Industrial and Organizational Psychology, **3** (2010), 465–476.

FOCAL ARTICLE

75 Years After Likert: Thurstone Was Right!

FRITZ DRASGOW
University of Illinois at Urbana-Champaign

OLEKSANDR S. CHERNYSHENKO
Nanyang Technological University

STEPHEN STARK
University of South Florida

Abstract

For over three-quarters of a century researchers and practitioners have analyzed rating scale data using methods that assume a dominance response process wherein an individual high on the trait assessed is assumed to answer positively with high probability. This approach derives from Likert's famous 1932 approach to the development and analysis of rating scales. In this paper, we argue that Likert scaling and related methods are misguided. Instead, we propose that methods that have evolved from Thurstone (1927, 1928, 1929) scaling provide a better representation of the choice process underlying rating scale judgments. These methods hypothesize an ideal point response process where the probability of endorsement is assumed to be directly related to the proximity of the statement to the individual's standing on the assessed trait. We review some research showing the superiority of ideal point methods for personality assessment and then describe several settings in which ideal point methods should provide tangible improvements over traditional approaches to assessment.

In a series of remarkable papers in the late 1920s, Louis Thurstone asserted ''Attitudes can be measured'' (1928); see also Thurstone (1927, 1929). Central to his approach was the assumption that a conscientious person would endorse a statement that reflected his or her attitude, but ''as a result of imperfections, obscurities, or irrelevancies in the statement, and inaccuracy or carelessness of the subjects'' not everyone would respond accurately (1929, p. 224). Using Thurstone's notation, suppose there were N_1 people with an attitude value of S_1; all should endorse a statement with scale value S_1 if they were conscientious and the item was perfect. In practice, Thurstone expected only n_1 to agree with the statement, where $n_1 < N_1$. Moreover, these people would endorse another statement with scale value S_2 (where $S_2 \neq S_1$) with probability p that is inversely related to $|S_2 - S_1|$. Figure 1, from Thurstone's 1929 paper, illustrates his theory.

In his 1928 paper, Thurstone used the example of a militarism–pacifism attitude with six statements representing a range of attitudes. Figure 2, from Thurstone (1928, p. 537), gives the locations of the six statements and shows the distribution of

Correspondence concerning this article should be addressed to Fritz Drasgow. E-mail: fdrasgow@uiuc.edu

Address: Department of Psychology, University of Illinois, 603 E. Daniel Street, Champaign, IL 61820

Fritz Drasgow, Department of Psychology, University of Illinois at Urbana-Champaign; Oleksandr S. Chernyshenko, Nanyang Technological University; Stephen Stark, University of South Florida.

https://doi.org/10.1111/j.1754-9434.2010.01273.x Published online by Cambridge University Press

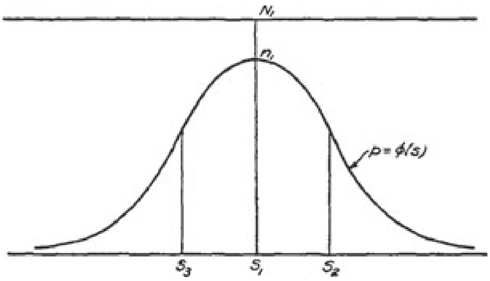

Figure 1. Response probabilities from Thurstone (1929). Reprinted from Thurstone (1929, p. 229).

the attitude in some population. Thurstone noted a pacifist ''would be willing to indorse all or most of the opinions in the range *d* to *e* and . . . he would reject as too extremely pacifistic most of the opinions to the left of *d*, and would also reject the whole range of militaristic opinions'' (p. 539). Of critical import is Thurstone's method of scoring: A person's ''attitude would then be indicated by the average or mean of the range [of statements] that he indorses'' (p. 539). For example, Person 1 might endorse statements *f* and *d* from Figure 2, Person 2 might endorse *e* and *b*, and Person 3 might endorse *c* and *a*. Although each endorses two items, their attitudes are quite different. By attending to *which* items are endorsed and not simply how many, Thurstone's scoring allows individuals who endorse the same number of items, but who have different attitudes, to be differentiated.

In 1932, Likert provided a much simpler alternative to Thurstone scaling. Although he examined several approaches, Likert found that using a 5-point response scale with options ''Strongly Approve,'' ''Approve,'' ''Undecided,'' ''Disapprove,'' and ''Strongly Disapprove,'' and integer scoring (''Strongly Approve'' = 5, . . ., ''Strongly Disapprove'' = 1), yielded ''the same reliability with fewer items, or higher reliability with the same number of items'' as Thurstone scaling (p. 34). Using an Internationalism scale as an example, Likert assigned the largest value of the response scale ''to the end [of each statement] which seemed to favor internationalism'' and thereby invented reverse scoring of negatively phrased items. After reverse scoring, an individual's total score could be taken as the sum or mean of the item scores.

