VALIDITY OF INTERPRETATION: A USER VALIDITY PERSPECTIVE BEYOND THE TEST SCORE

RAB MACIVER
Brunel University

NEIL ANDERSON
Brunel University

ANA-CRISTINA COSTA
Brunel University

ARNE EVERS
University Of Amsterdam

Correspondence Address for Lead Author: Brunel Business School, Brunel University, Uxbridge, UB8 3PH, UK. iain.maciver@brunel.ac.uk

Abstract

This paper introduces the concept of *user validity* and provides a new perspective on the validity of interpretations from tests. Test interpretation is based on outputs such as test scores, profiles, reports, spread-sheets of multiple candidates' scores, etc. The *user validity* perspective focuses on the interpretations a test user makes given the purpose of the test and the information provided in the test output. This innovative perspective focuses on how user validity can be extended to content, criterion and to some extent construct-related validity. It provides a basis for researching the validity of interpretations and an improved understanding of the appropriateness of different approaches to score interpretation, as well as how to design test outputs and assessments which are pragmatic and optimal.

Validity of Interpretation:

A User-Validity Perspective Beyond the Test Score

Test score interpretation rather than test scores are at the centre of contemporary definitions of test validity (e.g. American Educational Research Association, American Psychological Association, National Council for Measurement in Education, 1999; Kane, 2009; Messick, 1989). Yet to what extent are the current conceptions of validity and the practice of validation concerned with the validity of interpretation? Criterion-related validity, for example, and the practice of validating occupational assessments (including meta-analyses) is still to a large extent focused on establishing relationships between test scores and criteria. Test scores rather than actual interpretations are the focus of validation. There is a need to further develop and articulate a validity of interpretation to better address the shift in the definition of validity towards interpretation and the decisions individual test users are actually making based on test scores.

The test user for the purposes of this article is the interpreter of the test output: a manager reading a psychometric multi-trait narrative report summary on a candidate prior to interview; a trained test user interpreting a personality profile to aid the test taker's development; a test taker being given direct online feedback to aid their self-selection as to whether to apply for a particular role; an I/O Psychologist deciding what rules or equations to apply in a selection system for a particular role, etc. and what latitude (if any) to give other test users in influencing selection decisions. A broader definition of a test user might include an applicant who has not received feedback, however, for the purposes of this article such an individual would be regarded as a test taker rather than a test user.

Although evidence from different aspects of validity such as content and criterion-related validity provide a basis to support potential test interpretations, they generally do not focus on the validity of interpretation from the users' perspective itself. In their stead, they tend to

focus on providing general support for a particular proposed interpretation that a test user could make. This support for an interpretation may be achieved, for example, by demonstrating that there is a relationship between a test score and criterion. While establishing the relationship between a predictor test score and a criterion can be an important consideration in providing a justification for the efficacy of a measure being used to make interpretations (and decisions), it does not provide a direct basis for validating the actual interpretations that test users are making in the real world or for the effectiveness of other alternative approaches to interpretation they could take. It also does not consider how appropriate, effective and acceptable the method of interpretation is for a given application.

There is a long history of investigating the relative validity of different methods of forecasting ranging from the validity of algorithms (mechanical validity) to human users making decisions from test scores (clinical validity) and the validity of interpretations made by test outputs and reports.

Mechanical (or actuarial) validity is the criterion related validity where an algorithm or procedure is used and applied consistently (e.g. by a computer program). The result from the application of the procedure when correlated with the outcome criteria, gives evidence of the validity of the algorithm. In practice, the algorithm can be informed by experience (Goldberg, 1970) (lay or expert), or could be based on previous statistical evidence of relationships between a predictor (e.g. cognitive test) and a criterion based on the results of a previously conducted criterion-related validation study.

Clinical validity, by contrast, refers to the validity of test users' aggregation of data to form evaluations of individuals based on test scores (profiles or test outputs). The correlation between the interpretative evaluation a test user makes with a criterion measure or classification provides a direct form of validity evidence which supports 'clinical

interpretation' (Meehl, 1954; Grove, Zald, Lebow, Snitz, & Nelson, 2000). Clinical validity has been criticized as an inappropriate and potentially misleading label for this form of interpretation (e.g. Westen & Weinberger, 2004), primarily, as 'clinical' tends to confuse the method of aggregation of data with the validity of clinicians making judgements i.e. clinical diagnoses of psychopathology.

The findings comparing 'clinical' methods to mechanical approaches, where an algorithm is used to aggregate data, are consistent and clear. Mechanical validity is more effective in forecasting outcomes across a wide number of disciplines including medicine, mental health, personality and education and training settings than clinical interpretation (Grove *et al*., 2000). Given that the original superiority of algorithms over 'clinical' interpretation was reported by Meehl in 1954 in a review of 20 studies, it might be expected that these findings would have strongly influenced assessment practice in the field from which much of the work originated. However, a survey of the use of different approaches in clinical practice found that only 31% used mechanical approaches in comparison to 98% for clinical combination of data (Vrieze & Grove, 2009). Thus, it would seem important to both encourage the use of mechanical approaches where possible, but also to consider how to optimise the validity of test interpretations made by human users.

In the internet age interpretation of scores from assessments can be based on computer generated outputs e.g. profile chart or narrative reports for an individual or a spread-sheet summarising assessment scores for multiple individuals. Whether making interpretations directly from test scores, users drawing conclusions from the computer based test outputs, or narrative reports written by expert test users for others users consumption, interpretations are required to be made from test outputs.

It is argued in this paper that there is a need for a new perspective of validity which focuses on the validity of interpretation*: user validity*. Such a perspective needs to be integrated with and extend further contemporary conceptions of validity. It needs to focus on the validity of interpretations that are actually made by test users based on the information available to them. It must seek to consider, mechanical interpretation as well as clinical interpretation and seek to address the role of the test output in supporting interpretation. Such an approach, it is argued, will provide the basis for researching appropriate methods of interpretation for a particular application of a test in a particular context. This approach could potentially provide the basis for informing and evaluating the appropriate design for test outputs which best support valid interpretation. In essence, this paper seeks to advance the literature, by outlining a more coherent perspective on how to conceptualise the validity of interpretation. This is first achieved by providing a framework for the validity of different methods of interpretation. Secondly, the aim is furthered through a delineation of the concept of user validity and integrating this new concept into important established validity perspectives.

This paper begins with an overview of the common methods of validating test interpretations, in order to review the relevant literature that relates to this topic. Secondly, a framework is outlined for valid interpretation, highlighting the key factors to consider that can, for example, impact on the design of test outputs (including test scores) to ensure that they provide a basis for appropriate and valid interpretation. Thirdly, the perspective of user validity is introduced as a new formulation for the validity of interpretation. This is conceptualised in relation to different established forms of validity evidence and modern perspectives on validity to extend current perspectives on validity.

Validity and test output interpretation

As noted earlier, there are two different approaches that form the basis of interpretation of information generated from individuals completing tests: human test user (clinical)

interpretation and mechanical interpretation. Both of these forms of interpretation can be guided by expert judgement or by prior statistical evidence (e.g. criterion-related validity studies). Human user interpretation can also be guided by computer based test interpretation (CBTI). This provides support in the form of test outputs by supplying test users with interpretations. How individuals make interpretations and evaluations is one of the questions addressed by the heuristics and biases movement. In the following section different interpretations are reviewed concerning, mechanical interpretation, clinical interpretation and computer based test interpretations, followed by a brief overview of some issues arising from the heuristics and biases perspective related to interpretation.

