) contributions require adjustment to fit more comfortably within the current paradigm. To more appropriately place both contributions within the current concept of validity, it is necessary to a make singular concession: If validity is concerned with interpretations and use and score reports are the medium by which the outcomes of tests are communicated to their users, then evidence of the actual interpretability of score reports is essential in any judgement with regards to validity. Assuming such a concession has merit and is possible, an adjustment to both contributions is proposed. Messick's (1989b) seminal chapter unified the three distinct types of traditional validity and brought them under the banner of the unitary notion of construct validity. Thus, user validity should not be considered as a distinct form of validity nor should we think of the validity of reports. More appropriately, they should both be incorporated and we should consider formalizing Van der Kleij et al.'s (2014)

notion of interpretability of score reports as an aspect of validity. This could be achieved quite simply through the introduction of a new form of validity evidence focused on the effectiveness of interpretations and use resulting from score users' engagement with score reports.

Integrating actual score interpretations as an aspect of validity evidence

The changing notion of validity and the evidence required to support it have been focused on the improvement of and increased sophistication of the technical aspects of validity and its supporting evidence. These changes have brought about improvements to validity and validation in practice. However, despite the Standards having only recently been republished in 2014, there is a need to consider the next set of amendments thereto. One of the most significant and glaring omissions is the lack of a necessity to incorporate evidence of the interpretability of scores (Hattie, 2014). Such evidence should not only be included in the next edition of the set of Standards, rather it must be considered central to the notion of validity and validation.

Figure 1 visually positions validity, validation, and validity evidence in relation to each other. The current concept of validity is regarded as "the degree to which evidence and theory support the interpretations of test scores" (AERA et al., 2014, p. 11) and validation is the actual process of gathering the relevant and appropriate evidence in order to provide a "sound scientific basis for the proposed interpretation" (AERA et al., 2014, p. 11) with a sound validity argument integrating "'various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretations of test scores for specific uses" (p. 21).

[INSERT FIGURE1]

Within the current concept of validity, there are five identified types of validity evidence and, depending on the intended interpretations or use being evaluated, different combinations and/or permutations of evidence might be required through the validation process. These five forms of evidence can be used, as needed, with no formally mandated hierarchy or process.

We contend that, if validity is truly conceptualised with interpretations and use as being central, then evidence of the *actual* interpretations made and the use planned by score users must be included. Our view is that, regardless of how good the traditional forms of validity evidence are, if a score user does not interpret or use the resulting scores as intended, then the very validity of the intended interpretation and use should be in question.

Broadening validity evidence to incorporate a notion of evidence of interpretability could be achieved quite simply by including evidence of score report interpretability as one of the forms of validity evidence. This is represented graphically in Figure 2.

[INSERT FIGURE2]

While such an amendment to the current conceptualisation is appealing we do not believe it to be sufficiently ambitious. Essentially, there is an opportunity to not only broaden the types of evidence considered necessary, but also to re-evaluate the relative importance of each form of validity evidence and contemplate a potential hierarchy of evidence. Figure 3 presents a reframing of validity evidence for consideration.

[INSERT FIGURE3]

According to such a model, the broad definition of validity as "the degree to which evidence and theory support the interpretations of test scores" (AERA et al., 2014, p. 11) would remain. Rather, what is added is a distinction between the type and purpose of evidence. According to such conceptualisation, evidence is identified as being one of two meta-forms: evidence of interpretation/uses and technical evidence.

Technical evidence is that which is focused on assessing and evaluating whether a test is capable of achieving its set aims. Technical evidence is simply a categorisation for positioning the traditional forms of evidence that have been concerned with answering the key question: Is the test capable of achieving its certain state aims? That is, does the test achieve what it sets out to do? This evidence is the rigorous psychometric evidence that typically supports test development and currently dominates the literature on score validity.

Evidence of interpretations/use is that which is focused on assessing and evaluating the appropriateness, adequacy, accuracy, and effectiveness of user understanding of scores and the consequences of testing. In part, this incorporates aspects of the existing forms of validity evidence specifically those introduced by Messick (1989b) relating to evidence of consequences. What is added, though, is a clearer focus on the actual understanding of test users and the uses they plan. This type of evidence is focused on the interpretation, uses, and consequences of how scores are actually interpreted by users. Perhaps an easier explanation of this is that it is intended to help answer three key questions:

1. How are the outputs from tests (test scores via score reports) actually being interpreted?
2. Are these interpretations as intended?
3. Are the outcomes (read actions and uses), as a consequence of the testing, as intended?