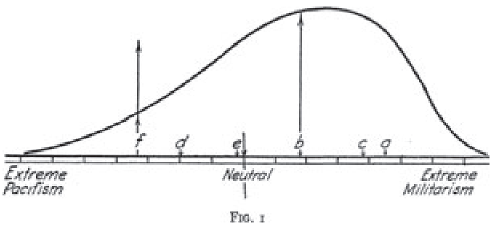

Figure 2. Locations of six militarism–pacifism attitude statements. Reprinted from Thurstone (1928, p. 537).

A crucial difference between Likert and Thurstone concerns intermediate statements like ''Compulsory military training in all countries should be reduced but not eliminated'' (p. 34). Likert argued that ''It is impossible to tell whether a person is agreeing or disagreeing with the 'reduction' aspect of this statement or the 'not eliminated' aspect'' and therefore this ''statement is double-barreled and of little value because it does not differentiate persons in terms of their attitudes'' (p. 34). Consequently, Likert recommended deleting items like this intermediate statement. Thurstone, on the other hand, viewed this statement, like statement e in Figure 2, as necessary for accurately measuring the attitudes of people with intermediate standings. Therefore, Thurstone deliberately wrote intermediate items and included them in his measures.

Likert based his conclusion that double-barreled items were worthless on item–total correlations. He argued that ''If a zero or very low correlation coefficient is obtained, it indicates that the statement fails to measure that which the rest of the statements measure'' (p. 48) and ''Thus item analysis reveals the satisfactoriness of any statement so far as its inclusion in a given attitude scale is concerned'' (p. 49).

Although Likert did not articulate a psychometric model for his procedure, his approach implies what Coombs (1964) called a *dominance response process*. Here an individual high on the trait or

https://doi.org/10.1111/j.1754-9434.2010.01273.x Published online by Cambridge University Press

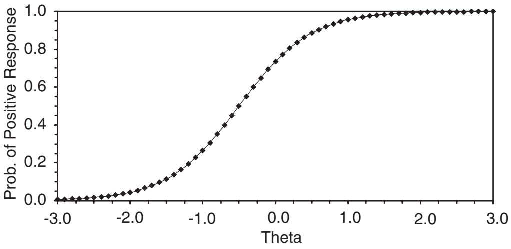

Figure 3. Probability of a positive response to a dichotomously scored statement as a function of the latent attitude (theta).

attitude measured by a scale is likely to ''Strongly Agree'' with a positively worded statement and ''Strongly Disagree'' with a negatively worded statement. Figure 3 illustrates a dominance response process for a dichotomously scored item. Note that as the attitude—labeled theta in Figure 3—increases, so does the probability of a positive response.

We believe that dominance models are most sensibly applied to domains in which an individual's capacity or maximum performance capability is pitted against the difficulty or extremity of an item (Stark, Chernyshenko, Drasgow, & Williams, 2006; Tay, Drasgow, Rounds, & Williams, 2009). Consider a weight lifter attempting to clean and jerk a series of increasing weights. The weight lifter's strength would be denoted as theta in Figure 3 and a response function—the smooth curve in the figure—could be created for each weight. A stronger weight lifter would be indicated by a theta value further to the right on the figure and he or she would have a higher probability of successfully lifting the weight.

Much of the work on dominance models has been in the context of cognitive ability testing. Here, an individual's ability is assessed by a set of items of varying difficulty. An individual with a high ability level is expected to answer all the easy items correctly, all the moderately difficult items correctly, and some of the most difficult items correctly. Thus, the individual *dominates* the easy and moderately difficult items.

We argue that psychometric models for dominance response processes, such as classical test theory, factor analysis, and the logistic item response theory (IRT) models, are ill suited for response processes requiring introspection (Chernyshenko, Stark, Drasgow, & Roberts, 2007; Tay et al., 2009). When individuals consider their militarism–pacifism or internationalism, we believe that they ask ''Does this statement closely describe me?'' The closer an item's location on the attitude continuum to the individual's location on that continuum, the greater the probability that the person will endorse the item. The maximum probability of endorsement occurs when the attitude level of the item equals the individual's attitude level. This is the idea underlying Thurstone's model shown in Figure 1. Coombs (1964) is credited with coining the term ''ideal point'' and wrote, ''We conceive, then, of representing an individual by a point in the same space containing the stimulus points, in such a way that the point corresponding to the individual is a point of his maximum preference in this domain of stimuli'' (p. 8). Coombs used the term ''unfolding technique'' for formal representations of this process because the probability of endorsement decreases in both directions from the individual's ideal point: statements representing lower and higher locations on the latent trait continuum have decreasing probabilities of being endorsed as they are further away from the individual's ideal point.