Differences in interpretations and the heuristics and biases perspective

Mechanical Interpretation

Mechanical interpretation provides a consistent mechanism for making decisions based on algorithms or sets of rules. It allows data from tests to be applied in a consistent mechanism to all individuals being tested. The fact that the validity of clinical interpretation has been demonstrated on average to be lower than mechanical algorithms, does not, however, provide evidence that statistical methods are better than experts in developing algorithms. The meta-analytic evidence does not provide a basis for a material difference in forecasting outcomes, based on whether the algorithm was created by a panel of experts or the algorithm was generated by applying a statistical method e.g. based on a previous validation (Grove *et al.*, 2000*).* One argument against the use of the mechanical approach is the cost of creating such an algorithm. The data indicates that it is possible to do this without the need for a large rigorous and potentially expensive validation study. That is a panel of experts can provide a valid algorithm. Even the expense of assembling a panel of experts may not always be necessary if we consider the use of a simple baseline model. There should not always be the presumption that the complex regression model (or expert model) is always optimum

(Hogarth, 2012). The Dawes rule provides a simple baseline model where all the predictors are correctly aligned in the direction of prediction and are added together to create a unit weighted sum (as opposed to a regression model for example, where the beta weights indicate different weights for different variables). Surprisingly, the performance of this simple equation has been shown to perform remarkably similarly to the optimised differential weights provided by a regression i.e. without an appreciable loss in forecast accuracy (Dawes & Corrigan, 1974; Wainer, 1976). Dawes and Corrigan (1974) proposed that it is more important in forecasting accurately to have the correct variables in the equation than to differentially weight them to optimise prediction. Thus there may be circumstances where cost and time do not permit, but where it is still possible to create simple mechanical equations which have robust validity (i.e. by unit weighting variables that are known/expected to be related to the criterion). However, the improvement in our capacity over time to manage larger databases of data is likely to make differential mechanical weighting of data more feasible, if it is coupled with better development of performance metrics that can be applied to measure work effectiveness of individuals (e.g. both role specific and across roles).

Clinical Interpretation

Meta-analytic empirical evidence indicates that the disparity between 'clinical' and mechanical can, in large part, be accounted for by the inconsistency of human judges (Karelaia & Hogarth, 2008). In fact, experts interpreting 'clinically' also fair more poorly than an algorithm they have themselves created (Goldberg, 1970). These findings could be argued to support more widespread use of mechanical approaches and they could also be viewed as arguing for the need to find approaches that make human user decision making

more consistent. The lack of consistency of experts could be seen to result from processing errors which computers tend not to make (Kahneman & Klein, 2009), or an alternative explanation is that at least some interpreters are trying different heuristics (i.e. experimenting) when they are making certain interpretations or forecasts (Todd & Gigerenzer, 2012).

Computer based test interpretation

Online technology offers many possibilities for the reporting of individual and group (multi-individual) data. Reports based on test scores can be written by expert test users for other stakeholders to interpret or generic computer generated reports are also available as a mechanism to support human user interpretation e.g. computer based test interpretation (CBTI) of test scores driven by algorithms (International Test Commission, 2005; Butcher, Perry & Atlis, 2000). It should be noted that computer generated reports may be restricted to appropriately qualified (trained) users, in some instances, for a particular application or may be available to an untrained user e.g. fed back automatically by an online computer assessment system to an online test taker. There is a distinction between generic reports and reports which are developed for a particular testing project/client with very specific testing aims. The more widely applicable and more generic reports may provide, for example, narrative interpretations of test scores or may also provide derived scores as a basis for further interpretation.

When outputs are validated then, the unit of analysis can vary from individual narrative statements/interpretations through to the evaluation of whole reports. Approaches to the validation of CBTI reports include user satisfaction studies, statistical modelling, expert modelling, and external criterion studies.

User satisfaction studies seek to evaluate the perceived accuracy and adequacy of the information contained in reports either from the perspective of the test taker or other

stakeholders (i.e. the user). Clearly such studies are open to the Barnum (or Forer) effects where interpreters are gullible to generalities of human personality (Dickson & Kelly, 1985). Butcher *et al.* (2000) highlighted, for example, that users are more satisfied with reports that have a higher proportion of non-specific statements. The fact that test outputs give results which readers interpret as valuable, useful or meaningful is of course not to say that the test outputs are differentiating meaningfully and validly between the people that have been tested. Clearly, the evaluation of acceptability of reports/test outputs needs to be tied to their adequacy in fulfilling their intended purpose (e.g. overall recommendation in selection or accurate descriptive statements of individuals' behaviour). While methods such as interweaving bogus and real narratives has been attempted (Hoover & Snyder, 1991), the evidence more generally provided from this type of study tends to fall into the category of user acceptability studies rather than providing direct evidence of the validity of interpretation. However, if particular outputs provide information which test users find more acceptable, it is reasonable to infer that such outputs are more likely to be used. Researching the acceptability of different methods of interpretation from different test outputs may help to clarify which outputs are most likely to be applied in practice.

In *statistical modelling studies*, the focus is on demonstrating that actuarial findings form a rationale for the interpretations in test outputs. The validity of the computer narrative is likely to be higher where the developer of the system closely conforms to the actuarial findings of validity for the instrument (Butcher, 1995). The development of such narratives can also be guided by experts. Vale and Keller (1987) recognised that *expert-opinion modelling* comprises only an initial validation strategy for computer based test interpretation. Indeed, simulating experts will tend to maintain the status quo and is only likely to lead to improvements in assessment accuracy where the experts themselves improve (Honaker & Fowler, 1990).

While the external *criterion-related validation* evidence can suffer from the methodological problems associated with criterion-related validation studies in general (e.g. the need for large sample sizes and reliable and valid criteria) they provide a method of correlating the interpretation provided with criterion score(s). To this end they provide a relatively direct indication of whether the interpretations in the report provide valid forecasts of criterion scores.

Computer Based Test Interpretative (CBTI) reports have not been extensively validated and the results have not yielded a consensus about the effectiveness of these reports in the literature (Snyder, 2000). Indeed, the studies on the reliability, validity and utility of CBTI have been sufficiently divergent in nature to permit starkly different interpretations of their effectiveness (Garb, 2000; Butcher *et al*., 2000). There is a need for more studies, and for these to be organised as part of a more coherent framework of validity which integrates and allows comparison of CBTI, clinical and mechanical interpretation. Test manuals should also seek to provide greater information which relates to the validity of interpretation that arises from test outputs not just that arises from instruments' test scores.

Heuristics and Biases Perspective

Human decision making provides an important perspective in understanding the evaluative component of test interpretation. The heuristics and biases movement (partly influenced by Meehl's early work on mechanical versus statistical interpretation) has identified many sources of bias in human interpretation and evaluation of data. A number of these biases have the potential to affect the quality of interpretative evaluations, in practice, including a failure to take account of either base rates or regression to the mean (Kahneman, Slovic, & Tversky, 1982). For example, where the base rate of job success is low, test interpreters will tend to overestimate the chance of success of new employees based on interpreting a high standardised predictor test score. Regression to the mean can also cause a systematic error in

estimation by a human interpreter e.g. a very cognitive high test score on average is likely to result in an elevated job proficiency score, but in these circumstances a human interpreter is also likely to consistently overestimate the degree of elevation in the criterion score.

