MEASUREMENT: INTERDISCIPLINARY RESEARCH AND PERSPECTIVES
2017, VOL. 15, NO. 2, 70–72
https://doi.org/10.1080/15366367.2017.1367245

Routledge
Taylor & Francis Group

How Can We Improve the Measurement of "Non-Academic" Competencies?

Peter F. Halpin

Department of Applied Statistics, Social Science, and Humanities, New York University

The target paper raises some interesting critical ideas, both old and new, about the validation of self-report surveys. As indicated by Dr. Maul, recent policy initiatives in the United States (e.g., ESSA) have led to a demand for assessments of "non-academic" skills that are suitable for applied educational contexts. It seems inevitable that self-report surveys will play a role in meeting this demand, because of their ubiquity and the relatively low bar that has been established for their validation in social science research. It is therefore a good time to take a hard look at self-reports and their validation methodology, and I would like to commend Dr. Maul for inviting us to do so.

I personally found the empirical studies to be the most interesting part of the paper. Although the examples were contrived, the results were nonetheless surprising and were successful in casting doubt (even more doubt) upon standard research practices associated with self-report surveys. However, I do not agree with Dr. Maul's interpretation of the implications of the empirical studies. He states that, because the survey items have been deliberately constructed to be meaningless, it is "reasonable to expect that results from the application of validation procedures should provide a clear falsification of the hypothesis that these items constitute a valid measure of anything at all" (p. 3). I interpret this in terms of the conventional (Popperian) falsification argument based on modus tollens, in which a scientific claim is the antecedent and an empirically observable implication of that claim is the consequent. In the present context, the conditional statement at the core of the falsification argument is something like the following.

Proposition 1: If the measure is valid, then the traditional validation methods should "work out." (If P then Q).

In Popperian-style falsification, Q is shown to be false and we conclude that P must also be false. Under this interpretation, a "failure of falsification" might occur if Q were shown to be false but P was nonetheless true. In terms of Proposition 1, this would happen if the validation methods led us to reject a measure that was indeed valid. It is not apparent to me what else a failure of falsification could connote in this context, since, as the reader will recall from elementary truth tables, a conditional statement does not provide any conclusion about the truth value of the antecedent when the consequent is true.

Dr. Maul asks us to consider a different form of argument. We start by assuming that P is false (i.e., assuming the measure is invalid) and are told to expect that Q is also false (i.e., the validation methods should not work out). However, this form of argument is a well-known fallacy, denying the antecedent. In other words, there is no logical contradiction in finding that the traditional validation methods do not lead to a falsification of his nonsense measures. For this reason, I must disagree with his conclusion, "If ever there were a time when a theory deserved to be falsified, this would appear to be it" (p. 7).

The foregoing analysis of the logic underlying the empirical studies depends on the conditional statement in Proposition 1. In this statement, the traditional validation methodology working out is a necessary condition for a measure to be valid. On the other hand, Dr. Maul's discussion seems to be based on the interpretation that the methodology is sufficient for a measure to be valid. As a criticism of the "extra-scientific factors" (p. 2) that support the uncritical application of these methods, this interpretation is certainly justifiable. However, under this interpretation, the examples serve merely as

CONTACT Peter F. Halpin peter.halpin@nyu.edu 246 Greene Street, Office 204, New York, NY 10003-6677.

a parody of those practices, and they may be readily dismissed as a straw man argument. The uncritical researcher might simply shrug it off and say "garbage in, garbage out" or some similar adage. As noted by Dr. Maul (e.g., p. 8), the more critical researcher will surely reject the premise that the traditional methods are sufficient for establishing validity and might point to argument-based interpretations of validity, the availability of more sophisticated methodology, and so on.

But I think we can learn something more from the empirical studies. I agree with Dr. Maul's interpretation that interindividual variation on the nonsense items must be considered irrelevant to the *mindsets* construct by definition (p. 3). What I found most surprising about the examples was the extent to which this construct-irrelevant variation was associated with variation on the genuine mindsets items ($r = .44$, as reported in the target article). In a multitrait, multimethod framework, this relationship would be attributable to methods variance, which, based on the design of the gavagai items, could be interpreted wholly in terms of response bias. This point was noted in passing by Dr. Maul (p. 8) before revisiting a number of philosophical arguments about validity and measurement. These arguments remain an interesting and important domain of measurement research, yet they do not seem to offer any concrete solutions to the problems raised by the empirical examples. From a more practical perspective, I think it would also be useful to treat the empirical studies as an opportunity to emphasize the problem of response bias. By offering pragmatic solutions to this problem, we can help applied researchers and policy makers avoid many of the pitfalls of self-report surveys. As discussed above and elsewhere (e.g., Duckworth & Yeager, 2015), these practical considerations are especially relevant given the current policy context concerning the measurement of non-academic attributes in educational settings. So, I'll conclude this commentary by mentioning some practical solutions.

Böckenholt (2017) and Bolt, Lu, and Kim (2014) provide two recent examples of how response bias on self-report surveys can be statistically modeled. These approaches offer a means for keeping the familiar design of Likert-type items but use a more sophisticated methodology for dealing with response bias. They also offer a data-based framework for thinking about how to redesign survey items to minimize construct-irrelevant variation. Alternatively, one may think about changing the response format altogether to mitigate the effects of response bias. This type of approach is exemplified by ipsative, forced-choice items, which have received attention from psychometricians (e.g., Brown & Maydeu-Olivares, 2012; Stark, Chernyshenko, & Drasgow, 2005) and researchers focusing on non-cognitive assessments (e.g., Kyllonen, 2015). As a third approach, we can avoid self-reports altogether by measuring non-academic outcomes using performance-based assessments—that is, assessments that are designed to require respondents to exhibit or utilize the attributes about which the assessor wishes to make an inference. For example, conventional mathematics items are designed to require students to "do" mathematics, whereas a self-report measure of, say, grit, does not require respondents to utilize or exhibit grit. Here I'll point to research on collaborative problem solving as an illustration of how nonacademic skills can be measured in a performance-based context (e.g., the spring 2017 special issue of the *Journal of Educational Measurement*).

These practical solutions to the problem of response bias are intended to be illustrative rather than exhaustive. They each require that we accept Dr. Maul's challenge to move beyond the uncritical application of self-report surveys. This is an important goal that applied researchers, policy makers, psychometricians, and measurement theorists can all work toward together.

References

Böckenholt, U. (2017). Measuring response styles in Likert items. *Psychological Methods*, *22*(1), 69–83.

Bolt, D. M., Lu, Y., & Kim, J. (2014). Measurement and control of response styles using anchoring vignettes: A model-based approach. *Psychological Methods*, *19*(4), 528–541.

Brown, A., & Maydeu-Olivares, A. (2012). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, *18*(1), 36–52.

Duckworth, A. L., & Yeager, D. S. (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*(4), 237–251.

Kyllonen, P. C. (2015). Designing tests to measure personal attributes and noncognitive skills. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 190–211). New York, NY: Routledge, Taylor and Francis.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise preference model. *Applied Psychological Measurement*, *29*, 184–201.