Figure 4 presents a more modern formulation of the Thurstone's model shown in Figure 1. Here, the probability of a positive response to a dichotomously scored item is given as a function of the trait or attitude it assesses; this curve is called an item response function. For example, an extraversion scale might contain the item ''I enjoy chatting quietly with a friend at a café.'' As shown in Figure 4, individuals who are too introverted might tend to disagree with the item because they are uncomfortable in public places. In direct contrast to a dominance response process, individuals who are high on the trait

https://doi.org/10.1111/j.1754-9434.2010.01273.x Published online by Cambridge University Press

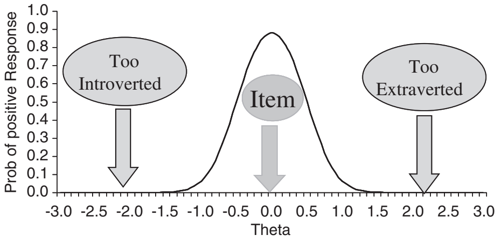

Figure 4. Item response function for an ideal point response process.

assessed—extraversion in Figure 4—also tend to disagree because they prefer more exciting settings.

What is the item–total correlation of items of the sort depicted in Figure 4? Individuals who are low in extraversion would tend to receive scores of 0 on this dichotomously scored item and individuals who are high in extraversion would also tend to receive 0's. Only individuals who are intermediate would tend to receive 1's. There is little *linear* trend in the item score (but certainly a nonlinear trend) as a function of the trait assessed and therefore the item–total correlation would be close to zero. Is this a bad item? No! It provides useful information about the trait assessed—a score of 1 indicates an intermediate degree of extraversion, whereas a 0 indicates high or low (but not intermediate) extraversion. We believe Likert was misled because, as is so common in psychology, he looked only for linear relations.

Personality Assessment and Ideal Point Models

In the late 1990s, we began fitting IRT models to data from personality scales. Based on earlier work by Reise and Waller (1990), we expected the two-parameter logistic model (2PLM) to fit well. To our surprise, fits of the 2PLM and other IRT models were noticeably worse than the fits of IRT models to data from cognitive ability tests. In retrospect, we should not have been surprised because personality items are essentially attitude statements about oneself (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001) and Thurstone's work on attitude measurement should be directly applicable to personality.

In our research, we have used a chi-square goodness-of-fit measure that compares the observed and expected (on the basis of the IRT model) frequencies of endorsing/not endorsing items. This measure can be computed for individual items, pairs of items, and triples of items. For example, the chi-square for a pair of items compares the observed and expected frequencies in the two-way table formed by crossing endorse/not endorse on the first item with endorse/not endorse on the second item (see Drasgow, Levine, Tsien, Williams, & Mead, 1995 for details).

Interestingly, we found (Chernyshenko et al., 2001) that IRT models—even misspecified ones—do a good job of reproducing the observed frequencies of single items. The chi-squares for pairs and triples of items present a more challenging test of the fit of the IRT model (we have found that higher order tables have less power to detect misfit, presumably because the sample sizes in some cells become too small).

When an IRT model is estimated using data from a large sample (i.e., 3,000 or more) and fit is evaluated in a cross-validation sample, we've found that a chi-square to degrees of freedom (*df*) ratio of less than 2 indicates an excellent fit, between 2 and 3 indicates a generally satisfactory fit, and over 3 indicates misfit.

Sixteen Personality Factor (16PF) Questionnaire

Chernyshenko et al. (2001) fit several IRT models to data from the 16PF (Conn & Rieke, 1994). Although they analyzed all 16 scales, we shall discuss just the Sensitivity scale. Data from 6,455 individuals were used to estimate item parameters and 6,456 individuals served as a cross-validation sample. The mean chi-square to *df* ratio for single items was 0.98 for the 2PLM, but was 4.05 and 5.45 for pairs and triples of items, which is clearly unsatisfactory. The three-parameter logistic model (3PLM) did little

https://doi.org/10.1111/j.1754-9434.2010.01273.x Published online by Cambridge University Press

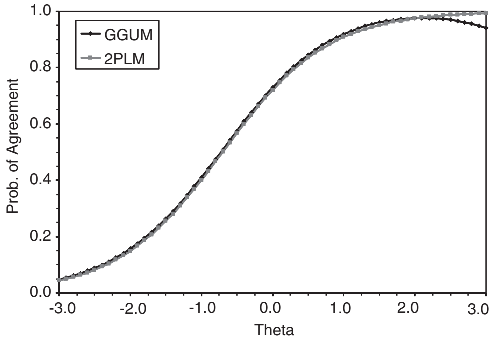

Figure 5. Two-parameter logistic model and generalized graded unfolding model item response functions for the same item.

better, with 0.87, 3.89, and 5.23 for singles, doubles, and triples of items, respectively. These findings led us to question the appropriateness of the logistic IRT models for personality data.