Strategies and methodologies which provide answers to these types of questions are exactly the sort that are required in the evaluation of validity and are necessary in moving the conversation around validity forward.

Implied, but perhaps not clear, within the model presented in Figure 3 is the interrelated nature of the two forms of validity evidence in relation to the process of validation. Neither form can truly survive without the other. No matter how much and how good psychometric evidence is in support of a diagnostic test, for mathematics, it is meaningless if the results are not interpreted in ways in which appropriate interventions based upon students' needs are ensured and then undertaken. Like the proverbial tree in a forest, the best developed test could fall entirely without being heard as intended or, worse, in ways which were not intended. Conversely, no matter how well the end users' understanding and interpretations are based upon a score report, the outcomes are dangerous if based on a device with poor psychometric evidence. Taking this view of validity evidence one step further, Figure 4 positions the two meta-forms as equals in the process of validation with a validity argument, requiring examples of each form in support of the interpretations of test scores.

[INSERT FIGURE4]

What is hopefully clear in Figure 4 is the interrelated and interdependent nature of the forms of evidence in support of validity. Evidence of each of the meta-forms (i.e. evidence of interpretation /use and technical evidence) are potentially available in the construction of a validity argument. Provision of each form is a necessary but insufficient condition in the construction of a validity argument with both forms required.

Concluding Comments

At its very core, validity (and validation) is about the appropriateness and adequacy of interpretations and use based on test scores. Currently, however, there is a tendency to focus validation efforts on the technical aspects of test development and intended interpretations and use rather than also considering the practicality of how tests and their scores are actually interpreted and used. Score reports, on which these actual interpretations are made, are fundamental to the process of communication between test developers and their audience. As such, the interpretability of score reports (that is, how well members of the intended audience are actually interpreting and using scores as reported) is of the utmost significance and is fundamental in claims about validity.

Both MacIver et al. (2014) and Van der Kleij et al. (2014) have progressed the literature on this topic by proposing notions of user validity and score report interpretability, respectively. Building on their work, this paper has proposed that the currently acceptable forms of validity evidence need to be expanded to explicitly include evidence of interpretability. This means that evidence of user understanding of the message being conveyed by score reports must also be included. Another explanation of this notion of interpretability is related to the level of alignment between intended interpretations and use and actual interpretations and use resulting from engagement with the output of tests, score reports. In our opinion, inclusion of evidence of actual interpretations and use as an

integral part of validity is, without doubt, necessary. As previously stated, while the difference between intended and actual interpretations might seem trivial, they are of the utmost importance. It is almost absurd to think that the validity of the intended interpretations and use of a test might be compromised because there is a lack of alignment with the actual interpretations made and the use planned by the audience.

Integrating evidence of actual interpretation and use within the concept of validity and the process of validation as a necessity would also, in our opinion, provide much greater clarity about the responsibilities of test developers. In particular it is argued that it is possible to address or minimise the unintended side effects of legitimate test use by enhancing the interpretability of scores (via score reports) by their intended audience. This responsibility must fall to those responsible for test and, consequently, score report design and development.

Further, more than simply grounding the notion of validity in the real world by collecting evidence of actual score users' interpretations and use, we believe that an introduction of evidence of interpretability as an aspect of validity potentially lends itself to a complete re-think of the test development process. Test development is highly evolved, with sophisticated statistical strategies available to be deployed as necessary in the creation and technical validation of tests. Unfortunately, as discussed, there is much evidence that score users have been misinterpreting test scores. Moving forward, a more logical approach to the test design process might be based on the notion of understanding by design (Wiggins & McTighe, 2008). This involves working backwards from the intended interpretations and uses through an iterative process of score report design process, utilising strategies to collect evidence of the interpretability of scores at the field testing phase in order to inform necessary improvements. An interpretation- and use-focused approach makes far more sense as poor interpretations have the capacity to cause more harm than good, even if they are made based on exceptionally well-designed tests (Hattie, 2014).

While a complete description of such an approach is certainly beyond the scope of this paper, for those interested, a suitable example is that of the design and development of the score reports for the asTTLe online assessment tool in New Zealand (for a full set of technical reports, visit https://e-asttle.tki.org.nz/Reports-and-research/asTTle-technical-reports). The design of this system involved an iterative approach involving the collection of various forms of evidence with regards to user engagement with score reports and refining score report designs until a satisfactory level of use comprehension was reached. For example, as part of the evaluation of score reports, a test of comprehension was created to identify whether the audience could accurately interpret and identify appropriate actions. The initial average score for all participants was 60% (Ward, Hattie & Brown, 203). This was considered unsatisfactory and the reports were redesigned. The redesigned score reports yielded an improvement in the average to over 90% (Hattie, 2009). Such an improvement in

the interpretability, arguably, also improved the overall validity of the intended interpretation and use.