The heuristics and biases movement generally compares human decisions with what are considered optimal solutions e.g. derived from statistical or perfect models. Heuristics focus on rules of thumbs that humans can use to make decisions. Humans have evolved to perceive and act with more urgency than they have a need to evaluate (Gilovich, Griffin & Kahneman 2002) and satisficing contends that humans have developed heuristics that are 'nimble tricks' to perform in the 'quirky structures' of the real world. These may not always be optimal, but they are adapted to the human needs and the situations humans often find themselves (Simon, 1956). Fast and frugal heuristics are 'satisficing' heuristics that humans can deploy that work with the minimum of time, knowledge and computation (Gigerenzer & Todd & the ABC Group, 1999). One example of a fast and frugal heuristic is 'Take the Best' (Gigerenzer & Goldstein, 1996). The decision making variable with the highest validity is used and the instance/subject with the highest value on that variable is selected (before moving to the next instance and then the next most valid variable or cue).

A key question is when different heuristics are likely to be appropriate and valid. Different data environments that are perceived (or received by the user) have different structures, and therefore to be able to interpret which heuristic is most effective requires an understanding of key aspects of each data environment (Simon, 1956; Todd & Gigerenzer, 2012). Todd and Gigerenzer (2012) have proposed features of the environments which help to determine which heuristic or heuristics are likely to be appropriate. First the *degree of uncertainty* which refers to the validity of the available cues (predictors) to predict a criterion. Kahneman and Klein (2009) distinguish certainty from *predictability* where the outcome can be uncertain, but highly predictable (a sporting event where there is a clear favourite is an

example of such an environment). A criterion validation study's ability to show that chosen predictors provide some accuracy of forecast in a future study (cross validated validity) would be a basis for a demonstration of *predictability* in the environment (the results of multiple correlation in one sample – not cross validated - would fall short of demonstrating this). The *number of alternatives* may pose problems as where the number of alternatives is high, the processing requirements will be heavy, if all the data is considered in analysing the information. Simpler strategies or heuristics are likely to perform well in forecasting under such circumstances. Developing heuristics on smaller *sample sizes* will favour simple heuristics; very large sample sizes in very predictable environments are more likely to favour statistical methods which weight the relative importance of data. *Redundancy* refers to how highly correlated cues are in the environment and where this is the case fast and frugal heuristics such as 'Take the Best' are more likely to be effective. Finally, the variability of the validity of cues can mean that approaches that use the best predictor will tend to be better (Hogarth & Karrelaia, 2005; 2006). Although, if it is not known which is the best predictor/cue prior to forecasting then other strategies which 'hedge their bets' on which is the most valid cue such as the Dawes rule are likely to be effective.

The effort-accuracy trade off hypothesises is that more complex processing of data will result in more valid forecasts (Shah & Oppenheimer, 2008). However, this trade-off is only likely to manifest itself under certain circumstances as some simple heuristics are likely to be good strategies when certainty is low, the number of alternatives is high, the sample size low, the redundancy in the variables high and the variability of predictors' validity is high.

Summary

In isolation, the present notions of validity of interpretation lack a coherent framework and can tend to pit one form of validity against another rather than look at what are the relative

advantages of the alternative methods of interpretation that are likely to be acceptable. A more coherent approach is to develop a framework that indicates the forms of interpretation that are likely to be effective in particular circumstances and therefore guide the design of the test output to best support valid interpretation. Given, that different forms of interpretations are supported by different outputs, a realistic approach is to seek to improve the validity of interpretation of each of these alternative approaches in different contexts. Current perspectives on interpretative validity are also not directly related to established aspects of validity, for example, content, criterion and construct. If we are to be able to evaluate the validity of test interpretations rather than test scores, this requires established views of validity to be extended to more actively consider the validity that arises when users interpret.

Firstly, then, there is a need for a framework for the validity of interpretation that indicates when different forms of interpretation are likely to be appropriate and this has to integrate test outputs as a fundamental driver of interpretation. Secondly, there is a need to provide a perspective on how existing aspects of validity integrate with an interpretative perspective on validity. The next two sections deal with these two issues in turn.

A framework for valid interpretation from test outputs

Figure 1 presents an overview of a framework for valid interpretation that is designed to result in appropriate test outputs. It provides an overview of the key variables that influence the appropriate form of interpretation which is likely to be effective given the purpose the test is being put to, the context and the structure of the data. And that the appropriate form of interpretation will impact on the design of test outputs (and possibly tests and test scores themselves).

- -

INSERT FIGURE 1 ABOUT HERE

- -

Purpose/Aims/Flexibility of Use

Firstly, it is important to explicitly set out what are appropriate and inappropriate uses of a test output (e.g. a long generic computer generated narrative report may be effective in providing a description of the results of a candidate on a personality questionnaire, but forecasting performance in teams is likely to be better served by displaying scores which result from forecast algorithms that have previously demonstrated efficacy in correlating with relevant team criteria). One consideration that impacts how an output is designed is the flexibility that is required from the output. If the output is required to match against one job role, then potentially one fit equation can be used actuarially, when a profile of test scores could be used for a very high number of different job roles then a much greater flexibility is required from the test output.

Data Structure/Environment

Once the purpose has been explicitly and accurately defined for the output for a particular use or set of uses, there can be a consideration of the data environment which is being interpreted or forecast. The data environment may be expected to be different depending on the purpose to which the test output is being put to. For example, in an overall selection decision, a number of test scores may need to be used in combination, whereas the simple behavioural prediction of frequency of a particular behaviour may merely require the inspection of one variable. The data structure may also vary in terms of predictability, number of alternatives, predictor cue variability, and redundancy.

Context/Need for:

Before deciding on what information is to be presented in an output, the context in which the testing is being used has also to be considered. A fundamental concern is likely to be the relative importance of the operational validity of the interpretation in forecasting outcomes in comparison to other considerations. While the merit of a test is normally indicated by concerns about validity, in practice, for particular applications the maximisation of validity cannot be achieved at any price. Validity has to be considered relative to other considerations which clearly include factors such as time and cost. The design of the output may, for example, have to take account of what will be acceptable to users. Stakeholders may also have an interest in maintaining their personal involvement in decision making (Hogarth, 2012). This may not just be experts protecting their vested professional interests, but may also come from a belief that their decision making is superior to an algorithm regardless of the validation evidence. The belief in the ability to effectively forecast an outcome is not a good indicator of actual success in predicting an outcome (Kahneman & Klein, 2009).

Method of Interpretation/Aggregation

These and other considerations may mean that a mechanical approach may not be readily accepted as an option and finding the most valid alternative may need to be considered. Semi-structured approaches where experts are provided with guidance are one alternative (Kahneman & Klein, 2009). It may also be beneficial to favour interpretations in outputs based more closely on heuristics that an expert might use. In practice, then, it may be more practicable to find a method which is acceptable and increases the consistency of decision making and which involves stakeholders, rather than always seek to replace experts with an algorithm. It may also be the case that given the data environment, that simple methods may be the most valid, such as the best single predictor (BSP), (McGrath, 2008), take the best

(TTB) (Gigerenzer & Goldstein, 1996) or the Dawes rule (Dawes & Corrigan, 1974). Although, in practice where mechanical approaches can be successfully implemented, the evidence is that a mechanical approach should be implemented to improve the validity of forecast (Grove *et al.*, 2000).

Test Report Output Decisions

The layout and information provided in reports are likely to vary as a result of which method of interpretation is appropriate. Where a fast and frugal heuristic has been selected to be used, for example, spread-sheets with candidate scores can be organised to enable 'Take the Best' to be more easily applied (e.g. order the columns of data by the validity of the predictors and sort the individuals based on their score on the highest validity predictor variable from high to low).

