

**Likert or not, survey (in)validation requires explicit theories and true grit**

Joshua A. McGrane

Oxford University Centre for Educational Assessment (OUCEA), Department of Education,  
University of Oxford

Trisha Nowland

Department of Psychology, Macquarie University

Correspondence concerning this article should be addressed to Joshua McGrane, Oxford  
University Centre for Educational Assessment (OUCEA), Department of Education,  
University of Oxford. E-mail: [joshua.mcgrane@education.ox.ac.uk](mailto:joshua.mcgrane@education.ox.ac.uk).

From the time of Likert (1932) on, attitudes of expediency regarding both theory and methodology have become apparent with reference to survey construction and validation practices. In place of theory and more theoretically-minded methods, such as those found in the early work of Thurstone (1928) and Coombs (1964), statistical models and methodological heuristics have come to dominate, with the result of gross simplification of complex techniques. Maul's focus article lampoons this atheoretical attitude with a rhetorical flair that is all too rare in this literature. More importantly, it draws stark attention to the scientific and ethical dilemmas that this attitude has created as self-report, survey-based instruments have become ubiquitous throughout psychological research, education systems, and modern society more generally.

In this commentary, we emphasise Maul's call for survey validation practitioners (and methodologists in the psychological sciences more generally - including ourselves) to critically reflect on current practices with the same attitude of openness, optimism, and perseverance toward conceptual and methodological growth that we increasingly expect of students in their own cognitive development as 21st Century learners. Such growth necessitates a spirit and process of critical inquiry, in order that survey validation methods find firmer scientific foundation through careful conceptual, logical, and empirical analysis (Petocz & Newbery, 2010). A primary tool of critical inquiry is logical analysis, which may be used to examine the clarity and coherence of current practices and their inherent (and often implicit) assumptions. In the following discussion, we direct such a logical analysis toward the empirical aspects of Maul's article. In doing so, we will somewhat disagree with Maul's empirical means of demonstration to strongly agree with his article's overarching message. No reverse-coding will be necessary.

*Having your completely absent cake and falsifying it too*

In setting forth his discussion of his findings, Maul states, “It would seem difficult to take seriously the claim that any of these sets of items constituted a valid measure of a psychological attribute, and if such a claim were made, one might reasonably expect any quality-control procedure worthy of the name to provide an unequivocal rejection. To state this in Popperian language: If ever there were a time when a theory deserved to be falsified, this would appear to be it” (p.7). While Maul’s argument has a strong rhetorical bite, it is questionable that it stands up to its own falsificationist logical demands.

Popper’s (1959) now ubiquitous theory of scientific demarcation drew upon a classical argument of propositional logic, modus tollens (denying the consequent). Popper pointed out that the following (simplified) scientific reasoning, which included two premises (P) and a conclusion (C), was logically sound:

P1: If Theory P is true [antecedent], then I will observe Q [consequent].

P2: Q is not observed.

C: Therefore, Theory P is false.

Whereas, if the scientist were to observe ‘Q’ and conclude that ‘Theory P is true’ on that basis, they would fall into the logically fallacious - but intuitively alluring - trap of affirming the consequent (‘Q’ may be observed for reasons that are unrelated to the truth of ‘P’, just as smoke may be observed in the absence of fire). Popper’s falsificationism was game-changing in the philosophy of science, as it did not demarcate rational scientific endeavour in terms of successful prediction and confirmation, but rather the capacity to precisely specify potentially falsifying phenomena and observations.

Maul’s discussion utilises this falsificationist logic as a rhetorical device by explicitly drawing the reader’s attention to the non-sequitur conclusion entailed by his (satirising) empirical premises:

P1: If the “Theory of Gavagai” items validly measure a psychological attribute, then I will observe a satisfactory Cronbach’s alpha coefficient and factor analytic solution.

P2: I observed a satisfactory Cronbach’s alpha coefficient and factor analytic solution.

C: Therefore, I cannot falsify that the “Theory of Gavagai” items validly measure a psychological attribute.

The rhetorical weight of Maul’s argument is born out of the fact that, like any good use of dramatic irony, the reader is already in on the ruse; the “Theory of Gavagai” is not a theory of anything and the items are not meant to measure anything. Maul explicitly reminds us of this when he states in the same paragraph, “...items were written in the complete absence of a theory concerning what they measured and how they worked” (p.7). But in this same sense, despite being thoroughly enjoyable, the argument carries no logical weight. This is because the antecedent is (by design) false, which violates the requirement of bivalence (i.e., both antecedent and consequent must be able to be either true or false) in a modus tollens argument, and so the antecedent cannot affirm (or deny) the consequent in any non-vacuous way. Given this breakdown in the propositional logic, Maul’s appeal to falsificationism to support his point does not stand on any logical footing and so may be guilty of invoking another logico-literary device; a red herring.

### *No free (inferential) lunch*

While lacking quite the same rhetorical flourish, Maul’s article initially sets out a similar propositional argument, but one that does not invoke any intentional falsehood in its premises. Given its position in his article, as well as its firmer logical foundation, we will consider the following as the primary argument of the article:

P1: If commonly used survey validation methods are robust quality-control procedures, then these methods will reject a set of nonsensical items.

P2: These methods did not reject this set of nonsensical items.

C: Therefore, commonly used survey validation methods are not robust quality-control procedures.