We also fit Levine's (1984) maximum likelihood formula score model (MFSM). MFSM is a nonparametric IRT model and thereby provides great flexibility: The best fitting item response function need not be logistic or even monotonic. Although the fit of this model to individual items was barely satisfactory (mean chi-square to *df* ratio of 2.91), the fit to doubles and triples of items was noticeably better (2.61 and 2.42, respectively) than the two logistic models. Interestingly, some of the MFSM item response functions showed nonmonotonicity: The probabilities of endorsement did not always increase, which is the hallmark of dominance models. Instead, the probability of endorsement increased in one part of the trait continuum but *decreased* in another part. We are indebted to Michael Levine for pointing out that this nonmonotonicity is suggestive of an ideal point response process.

Developers of personality scales compute item–total correlations, internal consistency reliability, and factor loadings and delete items with poor psychometric properties according to these dominance model analyses. Consequently, it was surprising to find evidence of unfolding (i.e., endorsement probabilities that increased at lower trait levels and then decreased at higher trait levels) with the 16PF scales. Note that by deleting intermediate items like the one depicted in Figure 4 and reverse scoring negative items, we should be left with items that ought to be fit well with the 2PLM: Figure 5 shows endorsement probabilities estimated with an ideal point model (the generalized graded unfolding model [GGUM] developed by Roberts, Donoghue, & Laughlin, 2000) and the 2PLM. These item response functions were estimated for the item ''Even when something bad happens, I can push negative thoughts out of my mind'' from the Well-Being scale of the Tailored Adaptive Personality Assessment System (TAPAS; Drasgow, Chernyshenko, & Stark, 2010). These two item response functions are virtually identical over almost all the trait continuum, differing appreciably only at trait values above 3.0 (IRT trait scores are given as standardized scores). Because these two item response functions are virtually identical, they cannot be differentiated on the basis of a goodness-of-fit measure.

Constructing New Personality Scales

Table 1 briefly summarizes instrument development for dominance and ideal point models. Note that there are important differences at each step. For example, within the ideal point framework, scores could be computed as the mean of the item locations of the items endorsed, rather than the proportion endorsed or sum of item scores as commonly used within the dominance framework.

To better understand the characteristics of items that have *not* been preselected by dominance model methods, Chernyshenko et al. (2007) reported the analysis of responses from 539 students to 50 new single-statement personality items written for the Order facet of the conscientiousness Big Five personality dimension. Items were written to assess the entire range of this dimension: low, intermediate, and high. For dominance model analyses, items written

https://doi.org/10.1111/j.1754-9434.2010.01273.x Published online by Cambridge University Press

Table 1. *Comparison of Dominance Model and Ideal Point Model Procedures*

Activity	Dominance model	Ideal point model
Item development	Write positive and negative items	Write positive, intermediate, and negative items
Item scoring	Reverse score negative items	Do not reverse score
Item analysis	Compute item–total correlations	Compute item–subtotal correlations[a]
IRT analysis	Use a logistic model	Use an ideal point model
Trait scores	Compute the proportion endorsed, the sum of item scores, or the logistic model trait estimate	Compute the mean item location of the items endorsed or the ideal point model trait estimate

[a] Subtotal scores for the negative, intermediate, and positive items can be computed and then correlations of each negative item with the negative item subtotal can be computed, correlations of the intermediate items with the intermediate item subtotal can be computed, and correlations of the positive items with the positive item subtotal can be computed.

to assess low trait values were reverse scored. Using an item–total correlation of .3 as a criterion, which was suggested by Nunnally and Bernstein (1994) as the cutoff for item retention, 13 items would have been rejected. Importantly, most of those items were designed to assess intermediate levels of orderliness. In contrast, when the items were analyzed with an ideal point IRT computer program (GGUM2000 developed by Roberts, 2001), only two items were found to have unsatisfactory discrimination parameter estimates (i.e., below 0.4). Thus, virtually all the intermediate items rejected in the dominance model analyses were found to have good psychometric properties in the ideal point analysis.

We have now developed personality statements representing 22 facets of the Big Five dimensions. The statements can be administered in various formats using the TAPAS software as an online computerized adaptive test. Results for the Well-Being facet of the emotional stability Big Five dimension are typical. Drasgow, Chernyshenko, and Stark (2009) reported findings for 20 single-statement items completed by 445 Army recruits. The items were administered using a 4-point response scale with 1 = ''Strongly Disagree,'' 2 = ''Disagree,'' 3 = ''Agree,'' and 4 = ''Strongly Agree.''

The five items designed to assess low well-being had a mean factor loading of −.50 before reverse scoring and a mean corrected item–total correlation of .41 after reverse scoring. Nine items intended to assess high levels of well-being had a mean factor loading of .52 and a mean corrected item–total correlation of .46. In contrast, the six items assessing intermediate levels of well-being (e.g., ''My life has had about an equal share of ups and downs'') had a mean factor loading of .02 and a mean item–total correlation of .07. Clearly, these items would have been deleted if dominance model methods were used. However, their mean GGUM2000 item discrimination was 0.87 after dichotomously rescoring the items, which is a bit lower than the 1.09 mean of the negative items and the 1.35 mean of the positive items, but is nonetheless quite satisfactory.

