Routledge
Taylor & Francis Group

Measurement: Interdisciplinary Research and Perspectives

ISSN: 1536-6367 (Print) 1536-6359 (Online) Journal homepage: http://www.tandfonline.com/loi/hmes20

How to Measure Nothing

Mijke Rhemtulla, Denny Borsboom & Riet van Bork

To cite this article: Mijke Rhemtulla, Denny Borsboom & Riet van Bork (2017) How to Measure Nothing, Measurement: Interdisciplinary Research and Perspectives, 15:2, 95-97, DOI: 10.1080/15366367.2017.1369785

To link to this article: https://doi.org/10.1080/15366367.2017.1369785

MEASUREMENT: INTERDISCIPLINARY RESEARCH AND PERSPECTIVES
2017, VOL. 15, NO. 2, 95–97
https://doi.org/10.1080/15366367.2017.1369785

Routledge
Taylor & Francis Group

How to Measure Nothing

Mijke Rhemtulla[a], Denny Borsboom[b], and Riet van Bork[b]

[a]Department of Psychology, University of California, Davis; [b]Department of Psychological Methods, University of Amsterdam

Maul's target article is insightful and enlightening, and it presents a number of very important recommendations for psychometric practice. While we overwhelmingly agree with the arguments made in the second half of the paper, however, we are not convinced that the results presented in the first half of the paper are really damning evidence against the adequacy of psychometric processes. The setup of Maul's research design appears to be based on the idea that, since the word *gavagai* doesn't refer to anything, the psychometric items that query participants' ideas about gavagai will not measure anything more or less by definition. Moreover, Maul appears to think that standing psychometric practices should reveal this: "If ever there were a time when a theory deserved to be falsified, this would appear to be it." From the fact that standard psychometric practices do not reveal any significant problems in the questionnaire, Maul concludes that there must be something deeply wrong with these practices. This conclusion, however, does not follow for two reasons. First, it is not clear that the gavagai questionnaire measures nothing, and hence, it is not obvious that the premise underlying Maul's argumentation is fulfilled. Second, most psychometric practices are based on the antecedent assumption that researchers are able to target a given attribute with a set of items (i.e., most psychometric practice assumes that test constructors have at least to some extent built validity into the test through item formulation and selection); they are not, however, designed to expose the falsity of that assumption. We therefore think it is useful to consider more deeply what these results reveal about response processes and validity.

Maul interprets his results as showing that it is so easy to get a set of well-behaving items that it can be done even without item content. In fact, he suggests that "favorable-looking results of covariance-based statistical procedures (such as high-reliability estimates and fit to unidimensional latent variable models) should be regarded more as a default expectation for survey response data than as positive evidence for the validity of an instrument as a measure of a psychological attribute." But experience tells us that's not right: Researchers typically have to try hard (pilot testing, cutting items, rewording items, etc.) to get item sets that behave well, and many scales that are proposed in the literature turn out to be best described by a multidimensional factor model. It is not at all a given that any random set of items subjected to a factor analysis will result in a well-fitting unidimensional model with high factor loadings.

So there are actually quite important questions here that Maul scarcely addresses—namely, Why do these nonsense items behave so well? Why are the responses so structured? What response process was tapped by the seemingly nonsensical items? There is a wealth of research showing that respondents in psychological research are willing to work with the investigator and behave according to the demand characteristics of the experimental design. What were the Mechanical Turk workers thinking as they filled out these nonsense items? What did they hypothesize that Andrew Maul wanted to know about them?

We do not know the answers to these questions, and as a result it is far from easy to determine what Maul's results really entail. It *may* be the case that the results show that standing psychometric practices are deeply deficient. For instance, participants might offer a truly random answer to the first question and stick with it for the following nine questions out of a desire to act consistently. This strategy is consistent with Maul's interpretation that "at least in the context of responding to

CONTACT Mijke Rhemtulla mrhemtulla@ucdavis.edu 1 Shields Avenue, Davis CA 95616.