Test Scoring and Content Decisions

Finally, it is logical that such considerations, may under certain circumstances, feed through into the design of the test itself. That is if a large number of variables are being excluded from being used as a basis interpretation due to their lack of contribution to valid decision making or redundancy, it calls into question the benefit of maintaining these variables in the test – and unless these variables have validity for other interpretations they can be removed to shorten administration with no loss of validity (the removal of non-valid cues may even serve to increase the validity of interpretation in certain circumstances such as unstructured human user ('clinical') profile interpretation or in the application of the Dawes rule).

The user validity perspective on interpretation

The second concern with the current conceptions of test validity outlined in this paper is that they do not focus adequately on a validity of interpretation. *User validity* is proposed as a

perspective on validity which specifically focuses on the validity of interpretations that users are responsible for making from different test outputs. The authors define user validity here as..

"User validity is the overall accuracy and effectiveness of interpretation resulting from the test output."

The start point for these interpretations is the purpose or aim of the test. These interpretations could be based on a set of test scores (or profile), descriptions of test scale/scores on profiles, expert user written narrative, computer based output reports or a spread sheet of multiple candidate scores. The user validity perspective places the focus on the validity of the interpretations in use and the decisions that form part of these test interpretations. This perspective is designed to be a lens with which to view and prioritise validity information with respect to how it supports (or falsifies) interpretations that users make from test outputs (including test scores). It is also designed to build on contemporary conceptions of validity and provide a focus for research enquiry that is user centric. To provide a richer understanding of the user validity perspective, the Trinitarian (Guion, 1980) conception of validity is now discussed with reference to the concept of user validity. The three pillars of the Trinitarian perspective are content related, criterion-related and construct-related evidence. *Content related evidence* is typically based on consensual informed judgments about the representative coverage of content in a test, in that a test appropriately samples a particular domain of interest (Messick, 1989). *Criterion-related evidence* is indicated by the relationship between a predictor and a criterion and is the dominant perspective in validity where an appropriate criterion is available e.g. personnel selection, classification and job placement. But where this is not the case there is a need to rely on other forms of validity evidence. Validity as a concept has evolved towards the evaluation of interpretations and the meaning of test scores rather than validation of test scores themselves (Sireci, 2009).

Another shift in the validity concept has been away from individual types of validity to a more unitary perspective of validity as an overall evaluation. The perspective that the last of the trinity, *construct evidence*, should be regarded as all of validity stretches at least as far back as Loevinger (1957) and the view that validity is unitary has been widely articulated (Messick, 1988; 1989; Shepard, 1993; Linn, 1997). Different forms of validity are now regarded as different forms of evidence contributing to on overall evaluation of a test's validity (Messick, 1989; AERA, APA & NCME, 1999).

User content-related evidence

From the perspective of user interpretation, where a test profile or output provides direct descriptions of the content of the test, then the descriptions of the content in the test outputs should provide both an appropriate sample of the domain of interest and be an accurate representation or summary of the nature of the test content itself. For example, if a personality questionnaire profile scale descriptions are not an appropriately representative sample of the content of the scales, the descriptions provided will have the effect of misleading users' interpretations. Thus subject matter experts' consensual judgments on whether user outputs accurately reflect the domain of interest and test content constitute a form of *user available content evidence*. That is to say, it is the validity of the content in the test output which is being made available to the user to interpret. The content detailed in the test outputs, could under-represent, mis-represent or even over-represent the domain of interest. Indeed, the content of the output may accurately reflect the domain of interest, where the test content does not – under such circumstance the test report would be making inferences which were not directly supported by the test content.

If the language in the description or narrative interpretation in the report is difficult, esoteric or not likely to be understood by the target group (test users who are interpreting the output)

then this presents content validity concerns with the output. Proper consideration needs to be given to how test reports and outputs are designed to ensure they accurately and straightforwardly convey score meaning to the user (see for example, Zenisky & Hambleton, 2012). A more direct form of evidence with regard to interpretation is to compare users' actual interpretations to the domain of interest and the content of the test, which would be to provide *user received content evidence*. This evidence provides a picture of whether test users are establishing or receiving an accurate interpretation of the content of the test (based on the test output) and that these interpretations are validly sampling the domain of interest. Thus interviewing users about their understanding of the content of a test scale could lead to a summary of the content users receive and this content can be related to the content actually assessed in the test and indeed in the domain of interest. It could be for example, that the test users are overweighting the importance of some negative interpretation provided in the test output. In practice, test user training and supporting documentation such as test manuals, should also be assessed for the accuracy of the content they make available to test users. These documents would also be expected to impact on the accuracy of interpretations that test users make (based on what they receive). One aim of the content provided in a test output is to help the user make accurate and valid interpretations. However, if the design and description in the output are not accurately emphasising (making salient), the most important points in a test output, then the content will not be accurately interpreted (and received) by the test user.

As opposed to user content evidence, standard content related evidence provides information about the content of a test as an instrument providing test scores, rather than evidence to support inferences made from individuals' test scores/test outputs. All forms of content related evidence described here, including the user content related evidence put forward in this paper can provide a justification for use when the interpretation is a simple summary of

observed performance and this form of evidence is almost universally appropriate for different uses, but is rarely in itself enough to justify most interpretations, particularly when those uses are related to assessing or forecasting performance (Kane, 2009; Messick, 1989). When the interpretations in the output go beyond the content assessed in the test they require further forms of validity evidence to justify their use. User content evidence then is useful as it can provide greater clarity as to when further forms of evidence are required to support the more ambitious interpretations made in test outputs.

User criterion-related evidence

User criterion-related evidence is a form of validity evidence which assesses the validity of the interpretation with respect to criterion outcomes. It is an extension of 'clinical' validity and is the validity of any interpretation that a human user makes from different test outputs or test scores. Again, two forms can be delineated. *User available criterion–related* evidence is the evidence which is available to the user prior to interpretation (i.e. the validity in the information that is presented in the test output). User available evidence can be established by correlating some aspect of the report itself (e.g. profiled test score, presence of a particular narrative statement describing an individual's likely behaviour) with a relevant outcome. This would be an evaluation of validity being made available to the user in the test score or output rather than the validity that the user has received and interpreted. Such available evidence provides support for the particular interpretations that are being proposed rather than direct evidence of the validity of interpretations themselves which is in the form of *user received evidence*. For example, where a generic test output such as a test score profile is being used by qualified test users, the validation of the individual test scores, provides an indication of the validity that can be made available to the test user, but does not provide an accurate basis for the evaluation of the operational validity of the test output (and often therefore by implication the test) in use.

Direct *user received criterion validity* evidence is provided by the relationship between the interpretation the user makes (the decision which stems from their interpretation based on the information they receive from the test output) and an appropriate criterion or outcome. The overall evaluation of personality profiles by users to recommend candidates correlated with a criterion of work effectiveness would be one example of a direct form of *user received validity evidence*. It could also be the decision a recruiting manager makes about the eligibility (or otherwise) of candidates for interview, that results from reading written narrative reports, written by an expert test interpreter – how the readers' evaluation correlates with a work criterion such as proficiency or tenure would be another example of this form of validity evidence. In practice, many interpretations have to result in dichotomous decisions: to decide to advance an individual to the next selection stage; to make an offer to hire; or to advise on a particular development action, require the interpreter (or mechanical decision rule) to make a binary decision (McGrath, 2001, 2008). And where the concern is for understanding the criterion-related validity of the decision, it is this dichotomous prediction rather than the test score which should be related to the criterion.