When the fits of the models were examined, we obtained an excellent 0.71 mean adjusted chi-square to *df* ratio for the GGUM analysis of pairs of dichotomously scored items. In contrast, the mean adjusted chi-square to *df* ratio was 2.39 for pairs of items when the 2PLM was used. It is important to note that the 2PLM failed to fit in a predictable and systematic way: It was unable to accurately model the responses to the intermediate items. When a 2 × 2 table delineating positive versus negative responses for two intermediate items were constructed, the 2PLM predicted that about 25% of the respondents would fall into each

https://doi.org/10.1111/j.1754-9434.2010.01273.x Published online by Cambridge University Press

cell. However, the observed frequencies in the positive/positive and negative/negative cells were much higher than 25% because people intermediate in well-being tended to agree or strongly agree with both items and people low or high in well-being tended to disagree or strongly disagree with both.

If the six intermediate items were deleted, the fits of the 2PLM and GGUM would be very similar. That is because item response functions (IRFs) for high and low well-being items are virtually indistinguishable across the observed ranges of the trait continuum—Figure 5 presents IRFs for one such item. As can be seen, the expected endorsement probabilities for the two models are nearly the same for all but extremely positive trait levels, perhaps reflecting the fact that highly optimistic individuals do not admit experiencing ''bad'' events and may begin to disagree with the item. This is reflected in the gradual descent of the GGUM IRF at trait values of 2.2 and higher. However, because the number of such individuals is very small, it has little material effect on trait estimation or fit.

Despite the fact that this item has similar IRFs under the two models, the terminology used to describe the item properties differs depending on which model is used. The GGUM location parameter for this item is 2.12 indicating that it is positive in location and wording. In ideal point terminology, item location refers to the point on the trait continuum where the probability of endorsement is highest. In Thurstone's Law of Comparative Judgment (1927), this is the scale value of the item. On the other hand, the 2PLM location parameter or, as it is often called, the ''difficulty'' parameter is −0.70. The term comes from cognitive ability terminology and indicates the point on the trait continuum where the probability of a correct response is .50. Readers unfamiliar with the applications of dominance IRT models in noncognitive assessment may erroneously assume that the item is negatively worded. In fact, all we can say is that the item is ''easy'' (i.e., individuals with trait levels of −0.70 are expected to endorse the item with a probability of 50%). As was noted by Chernyshenko et al. (2007), dominance model difficulty parameters (as well as *p* values) do not have easily understood relationships with item content in noncognitive domains.

Why Does the Choice of Psychometric Model Matter?

Currently available psychometric models for ideal point data are considerably more complicated than corresponding models for dominance data; see, for example, Equation 7 in Roberts et al. (2000, p. 6). Thus, it is reasonable to ask whether there are any tangible benefits that accrue from this added complexity. Wouldn't it be better just to delete intermediate statements, as we have done for the past 75 years, and work with easily scalable positive and negative statements?

A first answer is that from the perspective of basic science it is important to understand how and why people respond to assessment instruments. Understanding how people answer such assessments provides us with deeper insights into the nature of their responses. In this paper and elsewhere we have argued that responses to questions requiring introspection involve a comparison process. The individual considers his or her behaviors, attitudes, feelings, or whatever is assessed and then considers what the item asks. The decision to endorse or not endorse the item is then, we believe, driven by the psychological distance between the self-perception and the perception of the statement. A formal psychometric model of this process was given by Zinnes and Griggs (1974).

A second answer is that applying the wrong measurement model can lead to mistakes. For example, Davison (1977) showed that factor analysis of a unidimensional set of ideal point items produces two factors. Questions about the latent structure of emotion, comparisons of the Big Five versus the HEXACO model (Ashton & Lee, 2007), and other debates about constructs assessed by introspection may be clouded by the applications of misspecified models.

https://doi.org/10.1111/j.1754-9434.2010.01273.x Published online by Cambridge University Press

Kurt Lewin wrote ''There is nothing so practical as a good theory'' (1951, p. 169) and we believe that there are numerous important implications of and applications for a good measurement theory for self-report data.

