Building a Better Mouse Trap: Item Characteristics Associated with Rating Discrepancies in Multi-source Assessment Instruments

Robert B. Kaiser
Kaplan DeVries Inc.

S. Bartholomew Craig
North Carolina State University

Abstract

The authors hypothesized that the low level of convergence characteristic of 360-degree performance ratings is partly a function of the quality of the items used in these instruments. Considering rating items as "stimuli" to raters, three linguistic characteristics are identified that can cause raters to attach different meanings to the same item. In turn, these different interpretations can lead to discrepant ratings from multiple judges of the same target. The results of this study indicate that inter-rater reliability is lower for items that describe complex abstractions of behaviors and for items that refer to more than one discrete behavior. By contrast, inter-rater reliability is higher for items that describe less descriptively complex dispositional qualities and those that are grammatically complex enough to provide contextual cues yet focus on only one specific behavior. Thus, the authors of multi-rater instruments can enhance the quality of feedback provided with their tools by paying more attention to how the items are written.

The emerging research base on 360-degree feedback or multi-source assessment (MSA) ratings paints a discouraging picture. Specifically, analyses of the psychometric properties of these measures indicate that the ratings are plagued with measurement error. Not only are inter-rater reliability estimates dubiously low (Conway & Huffcut, 1997; Gregarus & Robie, 1998; LeBreton et al., 2001), but as much as 55% of the variance in the latent structure of MSA ratings can be attributed to idiosyncratic rater effects (Scullen et al., 2000). These data suggest that MSA ratings tell us more about the people who provided ratings than the people they were assessing.

One reason for the lack of convergence between multiple raters of the same individual may be that the items on MSA instruments are written in such a way as to invite different interpretations. Although rules of thumb for writing survey items have existed for some time (e.g., Crocker & Algina, 1986), we are aware of no published empirical research on the relation between the linguistic characteristics of rating items and their measurement properties. In this study, we empirically assessed the psychometric effects of the following item characteristics: syntax (number of linguistic elements), "multi-barreledness" (referring to more than one behavior), and degree of behavioral abstraction (versus specificity).

Theoretical Argument

MSA ratings can be thought of as raters' responses to presented stimuli (items). To respond to the item, it is assumed that raters use the stimulus as a cue to instigate an information retrieval

---

Inquiries about this research may be directed to Rob Kaiser at: rkaiser@kaplandevries.com.

Presented in R.B. Kaiser and S. Bartholomew Craig (Co-Chairs) *Beneath the Numbers: Factors Influencing the Psychometrics of Multi-source Ratings*, at the annual American Psychological Association Convention, Chicago, IL, August 23, 2002.

We are grateful to the Center for Creative Leadership for providing the multi-source assessment data and to Stéphane Brutus and Jeff Facteau for providing the Subject Matter Expert judgments of the linguistic characteristics of items.

process. Schemata that summarize the performance of the rating target stored in long-term memory are activated and compared to the standards of the rating scale (either "how often does the person do this…" or "how well does the person do that…"). Accordingly, the rater then assigns a scale value to represent the frequency or quality of the target's performance of the behavior connoted by the item.

Several reasons have been offered as to why two different raters of the same individual may provide discrepant ratings. These explanations tend to fall into one of three categories: raters have different opportunities to observer samples of a target's behavior, raters selectively attend to different aspects of the target's performance, and raters attach different levels of importance to the same observed behavior (Borman, 1997; Murphy & Cleveland, 1995; Tsui & Ohlott, 1988). Each of these explanations assumes that the cause of rating discrepancies lies in differences in the content of the raters' conception of the rating target. We would add to this list a factor having nothing to do with raters, but rather is a function of the quality of the stimulus used to elicit ratings: the linguistic characteristics of the items.

Survey method researchers have recently demonstrated how question wording, format, and context can have dramatic effects on the data obtained (e.g., Schwartz, 1999). What this body of research makes clear is that all items are not created equal—some are leading whereas others are neutral, some are precise whereas others are vague, some are concise whereas others are verbose. We contend that at least some of the lack of convergence between raters in MSA contexts is because the linguistic characteristics of some items render the stimulus ambiguous and cause raters to, in effect, attach different meanings to the same cue.

Item Linguistic Properties

In this study, we examine the influence of three linguistic characteristics of items on rating convergence: syntax (number of linguistic constituents; Chomsky, 1965), "multi-barreledness" (referring to more than one behavior), and degree of behavioral abstraction versus specificity.