This line of argumentation seems, at face value, quite reasonable and compelling. However, while the argument may be logically valid, the soundness (truth) of the conclusion is still contingent on the requirement that the antecedent can *only* be true if the consequent is true. Yet, Maul's article intentionally does not offer any theoretical or empirical justification for why this strict relationship holds between the antecedent and consequent in his empirical argument. Instead, the article repeatedly prompts the reader with intuition-piquing reminders that the items are atheoretical/nonsensical/gibberish/absent, and it seems intuitively unlikely that such items should fit any robust psychometric criteria.

Nonetheless, logical analysis pushes us beyond the constraints of intuition to explicitly acknowledge and question unsubstantiated assumptions. In the absence of a sound argument otherwise, it is logically possible that survey validation methods are robust quality-control procedures *and* they are insensitive to the nonsensical nature of the items in Maul's empirical context. On this point, as Maul also acknowledges, it has long been recognised in the survey methodology literature that reliability and validity statistics may be artificially, 'positively' inflated by various empirical factors, including response sets, the semantic overlap of items, the structural features of the response scale, impression management strategies, etc., as they tend to induce internally-consistent response patterns across survey items. Similar factors may be contributing to the 'positive' reliability and validity evidence observed in Maul's studies.

Consistent with this suggestion, recent findings on survey research using Mechanical Turk samples, which is the source of the data in Maul's studies, provide evidence that this cohort are attentive responders, but may also be particularly prone to superficial strategies for

responding because of time constraints and the financial incentives of this online platform. Such factors may induce people to ‘speed’ through a survey just to finish it (Hauser & Schwarz, 2016; Wood, Harms, Lowman & DeSimone, 2017). Similarly, Hamby & Taylor (2016) argue that Mechanical Turk samples show a stronger pattern of survey satisficing, where they will adopt a ‘good enough’ rather than an ‘optimising’ response strategy to survey items, which often includes not differentiating between similar items, and therefore leads to inflated reliability and validity indices.

This satisficing strategy shortcuts the cognitively demanding and time-consuming need for a survey respondent to interpret the content and deduce the intent of an item, consider relevant information, integrate the information, and then select the most appropriate response option for that integrated information (Krosnick, 1991). Considering several factors in Maul’s designs, including the use of Mechanical Turk and financial incentive; the forced-choice format; and the fact that the sets of nonsense items had no interpretable, deducible, or integrative meaning, satisficing may have ironically been an optimal response strategy for completing the task. This strategy, in turn, would give the appearance of reliable and valid responding, without necessarily leading to the general conclusion that these statistical processes are not robust.

While the above discussion is admittedly speculative, it is only intended to cast doubt on the intuitive basis for the assumed relationship between antecedent and consequent in Maul’s primary argument. In the absence of such a relationship, the implications of Maul’s empirical findings remain inherently ambiguous and do not substantiate conclusions such as, “favorable-looking results of covariance-based statistical procedures...should be regarded more as a default expectation for survey response data than as positive evidence for the validity of an instrument” (p.8). The point is, whether one is attempting to validate a survey instrument or invalidate the statistical procedures typically used for such purposes, explicit

theory and argument that satisfy logical analysis are necessary, or empirical inquiry is hamstrung from the outset. There are no free inferential lunches in the absence of well-specified, coherent theory, and assumptions (Rouder, Morey, Verhagen, Province & Wagenmakers, 2016).

### *The slow and humble path forward*

The empirical component of Maul's article adopts an intentionally atheoretical approach to rhetorically demonstrate the hollow, conceptual centre of common survey validation practices. In bringing a critical eye to these studies, we did not miss the parody, but primarily aimed to reinforce and supplement Maul's condemnation of these atheoretical practices. Integrating logical analysis into survey validation may help to insulate the field from continued slavish devotion to statistical rules of thumb in the important task of validating our survey instruments.

More broadly, we strongly agree with Maul that it is well and truly time for survey methodology practitioners to shrug off operationalism's illogical ghost. A commitment to 'true grit' as evidenced in a slow meticulous process of critical inquiry invites the potential for survey construction and validation practices to bear such virtues as: logical coherence, conceptual clarity, rigorous theory for both attributes and response processes, as well as intrinsic consideration of the influence of the empirical context in survey responding, all before a single piece of survey data is ever collected. These kinds of fundamental conceptual considerations may already be seen to varying degrees in the seminal methodological research of Thurstone (1928), Guttman (1944), Coombs (1964), and even in the original work of Likert (1932). In moving forward, survey researchers may well benefit from extending an eye to the past.

## References

- Coombs, C. H. (1964). *A Theory of Data*. New Jersey, NJ: John Wiley & Sons.
- Guttman, L. (1944). *Louis Guttman on theory and methodology: Selected writings*. Aldershot, UK: Dartmouth Publishing Group.
- Hamby, T., & Taylor, W. (2016). Survey satisficing inflates reliability and validity measures: An experimental comparison of college and Amazon Mechanical Turk samples. *Educational and Psychological Measurement*, 76(6), 912-932.
- Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, 48(1), 400-407.
- Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5, 213-236.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 1-55.
- Petocz, A., & Newbery, G. (2010). On conceptual analysis as the primary qualitative approach to statistics education research in psychology. *Statistics Education Research Journal*, 9(2), 123-155.
- Popper, K. (1959). *The Logic of Scientific Discovery*. New York, NY: Basic Books.
- Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E. J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science*, 8(3), 520-547.
- Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, 33(4), 529-554.
- Wood, D., Harms, P. D., Lowman, G. H., & DeSimone, J. A. (2017). Response speed and response consistency as mutually validating indicators of data quality in online samples. *Social Psychological and Personality Science*, 8, 454-464.