As a first application, consider the development of a new assessment tool. With an ideal point perspective, item writers can intentionally write items to assess low, intermediate, and high trait values and thus create an instrument that provides excellent measurement precision across a broad range of the latent trait continuum. In contrast, item writers do not create intermediate items when working within the dominance framework because they know that low item–total correlations will result. Paradoxically, it turns out that intermediate statements that are shunned in dominance frameworks tend to improve measurement at high and low trait values. Intuitively, one can see that adding an intermediate statement, such as ''My life has had about equal amounts of ups and downs,'' allows us to separate very high and very low well-being individuals from everyone else, because those individuals would disagree with the statement. This was illustrated by Chernyshenko et al. (2007) for the Order facet of conscientiousness: A substantial gain in test information, and a corresponding reduction in the conditional standard error of measurement, was obtained for a wide range of trait values. In research contexts where the focus is primarily on correlation coefficients, all that is needed is a rough separation of respondents into high and low trait groups, so traditionally used scales are more than adequate. In some important applications (e.g., personnel selection), however, correct rank ordering of individuals at the extreme trait ranges is of critical concern, so improving measurement precision would be beneficial.

Second, in our view, some constructs in organizational psychology might be better studied by embracing an ideal point perspective. For example, central to the notion of person–organization (P–O) fit is a comparison process involving what a person desires and what an organization provides. Interest usually revolves around the gaps between P and O on various dimensions. The magnitudes and directions of these gaps can, of course, be gauged by measuring personal needs and organizational supplies via separate assessments and deriving scores through simple mechanical (e.g., profile similarity correlations) or more complex statistical methods (e.g., polynomial regression [Edwards, 1994]). But an alternative is to write statements that explicitly capture the magnitude and direction of differences between P and O along various dimensions and scale the statements and persons for each dimension separately using unidimensional ideal point models (Chernyshenko, Stark, & Williams, 2009). For example, the statement ''I wish I had more autonomy at my current workplace'' reflects a positive gap between P and O on the fit dimension of autonomy, whereas the statement ''Managers give way too much freedom to employees here'' reflects a negative gap. Furthermore, the statement ''The amount of autonomy I get here is just perfect; I don't need any more or any less'' reflects near perfect congruence between personal needs and organizational provisions. Importantly, by scaling such statements using ideal point models, the resulting P–O fit scores have interpretations that are consistent with organizational theory; a score of 0 reflects excellent fit, whereas positive or negative scores indicate varying degrees of misfit.

Another domain where an ideal point perspective makes sense is performance appraisal. Statements describing low, moderate, and high levels of employee performance can be written and a rater can be asked to endorse statements that accurately describe an employee's performance level. Here the employee's performance would constitute the latent trait that we wish to estimate, and we believe that an ideal point response process describes the way in which a rater chooses which statements to endorse.

Borman et al.'s (2001) computerized adaptive rating scale (CARS) provides a

https://doi.org/10.1111/j.1754-9434.2010.01273.x Published online by Cambridge University Press

good example of the use of an ideal point model for performance ratings. Raters are given two statements reflecting different levels of performance and asked to choose the one that is a better description of the ratee. Zinnes and Griggs' (1974) ideal point model is used to describe the judgment process: Raters are assumed to compare each statement to the ratee's performance and then select the statement that is perceived to be closer. Borman et al. found substantially lower standard errors of measurement and higher validity for CARS in comparison with performance ratings obtained from graphic rating scales and behaviorally anchored rating scales.

The fourth, and perhaps most exciting application of ideal point models, is their use in conjunction with personality assessment instruments using forced-choice response formats.[1] Interest in forced-choice formats has been rejuvenated by relatively recent research suggesting that they have criterion-related validity and may be less susceptible to rater biases and response distortions commonly associated with Likert-type scales (Jackson, Wroblewski, & Ashton, 2000; see also Christiansen, Burns, & Montgomery, 2005; McCloy, Heggestad, & Reeve, 2005; Vasilopoulos, Cucina, Dyomina, Morewitz, & Reilly, 2006). For example, traditional single-statement items are often transparent and, when there are strong incentives to fake, respondents have been shown to increase their scores by as much as 1.5 standard deviations (Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001). Forced-choice measures, on the other hand, typically consist of two or more statements that are matched in terms of social desirability and often the items are multidimensional. For example, respondents might be asked to ''choose the statement that is more like you'' given a pair like:

— I get along well with others.
— I always get my work done on time.

By using this approach, not only is it harder to discern the correct answers (they may depend on the application), but also it is more difficult to raise substantially one's scores on all dimensions simultaneously.

Forced-choice measures involving items composed of pairs or tetrads of statements have been used for noncognitive assessment in the past (Edwards, 1954; Rounds, Henly, Dawis, Lofquist, & Weiss, 1981; White & Young, 1998), but difficulties arise. For example, the use of paired comparison items with traditional scoring is problematic because the results are ipsative (i.e., the total score on the assessment is the same for each respondent), which causes a variety of difficulties (Meade, 2004). Consequently, researchers had to rely on heuristics in scale construction or data coding to introduce variability into scale scores for interindividual comparisons in selection applications (Chernyshenko et al., 2009; Meade, 2004). To provide a model-based solution to these problems, we have developed and applied IRT models (Stark, Chernyshenko, & Drasgow, 2005; Stark & Drasgow, 2002) for paired comparison judgments.