From a user validity perspective rather than comparing two predictors, the comparison is for the two interpretations from the two predictor outputs (which may be as simple as two single standard test scores from two different tests!). From a content perspective, this would be subject matter experts evaluating which of the two test outputs (and test content) best reflect the content being targeted. Fundamentally, the desire is often to forecast a criterion or criteria. And the question is which of the two predictor outputs provide an interpretation which best measures the criterion? And, additionally, do they work better in combination – i.e. do they provide incremental validity. If the user validity is ostensibly equivalent for the two outputs (content overlap/strength of criterion relationship) other contextual variables (e.g.

cost) are likely to play a more significant role in assessing which output (and therefore predictor) to use.

Criterion-related validity evidence is not a direct measure of validity with respect to interpretation, unless it correlates the result of interpreting the test score (e.g. a users' evaluation) rather than a test score with a criterion outcome. Criterion-related evidence can provide a direct form of *user received validity evidence* only where the interpretations the human user makes based on test scores, reports written by experts, spread-sheets of multiple candidates scores or from human interpretation of CBTI are related to the criterion. In general, criterion-related validity evidence provides a metric to calculate and compare validities from different measures against the same criterion. It allows for the calculation of estimates of utility (Cronbach & Gleser, 1965) and allows for an evaluation to be made of a test's fairness in different groups (e.g. Cleary, 1968). Criterion-related evidence generally reflects the relationship between the test score and a criterion rather than a user outcome centred evaluation or forecast with a criterion. Criterion-related validation allows for the comparison of the relative validities of different methods of test interpretation from different outputs through meta-analyses and when appropriate co-validation studies are conducted. Criterion-related validity also allows for the modelling of how to make more validity available to test users (e.g. Vrieze & Grove 2009; Goldberg 1970).

In practice, criterion-related validation requires large samples to complete the predictor instruments, matched to the capability to establish independent criterion ratings and there is a danger of capitalising on sample specific effects and error effects which make cross validation in another sample desirable and the question of generalization of findings less than straightforward. The choice or development of a criterion, normally involves value judgments to be made by the validator (Cronbach, 1971). Criterion-related validation rests, then, on the adequacy and appropriateness of the criteria (Jenkins, 1946). To this end,

criterion-related evidence will to some extent rely on content judgements about the appropriateness of the criterion. The lack of the development of the criterion space (Landy, 2007) could provide a limitation with the argument that there is a lack of adequate criteria available to validate a particular interpretation from a test output. The argument may be made by the test developer or test user that there is no, one, appropriate criterion for a particular test output to be validated against. If an appropriate criterion is not available – then the developer of the test or test output or a proposer of a new use for a test could be seen as having a responsibility to devise an appropriate criterion to support the proposed interpretations of the test. For a specific interpretation of a test score for a particular purpose, it is argued in this paper that it is likely that an appropriate criteria can and should be devised which reflects the interpretation made in the test output. Thus, if a narrative statement in a personality report output makes a claim that an individual is likely to be a faster decision maker than most others, then this can be subject to test by independent raters evaluating a criterion of how fast a decision maker this and other individuals are.

Studies which show ecological validity, in accurately reflecting the environment where tests are actively used, can be difficult to conduct or often suffer from methodological weaknesses which are difficult to avoid. For example, test interpretations can lead to decisions which result in severe restriction of range on the predictor variables. Not using the scores for prediction can raise practical concerns about wasting time, money and not realising the benefits that might accrue from testing. Similarly, if you are not using the test for decision making there may be ethical obligations to inform candidates of this fact and this could impact on the candidates' motivation and therefore, for example, scores on a cognitive test. In practice, there is a need for studies which balance methodological rigour, while investigating the effectiveness of different interpretations based on test scores. This is part of a user validity perspective on criterion –related validity.

Construct related evidence relevant to interpretation

The user validity perspective contends that validity (and by direct implication construct validity) should not be focused primarily on the validity of test scores, but on the validity of the test interpretations.

The argument based approach (Kane, 2006; 2009) provides a logical extension of the work of authors such as Messick (1989), Cronbach (1988), and Shepard (1993) in providing support for interpretations. The argument based approach attempts to strike a balance between validity theory and the requirement to make a judgment about whether a test in use for a particular purpose is appropriate and defensible (Sireci, 2009). Shepard's focus (1993) on the most important evidence for a purpose and Kane's argument based approach allow for focused enquiry and a move away from a tick box exercise of establishing different sources of construct evidence which are generally expected to be collected for a given test (e.g. factor analyses, correlations between different measures). High priority should be given to major and likely intended consequences and plausible unintended consequences of test use (Shepard, 1993). The perspective that validity is an argument that will be the source of debate between different protagonists (Cronbach, 1988) does not preclude validity from being judged with a reasoned, logical approach. Claims about a test or assessment's validity (or rather the validity of claimed interpretations) can be assessed systematically (Kane, 2009). Toulmin (1958) has set out a structured approach to dealing with arguments. In focusing on the justificatory function of arguing, Toulmin (1958) set out that for a good argument to succeed, it has to give a strong justification for a particular *claim*. A proposed interpretation for a particular test will be making a *claim* and the *claim* will need to be supported with appropriate *grounds* in the form of evidence or data. A *warrant* provides rules of inference

(it could be a regression formula, an experts' judgement linking predictors or criteria or a mechanical algorithm based on a regression equation or a fast and frugal heuristic). The warrant gains *backing* from the evidence. The argument based approach articulated by Kane following Toulmin is that a well-directed case can be made that requires an interpretative argument for a particular use/interpretation of a test, the claim which results can be supported more or less well by the evidence.

There has in recent years been some criticism of the mainstream conception of validity (Borsboom, Mellenburgh, & van Heerden, 2004; Borsboom, Cramer, Kievit, Scholten, & Franic, 2009; Lissitz & Samuelsen, 2007). Borsboom *et al*. (2004; 2009) have called for a fundamental reconceptualization of validity. They argue for the adoption of trait (construct) interpretations with a strong causal model that indicates how the trait causes the observed performances or behavior as the standard model of validity. Lissitz & Samuelsen (2007) also focus on a conception of validity centred on content-related evidence and reliability, with considerations of other evidence being external to the validity argument. Critically, Borsboom *et al.* (2004; 2009) and Lissitz and Samuelsen (2007) both raise the issue of whether validity is a property of the test rather than test use and score interpretation. Both Borsboom *et al.* (2004) and Lissitz and Samuelsen (2007) reject construct validity: particularly, the complex conceptions such as nomological nets (Cronbach & Meehl, 1955) and they argue that several forms of evidence considered under the heading of validity should be considered under other headings (e.g. utility), while recognizing these issues have importance. However, the rejection of current conceptions of validity would seriously change the direction of travel in defining validity and would not free test users from the responsibility of providing a reasonable basis for test based interpretations and decisions (Kane, 2009).

Thus, Borsboom *et al.* (2004) provide a rationale for viewing validity from a causal perspective, and this can be viewed as relatively separate and independent from validity concerns focused on the effectiveness and accuracy of interpretations from outputs and also the utility that flows from these.

