Measurement: Interdisciplinary Research and Perspectives
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hmes20

Does Theory Drive the Items or Do Items Drive the Theory?

Robert M. Talbot [a] & Derek C. Briggs [a]
[a] University of Colorado , Boulder, Colorado, USA
Published online: 05 Dec 2007.

To cite this article: Robert M. Talbot & Derek C. Briggs (2007) Does Theory Drive the Items or Do Items Drive the Theory?, Measurement: Interdisciplinary Research and Perspectives, 5:2-3, 205-208, DOI: 10.1080/15366360701492906

To link to this article: http://dx.doi.org/10.1080/15366360701492906

Does Theory Drive the Items or Do Items Drive the Theory?

Robert M. Talbot and Derek C. Briggs
University of Colorado, Boulder, Colorado

At the heart of the argument-based approach to test validation as it has been presented by Kane (1992, 2004, 2006) is a relatively simple premise: test validity is demonstrated by linking the score that is observed from a test instrument to the use of that score for some subsequent inference. The devil is in the details. How does one craft a coherent interpretive argument? How does one systematically collect evidence to support that argument? To date there have been no examples of which we are aware of anyone explicitly applying Kane's validation approach to a test they have developed from the ground up. The collection of papers by the University of Michigan group is thus a very welcome development. The application of the argument-based approach to validation presented in these papers is framed by what the authors are introducing as two amendments to Kane's approach: (1) A clearer distinction between assumptions and inferences in the formation and evaluation of the interpretive argument; (2) Breaking up the interpretive argument into what the authors describe as elemental, structural and ecological pieces. Each of these pieces is then associated with specific methods for gathering the evidence needed to support a validity argument.

In the first of their amendments, the authors have defined inferences as the operationalization of specific assumptions. Because an inference can be tested empirically, it can also be falsified as part of an interpretive argument. Interestingly, Schilling, Blunk, & Hill (this issue) write, "discarding unsupported inferences does not necessarily mean discarding the assumptions to which they refer. Rather inferences arising from assumptions can be replaced or modified so they become consistent with both the assumptions and empirical data." One might worry that this is a prescription that advocates fishing for data that support ones assumptions. If assumptions cannot be supported by their inferences, is it not more advisable to more carefully scrutinize one's assumptions? Part of the

Correspondence should be addressed to Derek C. Briggs, University of Colorado, 249 UCB, Boulder, CO 80309. E-mail: derek.briggs@colorado.edu

problem here is that operationalizing an assumption with one or more testable inferences is itself not so straightforward. For example, consider inference D for MKT structural assumption (Schilling and Hill, this issue): "Teachers' reasoning for a particular item will reflect the type of reasoning (either CK or KCS) that the item was designed to reference." Note that in this inference teachers' reasoning is latent and the design of the items is unobserved by all but those who have designed the test. Unless the operationalization of both is very explicit, the evidence for or against this inference will necessarily be ambiguous. Schilling's (this issue) MIRT analysis seems to call this inference into question because some items do not load adequately upon the expected dimensions. But this could occur for two different reasons: (1) the items have been improperly designed, and/or (2) the underlying theory of teachers' reasoning is wrong. If it is the latter, then it is the underlying assumption that must be replaced or modified, if it is the former, it is the inference.

To the extent that it makes the formation of an interpretive argument more manageable and coherent, the notion of distinguishing elementary, structural and ecological assumptions and the methods that can be used to validate them might be viewed as a positive development. On the other hand, a strength of Kane's argument-based approach is that he avoids the use of terms such as content, criterion and construct validity, terms that bring with them a history of misuse and misunderstanding. Introducing new terminology has the potential of creating more confusion then it alleviates. The bigger problem is that the University of Michigan group appears to view a chain of reasoning in which elemental assumptions and inferences precede and inform structural assumptions and inferences. In our view it should be the other way around: the assumptions about the phenomenon to be measured should be the basis for the design of test items.