Our IRT model for the multidimensional pairwise preference (MDPP) format is called the multi-unidimensional pairwise preference model (MUPPM; Stark, 2002; Stark et al., 2005). It asserts that people first decide whether each of the two statements describes them. This first step is modeled by the GGUM (Roberts et al., 2000). Then, if one statement or the other (but not both) is judged to describe them, they select that statement. If both or neither statement is chosen, respondents reconsider each statement in a way akin to Thurstone's (1927) discriminal process, until exactly one statement is perceived as describing them. Our simulation and empirical studies (Chernyshenko et al., 2009; Stark et al.,

[1] Of course, it would be possible to use a dominance model to describe the response process underlying forced-choice judgments. However, model misspecification in this context might lead to serious problems.

https://doi.org/10.1111/j.1754-9434.2010.01273.x Published online by Cambridge University Press

2005) have shown that accurate latent trait estimates (i.e., normative scores) can be obtained when analyzing pairwise preference responses with MUPPM.

The use of IRT models capable of adequately representing forced-choice statement selection opens up a gamut of intriguing possibilities for developing new kinds of assessment instruments. For example, as the number of statements in a testing pool expands, the number of forced-choice items that can be formed increases exponentially, which is ideal for computerized adaptive testing (CAT). Thus, adaptive testing with the MDPP format enables modest-sized pools of statements (e.g., 40 or 50 per personality facet) to generate tens of thousands of items, even with matching constraints on statements' social desirability and location parameters. As with traditional IRT models, simulation studies have shown that CAT greatly reduces the number of MDPP items needed to achieve good measurement precision relative to nonadaptive tests (Stark & Chernyshenko, 2007). Importantly, having large numbers of potential pairings makes test compromise less of a concern in unproctored web-based testing environments.

Currently, in cooperation with the Army Research Institute, the Military Enlistment Processing Command, and the Defense Manpower Data Center, a personality test battery composed of MDPP items is being evaluated for military enlistment screening. Applicants for enlistment in the U.S. Army and U.S. Air Force as well as active duty personnel have taken either a paper-and-pencil version or a computer-adaptive version of the test and are being tracked on a variety of criterion measures to see if personality dimensions can add incremental validity to the existing selection process. Because parts of the sample have taken the test under operational conditions, it will be possible to evaluate the feasibility of using MUPPM in high-stakes testing environments.

Final Thoughts

Seventy-five years after Likert wrote his 1932 paper we believe there is compelling evidence that his approach does not do justice to the underlying processes by which people make introspective judgments. Certainly, as a rough and ready approach, a Likert scale works well. But for research and applications requiring a high fidelity representation of choice processes, the Likert approach has shortcomings.

As psychologists develop assessment tools needed for their research, it seems important for the approach to measurement to be in harmony with respondents' decision processes. P–O fit or performance appraisals provide good examples of domains where the approach to measurement can be designed to be consistent with the way people make judgments. This should make responding to questionnaires easier and more straightforward for research participants and thereby improve the quality of data that are collected.

In addition to making the task for respondents easier, researchers should be able to design better assessment tools. It has been 30 years since Frederic Lord (1980) published his seminal *Applications of Item Response Theory to Practical Testing Problems* and, today, it would be difficult to find a cognitive ability test that did not use a 3PLM or a Rasch model for its design, administration, or scoring. Perhaps, in another 30 years, the same will be said of the use of ideal-point IRT models in testing domains requiring introspection (e.g., personality, values, or performance).

In addition to the personality, P–O fit, and job performance variables described earlier, ideal point models may be usefully applied to many other important variables in industrial and organizational psychology. For example, job satisfaction, organizational commitment, leader behavior, subjective well-being, perceived organizational support, and many other variables may be fruitfully conceptualized and assessed via ideal point models.

In sum, we believe that using the right measurement model holds great promise for improved research and practice. Ideal point modeling should foster improved instruments, more straightforward linkage

https://doi.org/10.1111/j.1754-9434.2010.01273.x Published online by Cambridge University Press

between item content and psychometric parameters, and sophisticated applications such as CAT. Of course, many issues and problems remain to be solved; psychometricians have devoted great effort to dominance models during the past 100 years. In contrast, just a few psychometricians have worked on ideal point models and there have been few applications to applied measurement problems. The opportunity for new and creative research is enormous.

References

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, *11*, 150–166.

Borman, W. C., Buck, D. E., Hanson, M. A., Motowidlo, S. J., Stark, S., & Drasgow, F. (2001). An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. *Journal of Applied Psychology*, *86*, 965–973.

Chernyshenko, O. S., Stark, S., Chan, K.-Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, *36*, 523–562.

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

Chernyshenko, O. S., Stark, S., Prewett, M. S., Gray, A. A., Stilson, F. R., & Tuttle, M. D. (2009). Normative scoring of multidimensional pairwise preference personality scales using IRT: Empirical comparisons with other formats. *Human Performance*, *22*, 1–23.