Furthermore, there is the ongoing debate about the need for construct validity in support of a justification for the use of tests in applied use by I/O Psychologists in selection (Kehoe, 2012; Ployhart, 2012; Sackett, 2012; Schmidt, 2012a; Schmidt, 2012b). The need for construct validity to support measures in use is clearer where a theory or theories are being tested. However, to require complex construct validity evidence when, for example a cognitive test is being used for selection is much more questionable. The imposition of a requirement for construct validity across different occupational measures would impose a high hurdle that would exclude many assessments to I/O users that have demonstrated good criterion-related validity (Schmidt, 2012b). The requirement that test outputs provide evidence in support of their user validity is more achievable.

Both the arguments for stronger causal models for validity and the lack of need for construct validity to justify test use creates an argument for the delineation of user validity, as validity that relates to the accuracy of interpretation and the effectiveness of the test interpretation which directly underpins the utility of the test in use. While construct validity is not a direct form of user validity evidence (in the sense that criterion and content evidence can be used to directly justify test interpretations) that is not to say that certain forms of evidence that are generally considered to fall under the banner of construct validity are not useful in terms of an overall evaluative supporting argument for user validity. User validity can seek to establish sources of construct irrelevant variance which impact on the validity of interpretation and certain forms of construct validity evidence are of importance in developing better user validity. For example, the test developer's understanding the criterion

construct space (e.g. through factor analysis). A predictor structure well aligned to a criterion structure can lead to test outputs which are at the appropriate level of granularity and are transparently aligned with the criteria they are attempting to forecast. Better alignment between the criteria and the predictors is expected to result in an increase in the user validity of interpretation.

However, there are other forms of evidence that are generally less useful and should be viewed more sceptically. For example, indicating that scores on a report or on a new predictor (scale B) correlates with a score on another predictor (scale A) from a different established assessment only provides indirect support for a user interpretation from the test output from B. The new test output seeks to 'assume' the criterion-related validity evidence demonstrated by another test/test output as an argument for the user validity of the interpretation from B. From a user validity perspective this should be seen as weak evidence in and of its' own right and this provides inadequate support for most interpretations.

The user validity perspective advocated in this paper seeks to take a pragmatic stance as to which method of interpretation should be implemented given the purpose/aims of the test, the structure of the data/environment and the context in which the test is used. There needs to be a shift in focus to interpretations and outputs by reconstituting our notion away from construct validity to providing an argument based approach to support, justify and criticize alternative test interpretations and test outputs which support these interpretations. Separately, theory can advance to understand the scientific basis of the particular constructs and where and when it is appropriate it can inform user validity.

Discussion and Conclusion

This paper proposes the concept of user validity that provides a new conception of validity relating directly to the interpretations made from test scores and test outputs. It provides greater clarity on how aspects of validity such as content, criterion-related and construct relate to user validity, and separates out the notion of the validity that is made available to a test user in test scores and outputs (user available validity) from the validity of the interpretations a user makes (user received validity). Secondly, it provides a framework for the validity of test interpretations with regard to test outputs (including test scores). This framework is designed to help in the evaluation, development and research of interpretations which are optimal, appropriate and acceptable which it is argued will have implications for the design of appropriate test outputs to support valid test interpretations.

The user validity perspective should also provide a focus to investigate the different factors that impact on how test scores are interpreted in situ. Situational factors can have an effect on user validity. User validity can be impacted by sampling error, for example. This is likely to be a particular problem when there are very few applicants being considered for a single vacancy. In this case, the quality of the applicants in an applicant pool could be very poor, for example, but where there is a strong pressure to make an appointment, data may be re-weighted and expected standards reduced by the user in favour of selecting a candidate. A related effect may occur where two exceptional candidates are available from a pool of a handful of candidates, but only one is taken as there is only one vacancy. The next time that a vacancy occurs, in that post, the handful of applicants assessed may be much weaker. Summing across a couple of years of such a regularly recurring single vacancy post will clearly serve to weaken the user validity (in comparison to the criterion-related validity that would be expected - having selected for all the vacancies once from a larger applicant pool).

However, general, situational and individual effects all need to be considered from a user validity perspective. In financial investment decisions, for example, potential losses tend to be weighted more strongly than potential gains (Kahneman & Tversky, 1984). By extension, it might be expected that such loss aversion will generalise to test outputs. Interpretations associated with direct negative outcomes are likely to be given more weight by the user than indicators associated with positive outcomes. Where the negatively associated indicators have less (or even no) validity in comparison to the positively associated indicators this will lead to an overall loss in user validity.

A situational context, where the risk from negative consequences is perceived to be particularly high, could exacerbate the effect of overweighting variables which are assumed to have negative consequences in interpretation. It may also be expected to impact on the level of cut score that a test user may consider appropriate (i.e. increase the cut score). The test user's own personality may manifest itself with a preference to accept risk as opposed to being risk averse. This individual characteristic is hypothesised to impact on how a user weights the indicators associated with positive and negative consequences. Thus, the general phenomenon of loss aversion, the user's perceived risk of potential negative consequences, combined with a test user who is more risk averse, should all tilt the calibration of user validity towards the indicators which point to negative consequences.

Where the evidence of poor performance is clearly visible, the perception of risk by the user is likely to be greater. So where there are examples of errors leading to catastrophic consequences, or the potential loss of a large investment required to train an individual (e.g. when they leave straight after the period of training) this will inevitably serve to make the perception of risk by the user higher.

Another related line of investigation within the user validity perspective, is a consideration of the impact a users' personality can have on what they value when they aggregate data to make a decision. The degree to which a test user tends to value the importance of certain attributes they themselves possess in others (i.e. similar to me effect) may have relevance, for example. While, it is unlikely that many users are generally looking for a 'copycat' clone of themselves – they may be attracted to candidates that share some similarities to their own profile of scores or share attributes in common such as similar education or skills (Bagues & Perez-Villadoniga, 2012). Counter-intuitively, the effect of this in certain circumstances could be argued to enhance the validity of interpretation rather than reduce it (Sears & Rowe, 2003).

Indeed, there may be individual differences in the degree to which people recruit in their own image or alternatively exercise their judgement and valiantly attempt to judge which attributes are important for a particular job role (free from any contamination from their own personal attributes and values). In other circumstances, a test user's own personality and values could be hypothesised to lead to them favouring candidates that are quite different to themselves. For example, a dominant leader recruiting a member of staff may value submission (or a lack of dominance) in the subordinates they recruit (rather than favouring highly dominant people like themselves). The effect on user validity in such situations may have some degree of complexity, but the extent of these and other effects are nevertheless potentially discernible.

Test validity as a concept must serve many masters. On the one extreme, for example, there is the etymological drive to understand the causal underpinnings of a particular construct through the scientific method as outlined by Borsboom *et al.* (2004). On another extreme, there is a desire to investigate the validity of an assessment method in practice which will result in the most effective outcome (and is acceptable and usable). Both of these extremes

are important. However, from the perspective of the impact of the test in use, it is the latter which is the more relevant.

It is only the test outputs' interpretation that is the focus of enquiry for user validity. The two key forms of user validity evidence are content and criterion-related validity. A concern for the accuracy of interpretation in user validity based on content validity of the test output is not directly related to consequential validity. And while the effectiveness of test interpretation enshrined in user validity can have consequences in the same way as criterion-related validity (convergent and discriminant) can lead to consequences (from decisions that stem from the interpretation of test scores/outputs), neither form of validity evidence is directly synonymous with the broad, more amorphous conception of consequential validity.

