Industrial and Organizational Psychology, **3** (2010), 481–484.

Developing and Scaling Personality Measures: Thurstone Was Right—But So Far, Likert Was Not Wrong

FREDERICK L. OSWALD
Rice University

KRAIG L. SCHELL
Angelo State University

Although we are in agreement with Drasgow, Chernyshenko, and Stark (2010) on the appropriateness of considering ideal point models in psychological testing, we focus on a number of questions that need to be addressed before concluding that the theoretical appeal of the ideal point model translates into a consistent empirical advantage to organizations that use personality tests.

Ideal Point Models Add Complexity

To date, the dominance model has dominated the thinking of test developers. In particular, both the dominance model and Rensis Likert together have created a prescription for constructing personality test items worded at the extremes of the trait continuum. For instance, the items ''I am always neat and tidy'' and ''I am considered a very messy person'' are two items reflecting high and low levels of order, respectively. When items are worded at the extremes like this, both the dominance response model and the ideal point model yield essentially the same personality trait scores, as Drasgow et al. emphasize (also see Chernyshenko, Stark, Drasgow, & Roberts [2007] who found $r = .92$ for order; similarly, Oswald [2010], predicted ideal point personality scores very well from much simpler scoring methods). In these cases, there is no apparent empirical advantage for ideal point models applied to traditional personality measures. Perhaps this is fortunate, in the sense that the vast pool of personality test findings from past organizational research and meta-analyses do not need to be revised or reversed on the basis of scoring methods aligned with the dominance model.

The potential added value of ideal point models would come from those personality items that are situated in the middle of the trait continuum, such as an orderliness item that reads ''Sometimes I am neat and tidy, but other times I am not.'' We will refer to items like these as *moderate level*. Unlike items worded at the extremes, moderate-level items tend to be complex or multidimensional (e.g., see the attitude items in Roberts, Laughlin, & Wedell, 1999), yet they often rely on unidimensional Likert-scale responses. This is generally why complex items should be

Correspondence concerning this article should be addressed to Frederick L. Oswald.
E-mail: foswald@rice.edu
Address: Department of Psychology, Rice University, 6100 Main Street, MS 25, Houston, TX 77005
Frederick L. Oswald, Department of Psychology, Rice University; Kraig L. Schell, Department of Psychology and Social Work, Angelo State University.

https://doi.org/10.1111/j.1754-9434.2010.01275.x Published online by Cambridge University Press

avoided (Hinkin, 1998), unless such items allow for similarly complex responses.

Returning to the example of orderliness, most people are not obsessively orderly nor are they incredibly disorganized. In fact, the very extremes of orderliness or any personality trait may reflect impaired life functioning or psychopathology (Markon, Krueger, & Watson, 2005). Therefore, test development efforts that are focused on moderate-level personality items might pay off in terms of higher acceptability or face validity as perceived by test takers. However, this advantage is not enough in and of itself to justify the switch to ideal point models. The use of moderate-level personality items and the ideal point model used to score them ought to yield a practical increase in reliability and validity when compared with simpler items scored by the dominance model. In other words, one wants to avoid using a more complicated measurement paradigm corresponding benefit trade-off. Science prefers parsimony unless the added complexity is justified (i.e., Occam's razor). One could predict that reliability and validity would be especially pronounced for people located in the middle of the trait continuum when utilizing the ideal point measurement model—a lot of people, given that most personality traits approximate a normal distribution. This prediction might bear itself out because high agreement on a moderate-level item indicates that a person has a moderate trait standing, whereas high disagreement means that a person could stand either high or low on the trait.

Therefore, it is worthwhile for future research to investigate the content and response processes for moderate-level items, given their potential to improve our measurement and understanding of personality traits. There is much work to be done, however, because there are at least three interpretations when a person strongly and positively endorses a moderate-level item such as the orderliness item ''Sometimes I am neat and tidy, but other times I am not.'' One interpretation is straightforward and follows the tenets of good personality measurement, where the person is assumed to possess a moderate level of the trait of orderliness, and the prediction of orderly thoughts, feelings, and behaviors is stable, albeit imperfect. A second interpretation is that the person possesses a highly nuanced version of orderliness, such that orderliness is highly contingent on psychological judgments or perceptions of the particular situational features at hand (e.g., see the CAPS model of Mischel & Shoda, 1995, or the trait-situation interactionist model of Tett & Burnett, 2003). The third interpretation is that the person is indifferent to orderliness, and any thoughts, feelings, and behaviors pertaining to orderliness are inconsistent and unpredictable. Critically, the second and third interpretations imply a very high or very low level of situational judgment or decision making (respectively), perhaps as much or more than they imply anything about a person's standing on a personality trait. To the extent these three possible interpretations are all viable ones, there is the potential for a high level of ambiguity from the typical moderate-level personality item. Advances in measurement content, delivery, and response processes may help disentangle this interpretational problem.