Returning to the example where we began. The intended purpose of the hypothetical diagnostic educational test was to provide evidence to a teacher about their students in order to inform future instruction. The intended interpretations of such an assessment would be focused on the learning needs of students and the intended uses centred on planning appropriate interventions. When a teacher receives the output from such a test and is able to correctly identify the needs of their students and makes plans to meet those needs, then there is a clear alignment between intended and actual interpretation and uses. Assuming additional technical forms of validity evidence are sufficient, then the intended interpretations and uses might reasonably be deemed valid. If, however, the output was misinterpreted and the teacher did not correctly identify the needs of some (or all) of the students and planned interventions which failed to meet their specific learning needs, then the interpretations and uses would not be in alignment with what was intended. In such circumstances, regardless of the strength of other forms of validity evidence, the very validity of the intended interpretations and use should be questioned. As in the example of asTTLe, if this was the case, before proceeding, it would be necessary for test developers to engage in a report redesign process aimed at improving the interpretability of the score reports.

Moving forward, whilst there are examples of strategies for investigating users' interpretations of score reports, there is further work required for appraising how these methods might best service the provision of validity evidence of the interpretability of scores focused on score users' actual interpretation and uses of scores as an aspect of validity.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014) *Standards for educational and psychological testing (4th ed.)*. Washington, DC: American Educational Research Association

American Psychological Association (1954) *Technical recommendations for psychological test and diagnostic techniques.* Washington, DC: Author.

Bennett, R. E. (2010) Cognitively Base Assessment of, for, and as Learning (CBAL): A Preliminary Theory of Action for Summative and Formative Assessment. *Measurement, 8,* pp. 70 – 91.

Cizek, G. J. (2012) Defining and Distinguishing Validity: Interpretations of Score Meaning and Justification of Test Use, *Psychological Methods*, 17(1), pp. 31 – 42.

Cronbach, L. J. (1980) *Validity on parole: How can we go straight? New directions for testing and measurement -- Measuring achievement over decade* -- Proceedings of the 1979 ETS Invitational Conference (pp. 99-108). San Francisco: Jossey-Bass.

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

Cronbach, L. J. & Meehl, P. E. (1955) Construct Validity in Psychological Tests, *Psychological Bulletin*, 52, pp. 281 - 302

Goodman, D.P. & Hambleton, R.K. (2004) Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education,* 17(2), pp. 145 – 220.

Guion, R. M. (1977) Content validity: The source of my discontent, *Applied Psychological Measurement*, 1 (1), pp. 1—10.

Haertel, E. (2013) How Is Testing Supposed to Improve School?, *Measurement*, 11, pp. 1-18.

Haladyna, T.M. (2006) Roles and Importance of Validity Studies in Test Development. In S.M. Downing & T.M. Haladyna (Eds), Handbook of Test Development, pp. 739 – 750. Mahwah, NJ: Erlbaum

Hambleton, R. K., & Slater, S. (1997) *Are NAEP executive summary reports understandable to policy makers and educators? (CSE Technical Report 430).* Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Teaching.

Hattie, J. A. C. (2009, April 16) *Visibly learning from reports: The validity of score reports*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Hattie, J. A. C., & Leeson, H. (2013) Future Directions in Assessment and Testing in Education and Psychology. In APA Handbook of Testing and Assessment in Psychology: Volume 3. Testing and Assessment in School Psychology and Education, pp.591 – 622, Washington, DC: American Psychological Association.

Hattie, J. A. C. (2014) The last of the 20th Century Test Standards, *Educational Measurement: Issues and Practice*, 33(4), pp. 34 – 35.

Hattie, J. A. C. (2015) *What works best in education: The Politics of Collaborative Expertise*. Pearsons

House, E. R. (1980) *Evaluating with Validity*, Beverly Hills, CA: Sage

Ho, A. (2013) The Epidemiology of Modern Test Score Use: Anticipating Aggregation, Adjustment, and Equating, *Measurement*, 11, pp. 64 – 67.