The notion of the "Assessment Triangle" from *Knowing What Students Know* (Pellegrino et al., 2001) and Wilson's construct mapping approach (Wilson, 2005) would seem to suggest that validation arguments are best formed and strengthened on the basis of instrument designs that are driven by theory. Yet the theory behind the domain of MKT—how teachers develop MKT, and how this fits into broader conceptions of teacher knowledge—is strangely absent from the collection of papers by the University of Michigan group. In their discussion of measure development in the first paper, Schilling and Hill write, "because no map of 'mathematical knowledge for teaching' existed in 2000, and anticipating such a map would take years to build, the authors declined to criterion reference the instruments." While we also recognize that developing these sorts of "maps" for the construct to be measured is quite hard, in their absence the structural and elemental assumptions of any interpretive argument lack a solid foundation. Inferences for such assumptions can still be evaluated, but interpretations will always be equivocal—is the problem with the inference, or is the problem with

the assumption? When items and their responses are not carefully linked to a theory that informs their design, the resulting validity argument is necessarily weakened.

The design of the MKT measures seems to be very much item–driven rather than theory-driven. The researchers have made the decision a priori to use a multiple–choice item format because of the need to administer the instrument to a large number of teachers. Under a theory-driven approach to development, one would be more likely to begin by developing open-ended items as a means of discovering the universe of possible responses to the items. This would both inform the developing theory of what it means for a teacher to have more or less MKT, and help in the development and subsequent interpretation of high quality multiple-choice items. An explicitly stated theory about MKT development can help in making sense of the empirical difficulty of the items used in the instrument. Under an item-driven approach if an item is found to be very hard or very easy to answer correctly, or if the item does not support the essential unidimensionality of a subscale, one might choose to eliminate the item from the instrument because it does not fit on statistical grounds. But under a theory-driven approach this is precisely the item of greatest interest. Was the item designed to be hard or easy? Was it designed to measure one or two dimensions? What do we observe, and how does this relate to what we expect?[1]

To be more concrete in making this point, consider the two example items provided by the University of Michigan group in their appendix. Now, had a hypothesized theory of the MKT domain (or of MKT development) been made explicit, on this basis we might be able to predict that item 2 will be more difficult than item 1 for two reasons: (1) the content in item 2 (fractions) is more difficult to understand conceptually; (2) item 2 deals with representing this more difficult content to the students in alternative forms. Therefore, it may be reasonable to suspect that a respondent would need to have both pedagogical knowledge and content knowledge to respond correctly to item 2, but only content knowledge to respond correctly to item 1. This might serve to make item 2 harder than item 1, or prompt a related hypothesis about item dimensionality. One would hope that any subsequently specified measurement model would fail to reject these types of hypotheses. When the model rejects the hypothesis, one would expect an explanation for why this has occurred that is based on substantive knowledge of the phenomenon being measured.

The University of Michigan group views the lack of distinction between assumptions and inferences, and the lack of prescriptiveness in the formation and evaluation of interpretive arguments as the primary short-comings of Kane's

[1]Note that this is exactly the approach of Hill, Ball, Blunk, Goffney, and Rowan (this issue) in testing their ecological inference (3A) that "higher scores on the scale derived from our measures are positively related to higher-quality mathematics instruction" by using Spearman's Rho to correlate the rankings of teachers' MKT from their instrument and classroom observations.

argument-based approach to validity. In our view, both of these perceived shortcomings are symptoms of a deeper problem: the apparent disconnect between test instrument design and validation.

REFERENCES

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

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement*, *2*(3), 135.

Kane, M. T. (2006). Validation. In National Council on Measurement in Education, & R. L. Brennan (Eds.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.

Pellegrino, J. W., Chudowsky, N., Glaser, R., & National Research Council Division of Behavioral and Social Sciences and Education Committee on the Foundations of Assessment. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, N.J.: Lawrence Erlbaum Associates.