It is important to note that ambiguity not only exists for these moderate-level items that make the ideal point response model distinctive from the dominance model; it also applies to traditional Likert-scale items that are worded at the extremes. For Likert-scale items, however, the ambiguity shifts from the ''middling'' item content to the middle of the response scale. In fact, the research literature contains several empirical attempts to determine whether Likert-scale responses in the middle category are truly indicating a moderate level of endorsement of personality test items, as opposed to reflecting indifference or confusion about their content (Hernández, Drasgow, & González-Roma, 2004; Kulas & Stachowski, 2009; Rosenberg, Izard, & Hollander, 1955).

https://doi.org/10.1111/j.1754-9434.2010.01275.x Published online by Cambridge University Press

Issues in Need of Research

The Drasgow et al. paper and the present discussion together suggest several questions we view as important ones to research:

1. What is the practical impact when applying the ideal point versus dominance model approach to scoring measures of personality traits? Does it meaningfully change the magnitudes or patterns of validity for predicting organizational outcomes? Work on ideal point models has typically focused on reliability estimation, so the ongoing validity work mentioned by Drasgow et al. is very important because it can help to determine whether the validity of an innovative personality measurement system used in a high-stakes testing setting is driven by the forced-choice test format, by moderate-level personality items coupled with ideal point modeling, or both.
2. To improve the reliability and validity of people's trait scores in the middle of the trait continuum, is it better to add more moderate-level personality items (where the content is somewhat ambiguous) or more items at the high and low extremes (where a moderate-level response is somewhat ambiguous)? Impressively, Drasgow et al. found that over 25% of the items for their personality scale of order were better fit by the ideal point model because their moderate trait levels (Chernyshenko et al., 2007). It would also be valuable to know the size of the resulting increments in conditional reliability and validity along the trait continuum.
3. Does removing the middle category from a Likert scale reduce problems with responding to personality items with extreme wording, or does it merely disguise those problems or transform them into new ones?
4. How do subject matter expert judgments of item trait levels compare with the item locations empirically estimated from personality item data scored with ideal point item response theory models? What is the meaning of any large discrepancies, and does it have implications for item development and scoring?
5. How might supervisor or peer rankings of employees on personality traits improve the reliability and validity of those situated in the middle of a trait continuum, compared with self-report on moderate-level items with ideal point scoring? Sociometric ratings by peers and supervisors are consistent with the important concept of personality as reputation (Hogan, 1996). An approach to obtaining such ratings can be found in a current theory-driven method in the context of performance appraisal (Goffin, Jelley, Powell, & Johnson, 2009).

Thurstone was the person who proposed the ideal point model, but he also proposed the idea of simple structure in factor analysis (Thurstone, 1935), and he would likely agree that we should stick with simpler models until the data suggest that model complexity from ideal point scoring will improve model fit (Pitt & Myung, 2002; Preacher, 2006). Drasgow et al. suggest that we are in fact headed in this latter direction, as the substance of personality test development catches up with the statistical models and computer technology that allow for much greater flexibility in the delivery and scoring of personality tests. In our view, there is some distance to go, but we are making good headway as we enter an exciting era of research and applications involving personality testing in organizations.

References

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scores under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment, 19,* 88–106.

https://doi.org/10.1111/j.1754-9434.2010.01275.x Published online by Cambridge University Press

Drasgow, F. L., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *3*, 465–476.

Goffin, R. D., Jelley, R. B., Powell, D. M., & Johnson, N. G. (2009). Taking advantage of social comparisons in performance appraisal: The relative percentile method. *Human Resource Management*, *48*, 251–268.

Hernández, A., Drasgow, F., & González-Roma, V. (2004). Investigating the functioning of a middle category by means of a mixed-measurement model. *Journal of Applied Psychology*, *89*, 687–699.

Hogan, R. (1996). A socioanalytic perspective on the five-factor model. In J. S. Wiggins (Ed.), *The five-factor model of personality: Theoretical perspectives* (pp. 163–179). New York: Guilford.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods*, *1*, 104–121.

Kulas, J. T., & Stachowski, A. A. (2009). Middle category endorsement in odd-numbered Likert response scales: Associated item characteristics, cognitive demands, and preferred meanings. *Journal of Research in Personality*, *43*, 489–493.

Markon, K. E., Krueger, R. F., & Watson, D. (2005). Delineating the structure of normal and abnormal personality: An integrative hierarchical approach. *Journal of Personality and Social Psychology*, *88*, 139–157.

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, *102*, 246–268.

Oswald, F. L. (2010). *Practical recommendations for trait-level estimation in the Navy Computer Adaptive Personality Scales (NCAPS).* Millington, TN: Navy Personnel Research, Studies, and Technology.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, *6*, 421–425.

Preacher, K. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research*, *41*, 227–259.

Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement*, *59*, 211–233.

Rosenberg, N., Izard, C. E., & Hollander, E. P. (1955). Middle category response: Reliability and relationship to personality and intelligence variables. *Educational and Psychological Measurement*, *15*, 281–290.

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology*, *88*, 500–517.

Thurstone, L. L. (1935). *The vectors of the mind.* Chicago: University of Chicago Press.