The most basic component of Universal Grammar (Chomsky, 1986) is syntax. Syntax refers to the structure of sentences, which are combinations of elements called constituents (Chomsky, 1965). Constituents represent strings of words with a certain degree of internal cohesions, with certain formal properties. Put simply, syntax is the manner in which words are ordered and grouped together into constituents. The number of constituents in a given sentence is an indication of how long and complex the text is. Ceteris paribus, items with less constituents require less cognitive resources to process and are thus easier to comprehend (e.g., "This manager delegates important tasks"). More complex items, being harder to comprehend, are more vulnerable to differential interpretation (e.g., "This manager focuses more on managing and directing other people to accomplish a task than on getting personally involved and doing everything the work group does himself").

Another aspect of linguistic complexity refers to the number of behavioral referents contained in a given sentence. That is, how many distinct behaviors are described by the item text (e.g., "This manager praises people for a job well done" vs. "This manager provides feedback—praise for a job well done and constructive criticism to help employees improve"). Often referred to as "multi-barreled," items that describe more than one specific behavior pose a dilemma to the rater: How to arrive at one summary rating? Does the rater go through a routine of cognitive

gymnastics by assigning a numerical value to each behavioral referent and then arrive at an average? Or does the raters zero in on only one of the behaviors? What if multiple raters of mulit-barreled items attend to different behavioral referents?

A deeper linguistic property of rating items is their degree of specificity in describing the job-relevant behavior to be evaluated. Does the text concretely describe a narrowly defined and observable behavior in some contextualized frame of reference (e.g., "This manager is receptive to negative feedback about how she treats subordinates")? Or does it require the rater to either infer some internal characteristic (e.g., motives), summarize across a wide-range of behaviors, or otherwise form an abstract judgment ("This manager is defensive")? Items that are less specific require raters to impute definitional boundaries to delimit the information search. Climbing the ladder of abstraction is a highly subjective journey that introduces ample opportunity for idiosyncratic tendencies to influence the rating process.

Hypotheses

As noted at the outset, one troubling finding in the MSA literature is the widespread lack of convergence in the ratings furnished by multiple raters of a common target. We have outlined a relatively neglected explanation for this divergence in the linguistic properties of rating items. This view holds promise for improving the measurement quality of MSA ratings because it considers the design of the rating instrument, which can easily be changed or modified, rather than the cognitions of the raters. To ascertain the viability of this approach to improving the psychometrics of MSAs, we tested the following hypotheses:

1. The number of constituents in a given rating item (syntax) will be negatively related to inter-rater agreement and reliability.

2. The number of behavioral referents in a given rating item ("multi-barreledness") will be negatively related to inter-rater agreement and reliability.

3. The degree of behavioral abstraction (versus specificity) in a given rating item will be negatively related to inter-rater agreement and reliability.

Method

Sample

The data for this study were collected from managers (GMs through senior executives) who participated in various leadership development courses offered at a non-profit training institute between 1992 and 1997. Target managers were asked to fill out a self-rating form of a 360-degree feedback questionnaire and to ask several coworkers to also complete the form prior to beginning the courses. The target managers represented firms in the manufacturing, financial services, insurance, wholesale trade, transportation, communications, and utility industries. Data were randomly extracted from a larger database such that, for each target, two superior ratings, two peer ratings, and two subordinate ratings were available. This yielded performance ratings for a total of 1,404 target managers.

MSA ratings were collected with Benchmarks (Lombardo & McCauley, 1994), a 360-degree instrument developed from a series of studies of executive development. An overview of the

content considerations and construction of the instrument can be found in McCauley and Lombardo (1990). Benchmarks contains two general sections containing a total of 164 items. For the purposes of the present study, only the 106 items in the first section were used (cf. Fleenor, McCauley, & Brutus, 1996). These items consist of a stem containing a behavioral description of a skill, competency, or capacity, and the target manager is rated on a five-point scale (low to high) for the degree to which the item is characteristic of him or her. This instrument has been subjected to multiple validation efforts (for a review, see McCauley & Lombardo, 1990) and has received favorable overall evaluations as a reliable and valid measure of important aspects of leadership related to executive development (e.g., Zedeck, 1995).

Because items on this instrument were selected on a relatively rigorous psychometric basis (e.g., factor analyses, internal consistency analyses, criterion-relatedness, SME judgments of content appropriateness), the present database may suffer from range restriction. That is, items with complex syntax and multiple behavioral referents may be underrepresented in this study compared to the items commonly used in organizations.

Measures

The measures of the linguistic characteristics for each rating item were provided to us by Brutus and Facteau (2002) who used the data in a separate study relating item characteristics to error variance (defined as the amount of variance in an item accounted for by its underlying construct). Brutus and Facteau had two groups of experts assess the linguistic characteristics of the Benchmarks items. Item syntax (number of constituents) was evaluated by two graduate students in the linguistics program at a Canadian university. "Multi-barreledness" and "behavioral abstraction" were evaluated by nine people with Ph.D. degrees in I/O psychology and an average of 11.6 years of experience in performance assessment.