survey questions, respondents often choose to behave consistently unless there is a clear reason not to do so." But it may also be the case that, unbeknownst to Maul, the items tapped meaningful response processes that result in relatively homogeneous response behavior across items. For instance, if every person fills in the concept gavagai with a personality attribute of their own choice, then it may well be that the responses will reveal the systematicity that Maul finds in his data. We think it is worth considering Cattell's "bloated specific"—that is, the idea that when a set of items are all essentially interchangeable versions of the same question, the resulting scale will be highly reliable (but have questionable validity) because similar response processes result in similar responses to interchangeable questions (1978). Factor analyses and reliability coefficients are entirely determined by the pattern of correlations among items. As such, whenever these procedures are applied to a set of interchangeable items, their results will look excellent, no matter how individuals interpret them. Now, Maul's nonsense scales involve items that are about as perfectly exchangeable as items can get (in the case of Experiment 3, the items are literally identical).

If individual differences in response processes result in meaningful between-person variability in the data, does it entail that psychometric practices should have identified a flaw? Perhaps it is not surprising at all that a set of interchangeable nonsense items produces a high alpha value and that a one-factor model fits. In fact, it seems to us that this type of response process may actually be accounted for in terms of a traditional psychometric latent variable (e.g., to what extent person *i* believes that the concept he or she filled in for *gavagai* is malleable, etc.), even if this latent variable is an untraceable amalgam of idiosyncratic interpretations. Psychometric theory, after all, is based on the assumption that individual differences on *some* dimension determine individual differences in responses to items. This assumption does not require that the dimension in question is sensible, intelligible, interesting, or substantively meaningful.

One approach to understanding participants' response processes is to consider in which ways items with meaningful content differ from those without it. In this respect, it is precisely the fact that the items relate to a meaningless notion that undermines Maul's conclusions. We would find it much more damning (and truly surprising) if a set of *meaningfully different* items—for example, a set of items randomly culled from existing scales—were to produce similarly impressive results. However, randomly chosen self-report measures would be unlikely to behave so well, because meaningful differences in item content would lead participants to give meaningfully different responses across items. In this sense, it is exactly the fact that the gavagai items do *not* have a clear semantic target (i.e., are idiosyncratically but consistently understood in different ways by different participants) that leads them to align empirically with the psychometric norms of reliability and unidimensionality.

This brings us to the concept of validity. Maul's declaration that "if ever there were a time when a theory deserved to be falsified, this would appear to be it" demands that we consider precisely what validation procedures are supposed to falsify. If validity means that variation in some psychological attribute leads to variation in item responses, then the hypothesis that "the questionnaire is a valid measure of a psychological construct" is true when there is a psychological attribute that causes variation in item responses and false when there is no such attribute. As such, Maul argues that the finding that his nonsense scale scores correlated with *anything* would ordinarily be taken as positive evidence for their validity. But validity may be taken as a stronger hypothesis—namely "that a given attribute has successfully been measured by a given survey instrument," as Maul writes. This hypothesis specifies not only that *some* psychological attribute is causally responsible for variability in item responses but that a *particular* attribute is responsible. Thus, validating the "theory of intelligence" scale should involve an investigation of whether variation in scale scores corresponds to real differences in people's theory of intelligence, and validating the "theory of gavagai" scale should require searching for a correspondence between scores and real differences in people's theory of gavagai. To the extent that no such thing exists, it should prove impossible to find any evidence for the validity of this scale, over and above evidence for its reliability. Maul's results bear out this prediction.

If our current procedures cannot falsify the validity of a measure, the question of how we should go about it is, of course, pressing. In this respect, we think Maul's findings highlight the importance of reaching beyond self-reports, end extending validation research toward the inclusion of behavioral/observational measures. If Maul's results show one thing, it is that self-reports can form a hermeneutical system in which virtually tautological item formulations can generate consistent item response behavior, even if no two individuals understand the items in the same way. As such, we agree with Maul's conclusion that self-report measures may be particularly difficult to validate and with his recommendation that researchers work to explicitly uncover the processes that lead to variation in item responses.

In sum, Maul's attempts to measure nothing raise many questions about response processes, measurement, and validity, and we are grateful for this opportunity to consider them. The outcomes of these surveys and the question of what they mean is highly thought provoking. The challenges to psychometric practices presented in the target article evince substantial gavagai.

References

Cattell (1978). The Scientific Use of Factor Analysis in Behavioral and Life Sciences. New York: Plenum Press.