Likert Scales and Performance Appraisals

When a researcher is attempting to measure more than a single item, she/he might use a scale, “a number of items that are combined to form a composite score on a variable” (Monette et al., 2002, 356). For instance, inquiry about marital status or age is rather simple, entailing only one item. However, when trying to determine effectiveness in job functioning, a number of variables would be far more helpful in yielding pertinent information since no single variable would capture that aspect. In that genre, a legitimate and quantitative measure is that of the Likert Scale. Developed in 1932 by Rensis Likert, the scale is ubiquitous in the research methods literature and familiar to the general public (Monette et al., 2002, 360). Evaluations of courses and college professors, among others, typically employ this scale.

Used to measure attitudes, the scale may have anywhere from two to nine or more categories. Respondents locate their attitudes along a continuum such as this one: Strongly Disagree—Disagree—Neutral—Agree—Strongly Agree (Wolfer, 2007, 530). Researchers disagree on the actual number of the varying degrees of intensity measured but most settle on a minimum of 4-5 dimensions (Isaac and Michael, 1995, 149; Wolfer, 2007, 530; Neuman 2004, 127). A host of variables correlated with job performance, for example, can then be assessed with this format. Moreover, a tally of respondent scores can occur, making the Likert scale a summated rating scale: “... a person’s score is determined by summing the number of questions answered in a particular way” (Monette et al., 2002, 360). On a scale with ten items, for instance, and four possible responses, a person can have a score that ranges from ten to forty. In this sense, researchers describe
the scale as ordinal in nature, that is, rankings range from Strongly Disagree to Strongly Agree (360).

As with most measures, there are both advantages and limitations to the Likert scale. According to Isaac and Michael, they are most useful in behavioral research, offering greater variance (148). Improved validity and reliability, increased level of measurement and greater efficiency in handling data are the positive aspects of scaling, in general (Monette et al., 2002, 356-357). Providing respondents with a range of options along with yielding ordinal data which is superior to nominal data and the relative ease of construction speak to the popularity of the scale. They are simple and relatively easy to utilize. The downside to this instrument, however, is that of the response set or response bias: “. . . the tendency of some people to answer a large number of items in the same way (usually agreeing) out of laziness or a psychological predisposition” (Neuman 2004, 128). There are a number of strategies designed to correct some of these potential errors, for instance, altering the response patterns in varying items. At the same time, composite scores offer an overall picture of a respondent but do not provide how that individual may have responded to a particular question without exhaustively retrieving singular item answers.

Construction of a Likert scale requires much work and revision. In fact, Monette et al. offer this caution: “The major danger in the uninformed use of scales is that a poorly constructed measuring device will provide false or misleading information about the world” (Monette et al., 2002, 374-375). Their advice is to employ already developed scales, tested and pre-tested for both validity and reliability, rather than merely attempting to independently manufacture such a measure (376). In the event that
researchers construct their own scale, it is imperative to avoid the use of nondiscriminating items: “Nondiscriminating items are those that are responded to in a similar fashion by both people who score high and people who score low on the overall scale” (Monette et al., 2002, 361). Computation of these items yields a discriminatory power score (DP), allowing the researcher to identify and remove problematic items. Detailed descriptions of the methodology for determining the discriminatory power score are beyond the scope of this review but can be found in the literature (Wolfer, 2007, 529-533; Monette et al., 2002, 360-362). A critique of some of the myths about Likert scales with corresponding corrections is also available in the literature (Carifio and Perla, 2007).

Intriguingly, cultural considerations may play a prominent role in the implementation and outcomes of the Likert scale. McQuiston et al. utilized this instrument with immigrants who had recently come to this country in an effort to determine their degree of knowledge about AIDS (McQuiston et al., 2002). Respondents had all received 7 weeks of training relative to AIDS. At the same time, they administered a more qualitative measure, viz., interviewing respondents, asking them to respond to vignettes based on the training material. Discrepancies appeared in these two measures; the Likert scale did NOT reflect the degree of knowledge of the training material that the qualitative interviews yielded. They concluded:

“Findings from this study suggest that Likert-type scales should not be used with recently arrived Mexicans. Furthermore, when using quantitative measures in cross-cultural research, we commend qualitative strategies to ascertain whether quantitative measures accurately reflect knowledge” (McQuiston et al., 2002, 270-271.)
Given the cross-cultural nature of the Migrant Education Program, including the backgrounds of the recruiters, it would seem that a careful reading of this research is paramount.