Although, consequences are important in testing, whether consequential validity evidence should be considered a direct part of the validity concept and standard validation practice is open to debate (Cizek, Rosenberg & Koons, 2008; Cizek, Bowen & Church, 2010; Mehrens, 1997; Popham, 1997). Certainly, consequential validity evidence has not made it into the mainstream validation practice of test developers (Cizek *et al.*, 2008; Cizek *et al*.. 2010). An important distinction in relation to the practical benefit of testing is between efficacy and effectiveness research. Efficacy research investigates whether a particular measure can work e.g. as it generally correlates with the criterion of interest. Effectiveness research, by contrast, focuses on whether the use of the tool in the real world, leads to improved outcomes (e.g. elevation of graduates on a criterion of work effectiveness following the use of the test in selection on a criterion could provide an indicator of the effectiveness of a tool's interpretation in practice).

McGrath (2001) has related efficacy and effectiveness research to the distinction between validity research and utility research in assessment. McGrath (2001) called for more relevant

research which focuses on utility concerns and therefore for the need for some research to be more ecologically valid in representing the applied use of the test in real settings. This, for example, includes the need to consider the relevance of the population studied in the research, the fact that variables which are categorical rather than continuous are more representative of decision making in practice, and the need to address clinical interpretation which reflects the actual use of assessments in many contexts. Another consideration identified by McGrath (2001) was the importance of incremental validity of one test or assessment method in comparison to other alternatives that are used for the same purpose. These are important considerations from the perspective of test utility.

Estimating the utility of tests includes evaluating economic factors (e.g. the standard deviation of performance in financial units such as dollars) and is also related to factors such as base rates where they are applicable. Utility then is separate from but strongly driven by the validity of the test. The validity of tests, in practice, that underpin the effectiveness and utility of test score interpretations are best seen as an individual aspect of validity, and this paper argues that greater clarity and transparency is provided to the concept of validity through delineating the concept of user validity. User validity is a direct attempt to investigate and evaluate the validity of the test scores and other test outputs in use (i.e. their effectiveness when they are being interpreted to make decisions) that directly underpins the utility of the test in situ.

The acceptability and faith that people put into different methods is also relevant to the interpretation of test scores in practice. While user validity is not a perceived form of validity such as perceived predictive validity (Smither, Reilly, Millsap, Pearlman, & Stoffey, 1993), user validity is likely to be impacted by users perceptions about the job relatedness and predictive validity of different tests/test outputs: an individual interpreter is likely to take less account of information that he or she perceives to be from a weaker (less valid) method of

forecasting performance than from one he or she puts a great deal of faith in. It is argued in this paper, that both the incremental validity and acceptability concerns should also routinely be extended to different methods of interpretation from test scores. For evidenced based practice, validity research must move from a focus on efficacy to a greater consideration on effectiveness in practice. User validity can support very finely grained interpretations through to very coarse dichotomous decisions. We can have experimental designs to investigate likely influences on user validity and field work which looks at the entire selection system.

While definitions of validity centre on interpretations and inferences from test scores, validity practice is often centred on individual test scores themselves. In practice, this means that what is being validated (e.g. test scores) may not in many circumstances reflect the interpretations made from the test scores and therefore the actual operational validity that resulted from the use of an assessment in a particular sample. Without such a user validity perspective, there is a danger that test scores and test outputs have too limited a justification e.g. on the basis of the criterion-related validity of individual test scores (i.e. efficacy). Anywhere, test scores are provided by a test publisher or a test output designer directly to users (e.g. in a test score profile) will require the user to make some interpretations and/or evaluations of the information and these interpretations should be justified. The fallacy of "begging the question" is when one is asked to accept a conclusion (or claim) without critical examination (Kane, 2006). Begging the question can occur, for example, if a relatively modest interpretative argument is assumed for the purposes of validation, but more ambitious interpretations are made in practice. For example, when narratives in a test output go beyond a simple summary of the content of the test scale (or are based on an interpretation based on a certain configuration of scores) then evidence needs to be provided which goes beyond content validity (Messick, 1989). Other forms of validity evidence may be used to justify the validity of the test score, but do they support the validity of the far more ambitious

interpretation being made in the narrative statement in the test report? Another example of the fallacy in action, would be users' interpreting an output to make an overall evaluation. In this case providing justification based on individual test scores criterion-related validity with different individual criteria would be inadequate as it does not deal with how the overall evaluation the user make impacts on the criterion or criteria. In each case, the justification provides evidence for a different (and less ambitious) interpretation, than is being made and the evidence provided does not provide a direct justification for the claim. User validity reminds us that certain analyses such as regression, multiple or canonical regressions, or results of meta-analyses performed on a set of predictor test scores, do not provide an accurate basis for the operational validity for interpretations made from generic test outputs. When it comes to the many profiles of scores and outputs available for different purposes, this paper argues that 'begging the question' is relatively commonplace.

Furthermore, validity from the argument based perspective provides a basis for making justificatory arguments (Sireci, 2009; Kane, 2009) for test interpretations and these, it is argued, should be more routinely focused on the validity of the interpretation that results from the use of test outputs by test users. The user validity perspective emphasises that test scores are in practice interpreted and that these interpretations and methods of interpretation require justification. The user validity perspective then seeks to take a rational approach to maximising validity given the testing aims, the data environment and the context of testing. Taking a user validity perspective has implications for what is presented in test outputs. The output report provides the basis for user interpretations and therefore it should accurately present and not misrepresent the content of the test or make ambitious claims or predictions which have little justification and leave questions begged. The layout of a report should seek to emphasise relevant content and where applicable make explicit key variables such as criterion forecasts which are directly related to the testing/outputs aims.

In terms of tests and assessments themselves there may be implications too. If there is either a degree of redundancy or invalidity for particular predictor scales when used for interpretation, then reference to any such variables are better removed from the test output. Where this applies across outputs related to the assessments testing aims, this could make the argument for amending or removing the redundant variables from the assessment itself. Such redundancy is likely to become more apparent with the focus on user validity as interpretations which overall are the most effective, efficient and acceptable in certain circumstances may be the Best Single Predictor or Take the Best which may require one or a small number of variables.

It is also logical that where a particular variable has prominence in interpretation due to its high validity for decision making e.g. conscientiousness (Barrick & Mount, 1991), the user validity perspective would argue that a particular focus should be put in test development to improving/maximising the criterion-related (available criterion-related) validity of such key variables.

In the present paper the construct of user validity is proposed as providing an important focus for test designers, users, and psychometricians actively researching test use in practice. It is our hope that this perspective spurs future theoretical contributions and stimulates additional empirical research and development in the practical usage of tests internationally.

References

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

Bagues, M. & Perez-Villadoniga, M. J. (2012). Do recruiters prefer applicants with similar skills? Evidence from a randomized natural experiment, *Journal of Economic Behavior & Organization,* 82, 1, 12-20.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: a meta-analysis, *Personnel Psychology*, 44, 1, 1–26.

Borsboom, D., Mellenburgh, G. J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, 111, 1061-1071.

Borsboom, D., Cramer, A. O. J., Kievet, R.A., Scholten, A. Z., & Franic, S. (2009). The End of Construct Validity. In: R.W. Lissitz (Ed.) *The concept of validity*. Charlotte, NC: Information Age Publishing.

Butcher, J. N. (1995). How to use computer-based reports. In: J. N. Butcher (Ed.), *Clinical personality assessment: Practical approaches*. New York: Oxford University Press.

Butcher, J. N., Perry, J. N., & Atlis, M. M. (2000). Validity and Utility of Computer-Based Test Interpretation, *Psychological Assessment*, 12, 1, 6-18.

Cizek, G. J., Bowen, D., & 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., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, 68, 397-412.