Chernyshenko, O. S., Stark, S., & Williams, A. (2009). Latent trait theory approach to measuring person-organization fit: Conceptual rationale and empirical evaluation. *International Journal of Testing*, *9*, 358–380.

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance*, *18*, 267–307.

Conn, S., & Rieke, M. L. (Eds.) (1994). *The 16PF fifth edition technical manual.* Champaign, IL: Institute for Personality and Ability Testing.

Coombs, C. H. (1964). *A theory of data.* New York: Wiley.

Davison, M. L. (1977). On a metric, unidimensional unfolding model for attitudinal and developmental data. *Psychometrika*, *42*, 523–548.

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2009). Test theory and personality measurement. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 59–80). New York: Oxford University Press.

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). *Tailored Adaptive Personality Assessment System (TAPAS).* Urbana, IL: Authors.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement*, *19*, 143–165.

Edwards, A. L. (1954). *Personal preference schedule.* New York: Psychological Corporation.

Edwards, J. R. (1994). The study of congruence in organizational behavior research: Critique and a proposed alternative. *Organizational Behavior and Human Decision Processes*, *58*, 51–100 (erratum, *58*, 323–325).

Hough, L. M., Eaton, N. L., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities [Monograph]. *Journal of Applied Psychology*, *75*, 581–595.

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced-choice offer a solution? *Human Performance*, *13*, 371–388.

Levine, M. V. (1984). *An introduction to multilinear formula score theory* (Personnel and Training Research Programs, Office of Naval Research, Measurement Series No. 84-4). Arlington, VA: Personnel and Training Research Programs.

Lewin, K. (1951). *Field theory in social science; selected theoretical papers.* D. Cartwright (Ed.), New York: Harper & Row.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*, 1–55.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

McCloy, R. A., Heggestad, E. D., & Reeve, C. L. (2005). A silk purse from the sow's ear: Retrieving normative information from multidimensional forced-choice items. *Organizational Research Methods*, *8*, 222–248.

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology*, *77*, 531–551.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory.* New York: McGraw-Hill.

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement*, *14*, 45–58.

Roberts, J. S. (2001). GGUM2000: Estimation of parameters in the generalized graded unfolding model. *Applied Psychological Measurement*, *25*, 38.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement*, *24*, 3–32.

Rounds, J. B., Henly, G. A., Dawis, R. V., Lofquist, L. H., & Weiss, D. J. (1981). *Manual for the Minnesota Importance Questionnaire.* Minneapolis, MN: University of Minnesota, Vocational Psychology Research.

Stark, S. (2002). *A new IRT approach to test construction and scoring designed to reduce the effects of faking in personality assessment.*

https://doi.org/10.1111/j.1754-9434.2010.01273.x Published online by Cambridge University Press

Unpublished doctoral dissertation, University of Illinois at Urbana–Champaign.

Stark, S., & Chernyshenko, O. S. (2007, June). *Adaptive testing with the multi-unidimensional pairwise preference (MUPP) model*. Paper presented at the 2007 Graduate Management Admissions Council conference on Computerized Adaptive Testing. Minneapolis, MN.

Stark, S., Chernyshenko, O. S., Chan, K.-Y., Lee, W. C., & Drasgow, F. (2001). Effects of the testing situation on item responding: Cause for concern. *Journal of Applied Psychology*, *86*, 943–953.

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*, 184–203.

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology*, *91*, 25–39.

Stark, S., & Drasgow, F. (2002). An EM approach to parameter estimation for the Zinnes and Griggs paired comparison ideal point IRT model. *Applied Psychological Measurement*, *26*, 208–227.

Tay, L., Drasgow, F., Rounds, J., & Williams, B. A. (2009). Fitting measurement models to vocational interest data: Are dominance models ideal? *Journal of Applied Psychology*, *94*, 1287–1304.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273–286.

Thurstone, L. L. (1928). Attitudes can be measured. *The American Journal of Sociology*, *33*, 529–554.

Thurstone, L. L. (1929). Theory of attitude measurement. *Psychological Review*, *36*, 222–241.

Vasilopoulos, N. L., Cucina, J. M., Dyomina, N. V., Morewitz, C. L., & Reilly, R. R. (2006). Forced-choice personality tests: A measure of personality and cognitive ability? *Human Performance*, *19*, 175–199.

White, L. A., & Young, M. C. (1998, August). *Development and validation of the Assessment of Individual Motivation (AIM)*. Paper presented at the 106th Annual Convention of the American Psychological Association, San Francisco, CA.

Zinnes, J. L., & Griggs, R. A. (1974). Probabilistic, multidimensional unfolding analysis. *Psychometrika*, *39*, 327–350.

https://doi.org/10.1111/j.1754-9434.2010.01273.x Published online by Cambridge University Press