When it comes to self-appraisals of job performance, the Likert scale is a standard tool of measurement (Meyer, 1980; Fox and Dinur, 1988; Farh et al., 1988; Williams and Leavy, 1992; Cheung, 1999; Miller and Cardy, 2000). In these studies, it is evident that the summated scales implemented undergo exhaustive mathematical calculations in order to draw preliminary conclusions. Data bases and computer software packages are capable of performing such functions.

In an employee’s assessment of self relative to job performance, there is no small measure of interest and ongoing research. Meyer’s pivotal article found that employees had typically inflated views of their functioning, particularly when compared to the performance of others (Meyer, 1980). When asked to confidentially rate their performance relative to their colleagues, individuals saw themselves as doing better than almost 80% of their peers (Meyer, 1980, 292). These findings applied to various groups of employees, both white collar and blue collar. In research terms, the leniency error (tendency to over-rate oneself) is powerful.

Other researchers had similar results but another key study determined that when employees learned that their self-evaluations would be compared with independent criteria, they became more accurate (Mabe and West, 1982). Prior experience in self-appraisal, detailed instructions on that format and the awareness that their evaluations would be cross-checked with other measures served to enhance convergence between self-appraisal and supervisory appraisal. Still others have found that performance self-
appraisals are notoriously discrepant when compared to ratings of others within the organization, e.g., superiors, peers, or subordinates (Harris & Schaubroeck, 1988; Landy and Farr, 1980).

On the other hand, a body of research exists which illustrates the value of self-appraisals when accounting for key factors. Under certain conditions, the use of self-appraisals was found helpful in personnel selection: “These findings favor the notion that individuals possess the capability to reliably evaluate themselves in a manner similar to that of others and in a way that can predict subsequent performance” (Fox and Dinur, 1988, 590). In a study of university professors and their supervisors, there was a great deal of congruence found between self-evaluations and supervisory evaluations leading the researchers to conclude that the following conditions could contribute to this agreement: a management style which is participatory in nature; a setting in which employees work independently; and an organization in which there are clear definitions of performance criteria (Fahr et al., 1988). Similarly, supervisory ratings and employee ratings were in greater synchrony when employees embraced the organizational culture, understood the evaluation process and comprehended job expectations, termed Perceived System Knowledge (Williams and Levy, 1992). High ratings on this dimension translated into perceptions of self that more accurately reflected supervisory ratings.

Exploring the greater complexities of the degree of match between self-appraisal and other-appraisal, Yammarino and Atwater developed a model of self perception accuracy in which they described 3 categories: the over-estimator, the under-estimator, and the accurate estimator (Yammarino and Atwater, 1993). They described these groups in the following way:
“First, accurate estimators are those focal individuals whose self-ratings are in agreement with the ratings of the relevant others. Second, over-estimators are those focal individuals whose self-ratings are significantly inflated above the ratings of the relevant others. Third, under-estimators are those focal individuals whose self-ratings are significantly deflated below the ratings of the relevant others” (Yammarino and Atwater, 1993, 233).

When they discuss self-perception accuracy, they are referring to the degree of agreement between the self-appraisal and the appraisal of other personnel, viz., superiors, subordinates, peers, or customers (232). Interestingly, over-estimators (those who rated themselves higher) had ratings from others that were typically lower. Under-estimators (those who rated themselves lower), however, had higher ratings from others, leading the researchers to conclude that perceived humility or the absence of arrogance results in more positive perceptions (240). Consequently, these researchers place self-appraisal in the context of the relationship system as they consistently situate self-appraisal in the context of ‘with respect to whom.’ Identifying discrepancies between self and other ratings can lead to employee counseling and training since there are clear implications for the entire organization. Over-raters result in diminished organizational and individual outcomes while accurate estimators lead to enhanced outcomes for the organization and the employee. They conclude: “In general, the more accurate individuals’ self-perceptions are, the greater the likelihood of enhanced outcomes for the individuals and the organizations of which they are a part” (244).

In conclusion, Likert scales are a valid and reliable tool in the evaluation process of employees. At the same time, self-assessments are subject to leniency error in that employees tend to view themselves more favorably than others. The research indicates, however, that accounting for certain factors can mitigate this effect. Cultural considerations also evidently play a major part in how individuals respond to this item.
However, under the right conditions and perhaps coupled with qualitative measures such as interviews, a greater congruence between employee self-evaluations and supervisory evaluations can occur. When that happens, the individual employee and the organization benefit.
References


References