Cleary, T. A. (1968). Test bias: prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement*, 5, 2, 115–124.

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

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

Cronbach, L. J., & Gleser G. C. (1965). *Psychological tests and personnel decisions*. (2nd Ed.). Urbana: University of Illinois.

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

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, 81, 95-106.

Dickson, D. H., & Kelly, I. W. (1985). The "Barnum effect" in personality assessment: A review of the literature. *Psychological Reports*, 57, 367-382.

Garb, H. N. (2000). Introduction to the Special Section on the Use of Computers for Making Judgments and Decisions, *Psychological Assessment*, 12, 1, 3-5.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality, *Psychological Review*, 103, 650-669.

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple Heuristics That Make Us Smart*. New York: Oxford University Press.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and Biases: the Psychology of Intuitive Judgment*. Cambridge: Cambridge University Press.

Goldberg, L. R. (1970) Man versus model of man: a rationale plus evidence for a method of improving on clinical inferences, *Psychological Bulletin*, 73, 422 -432.

Grove, W. M., Zald, D. H., Lebow, B., Snitz, E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis, *Psychological Assessment*, 12, 19–30.

Guion, R. M. (1980). On Trinitarian conceptions of validity, *Professional Psychology*, 11, 385-398.

Hogarth, R. M. (2012). When simple is hard to accept. In: P. M. Todd, G. Gigerenzer, & The ABC Research Group (Eds.). *Ecological rationality: Intelligence in the World*. Oxford: Oxford University Press.

Hogarth, R .M., & Karelaia, N. (2005). Ignoring information in binary choice with continuous variables: When is less "more"? *Journal of Mathematical Psychology*, 49, 115-124.

Hogarth, R. M., & Karelaia, N. (2006). Take-the-best and other simple strategies: Why and when they work "well" in binary choice. *Theory and Decision Analysis*, 3, 124-144.

Honaker, L. M., & Fowler, R. D. (1990). Computer-assisted psychological assessment. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (2nd Edition). New York: Pergamon Press.

Hoover, D. W., & Snyder, D. K. (1991). Validity of the computerized interpretive report for the Marital Satisfaction Inventory: A customer satisfaction study. *Psychological Assessment*, 3, 213-217.

International Test Commission (2005). International Guidelines on Computer-Based and Internet Delivered Testing. http://www.intestcom.org/guidelines

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

Kahneman, D., & Klein, G. (2009). Conditions of Intuitive Expertise: A failure to disagree. *American Psychologist*, 64, 6, 515–526

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Kahneman, D., & Tversky, A. (1984). Choices, values and frames. *American Psychologist*, 39, 4, 341–350.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies, *Psychological Bulletin*, 134, 404–426.

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

Kane, M. T. (2006). Validation. In: R.L. Brennan (Ed.), *Education Measurement* (4th Ed.).Westport: American Council on Education/Praeger.

Kane, M. T. (2009). Validating the interpretations and uses of test scores. In: R.W. Lissitz (Ed.*). The Concept of Validity*. Charlotte, NC: Information Age Publishing.

Kehoe, J. F. (2012). What to make of content validity evidence for cognitive tests? Comments on Schmidt (2012). *International Journal of Selection and Assessment*, 20, 14–18.

Landy, F. (2007). The validation of personnel decisions in the twenty-first century: Back to the future. S. Morton McPhail (Ed.), *Alternative validation strategies: Developing new and leveraging existing validity evidence* (pp. 409-426). San Francisco, CA: John Wiley and Sons.

Linn, R. L. (1997). Evaluating the Validity of Assessments: The Consequences of Use. *Educational Measurement: Issues and Practice*, 16, 2, 14-16.

Lissitz, R. W. & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36, 437-448.

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

McGrath, R. E. (2001). Toward more clinically relevant assessment research. *Journal of Personality Assessment*, 77, 307–322.

McGrath, R. E. (2008). Predictor combination in binary decision-making situations. *Psychological Assessment, 20,* 195-205.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota.

Mehrens, W. A. (1997). The Consequences of Consequential Validity, *Educational Measurement: Issues and Practice,* 16, 2, 16-18.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In Wainer, H. & Braun, I.H. (Eds.). *Test Validity*. (pp. 33-45). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Messick, S. (1989). "Validity". In R.L. Linn (Ed.), *Educational Measurement* (3rd Ed.). New York: MacMillan.

Popham, W. J. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice,* 16, 2, 9-13.

Ployhart, R. E. (2012). The content validity of cognitively oriented tests. Commentary on Schmidt (2012). *International Journal of Selection and Assessment*, 20, 19–23.

Sackett, P. R. (2012). Cognitive tests, constructs, and content validity. A commentary on Schmidt (2012). *International Journal of Selection and Assessment,* 20, 24–27.

Schmidt, F. L. (2012a). Cognitive tests used in selection can have content validity as well as criterion validity: A broader research review and implications for practice. *International Journal of Selection and Assessment,* 20, 1–13.

Schmidt, F. L. (2012b). Content Validity and Cognitive Tests: Response to Kehoe (2012), Ployhart (2012), and Sackett (2012). *International Journal of Selection and Assessment,* 20, 28–35.

Sears, G. J., & Rowe, P. M. (2003). A personality-based similar-to-me effect in the employment interview: Conscientiousness, affect-versus competence-mediated interpretations, and the role of job relevance. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement,* 35, 1, 13-24.

Shah, A. K. & Oppenheimer, D. M. (2008). Heuristics made easy: An effort reduction framework. *Psychological Bulletin*, 134, 207-222.

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

Simon, H. A. (1956). Rational choice and the structure of environments. *Psychological Review*, 63, 129-138.

Sireci, S. (2009). Packing and Unpacking Sources of Validity Evidence. In: R.W. Lissitz (Ed.) *The Concept of Validity*. Charlotte, NC: Information Age Publishing.

Smither, J. W., Reilly, R. R., Millsap, R. E., Pearlman, K., & Stoffey, R. W. (1993). Applicant reactions to selection procedures. *Personnel Psychology*, 46, 49-76.

Snyder, D. K. (2000). Computer-Assisted Judgment: Defining Strengths and Liabilities. *Psychological Assessment*, 12, 1, 52-60.

Todd, P. M., & Gigerenzer, G. (2012). What is Ecological Rationality? In P. M. Todd, G. Gigerenzer, & The ABC Research Group (Eds.). *Ecological rationality: Intelligence in the World*, 61-79 .Oxford: Oxford University Press.

Toulmin, S. E. (1958). *The Uses of Argument.* Cambridge University Press. New York, USA, 2nd edition 2003.

Vale, C. D., & Keller, L. S. (1987). Developing expert computer systems to interpret psychological tests. In J. N. Butcher (Ed.). *Computerized psychological assessment: A practitioner's guide*, 64-83. New York: Basic Books.

Vrieze, S. I., Grove, W. M. (2009). Survey on the Use of Clinical and Mechanical Prediction Methods in Clinical Psychology. *Professional Psychology: Research and Practice*, 40, 5,525-531.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, 83, 213-217.

Westen, D., & Weinberger, J. (2004). When clinical description becomes statistical prediction. *American Psychologist*, 59, 595– 613.

Zenisky, A. L. & Hambleton, R. K. (2012). Developing Test Score Reports That Work: The Process and Best Practices for Effective Communication. *Educational Measurement: Issues and Practice*, 31, 2, 21–26.

Figure 1: A framework for valid interpretation from test outputs