Hubley, A. M., & Zumbo, B. D. (1996) A dialectic on validity: Where we have been and where we are going, *The Journal of General Psychology,* 123, 207 – 215

Hubley, A. M., & Zumbo, B. D. (2011) Validity and the consequences of test interpretation and use. *Social Indicators Research,* 103 (2) , pp. 219 – 230

Jaeger, R. M. (1998) *Reporting the results of the National Assessment of Educational Progress (*NVS NAEP Validity Studies). Washington, DC: American Institutes for Research.

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

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

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

Linn, R. L. (1989) Validity. New York: Macmillan.

Lizzitiz, R. W., Samuesen, K. (2007) A Suggested Change in Terminology regarding Validity and Education, *Educational Researcher,* 36 (8), pp 437-448.

MacIver, R., Anderson, N., Costa, A. & Evers, A. (2014) Validity of Interpretation: A user validity perspective beyond the test score, *International Journal of Selection*, 22(2), pp. 149 – 164.

Messick, S. (1989b) Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.), 13-103). New York: Macmillan.

Messick, S. (1989a) Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5 – 11

Messick, S. (1998) Test Validity: A Matter of Consequence, *Social Indicator Research*, 45 (1), pp 35 - 44.

Messick, S. (2000) Consequences of test interpretation and use: The fusion of validity and values in psychological assessment. In R. D. Goffin and E. Helmes (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy,* pp. 3 - 30. Boston: Kluwer Academic Publishers

Sheppard, L. A. (1993) Evaluating Test Validity. In L. Darling-Hammond (Ed.), *Review of Research in Education* Vol. 18 (pp. 405 – 450). Washington, DC: American Educational Research Association.

Sierci, S. G. (2009) Packing and unpacking sources of Validity Evidence: History Repeats Itself. In R. W. Lizzitz (Ed.) *The Concept of Validity: Revisions, New Directions, and Applications*, pp. 19 - 38.

Tenopyr, M., L., (1977), Content- Construct Confusion, *Personnel Psychology*, 30(1), pp. 47 – 54.

Van der Kleij, F. M., Eggen, T. J. H. M, & Engelen, R. J. H. (2014). Towards valid score reports in the Computer Program LOVS: A redesign study, *Studies in Educational Evaluation, 43*, pp. 24 – 39

Wiggins, G. & McTighe, J. (2008). Put understanding first, *Educational Leadership*, 65(8), pp. 36-41.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & Sinharay (Eds.) *Handbook of statistics, vol 26: Psychometrics*, pp. 45 – 79, The Netherlands: Elsevier Science B. V.

Zumbo, B. D. (2009). Validity as a contextualised and pragmatic explanation, and its implications for validation practice. In R. W. Lizzitz (Ed.) *The concept of validity: Revisions, New Directions, and Applications* (pp. 65 - 82) Charlotte, NC: IAP - Information Age Publishing, Inc.

Zumbo, B.D. (2015, November). Consequences, Side Effects and the Ecology of Testing: Keys to Considering Assessment 'In Vivo'. Keynote address, the annual meeting of the Association for Educational Assessment - Europe (AEA-Europe), Glasgow, Scotland. [http://brunozumbo.com/aea-europe2015/]

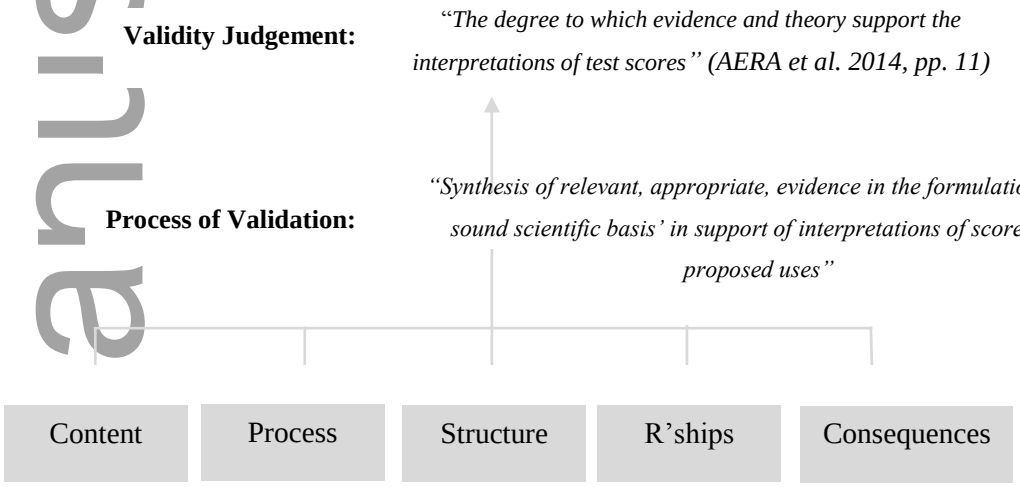

Figure 1 Current Conception of Validity & Validity Evidence

Types of Validity Evidence:

Figure 2 Expansion of validity evidence to incorporate evidence of user interpretation

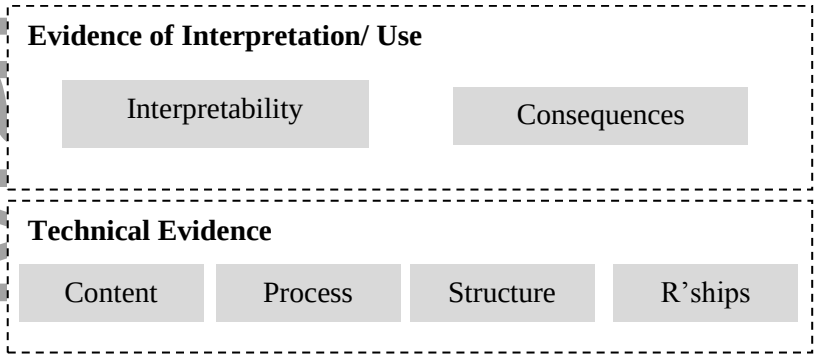

Figure 3 Validity evidence including user interpretations

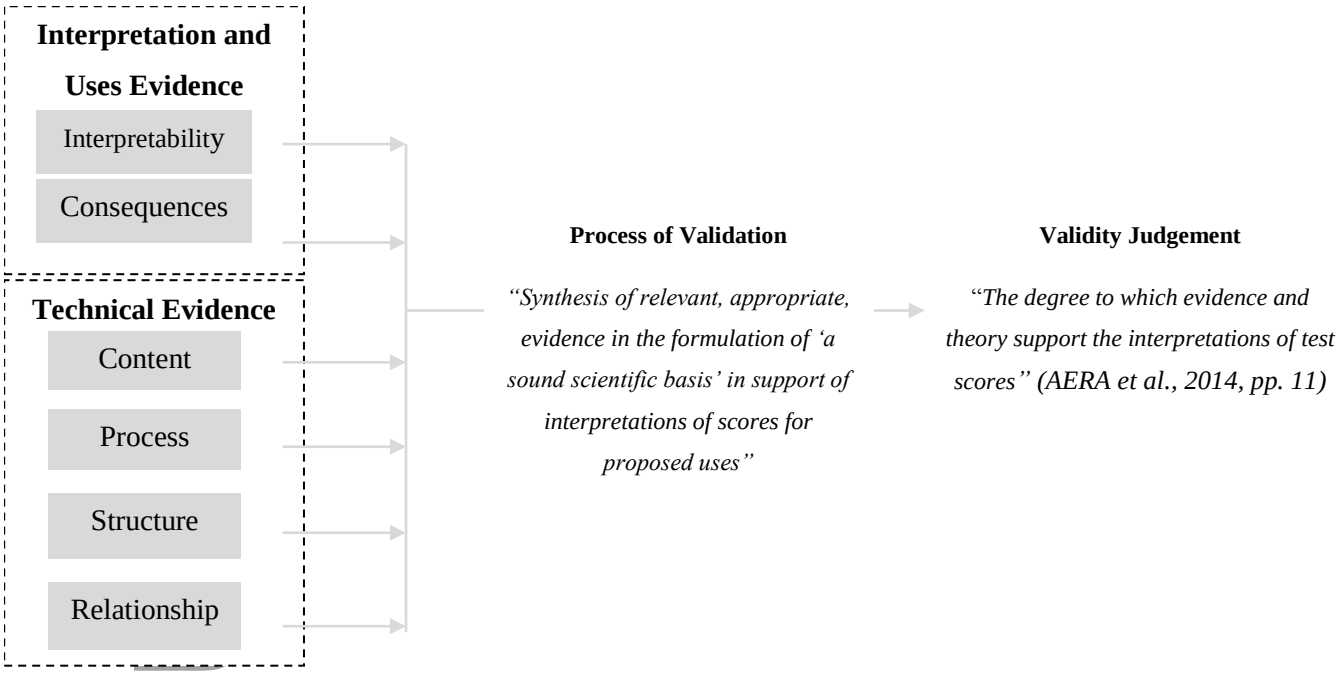

Figure 4 Current Conception of Validity & Validity Evidence