Syntax. Syntax ratings were made for each item according to Chomsky's (1965) protocol. Items that contained only a verb followed by an object of that verb are coded with a 1 in this scheme, representing the fewest number of constituents possible in a grammatically complete sentence (e.g., "this manager [[emphasizes$_{verb}$] [cooperation$_{obj}$]]"). Items that contain one optional complement or verbal modifier—which add supplementary information about the subject described by the verb but are not necessary for grammatical completeness—were assigned a rating of 2 (e.g., "this manager [[responds$_{verb}$] [*with sensitivity*$_{opt1}$] [to the feelings of others$_{obj}$]]"). In a similar way, another point is counted for each additional modifier or adjunct (i.e., optional complement) that is added to the obligatory object. This coding scheme assumes that the syntactic complexity of rating items increases with each constituent (Chomsky, 1965). The two raters exhibited an adequate level of agreement for this index (r_{wg} = .83). Their ratings were therefore averaged to create a composite index of "syntax" (M = 3.41, SD = 1.38).

Number of behavioral referents ("Multi-barreledness"). The psychologists were asked to rate the number of behaviors referred to in each item. Because only one item referred to more than two discrete behaviors, a dichotomous coding (1 = uni-dimensional, 2 = multi-dimensional) was used. The inter-rater agreement for this index was acceptable (r_{wg} = .79). Just under half of the items (43%) were judged by all nine raters to be unidimensional; only seven (6.6%) were unanimously deemed multidimensional. A composite "multi-barreledness" index was created by computing the average ratings across all nine judges for each item (M = 1.28, SD = .34).

Behavioral abstraction/specificity. The degree of abstraction versus behavioral specificity was rated by the nine I.O psychologists. They were first presented with a brief definition of this

dimension and asked to rate each item on the following scale: 1 = not specific at all, 2 = slightly specific, 3 = moderately specific, 4 = quite specific, 5 = extremely specific. Inter-rater agreement for this rating across the nine judges was somewhat low (r_{wg} = .68). Reliability was enhanced by computing the average rating across all nine judges to derive an abstraction index for each item (*M* = 3.21, *SD* = .56; this scale was reflected so higher scores would indicate more abstraction).

Psychometric properties. Because of the alarming finding that MSA ratings say more about the raters than the rated, we used measurement criteria having to do with rating similarity. Two conceptually and empirically distinct kinds of similarity indices were used: inter-rater reliability and inter-rater agreement (cf. Fleenor, Fleenor, & Grossnickle, 1996; LeBreton et al., 2001). Inter-rater reliability coefficients reveal the similarity or consistency of the pattern of responses, or the rank-ordering of responses between two or more raters, independent of the level or magnitude of those ratings. On the other hand, to index the degree that ratings are similar in level or magnitude, an inter-rater agreement technique is required. As Kozlowski and Hattrup (1993) noted, an inter-rater agreement index is designed to "reference the interchangeability among raters; it addresses the extent to which raters make essentially the same ratings" (p. 163). Thus, inter-rater agreement assesses similarity without controlling for idiosyncratic tendencies to rate severely or harshly whereas inter-rater reliability effectively controls for elevation bias in the estimation of rating similarity (James, Demaree, & Wolf, 1984; 1993).

The inter-rater reliability for each of the 106 items was calculated separately for superiors, peers, and subordinates using one-way random effects intraclass correlation coefficients to estimate the reliability of the mean of 2 raters in each group (i.e., ICC[2,2]; cf. Bartko, 1976; Shrout & Fleiss, 1979). The inter-rater agreement for each item was also calculated separately for superiors, peers, and subordinates. James' r_{wg} statistic (James et al., 1984; 1993) was used for this index. Following James et al., all negative r_{wg} values were reset to 0. Descriptive statistics for each measure of rating similarity are presented in Table 1.

Table 1. Descriptive statistics for inter-rater reliability and agreement indices by rating source.

	ICC (2,2)		r_{wg}	
Rating Source	*M*	*SD*	*M*	*SD*
Superiors	.41	.09	.75	.03
Peers	.31	.08	.71	.04
Subordinates	.32	.08	.69	.05

Results

The correlations between the three linguistic characteristics of the items and the six estimates of rating similarity are presented in Table 2. Three trends stand out: first, none of the correlations with inter-rater agreement are significant. This may be due to the relatively low amount of variance in these indices. (The variance observed for the r_{wg} values was more than 30% lower than the variance of the ICCs.) Second, none of the correlations with the inter-rater reliability for superior ratings are significant—the link between item characteristics and inter-rater reliability is only present in peer and subordinate ratings. Finally, all of the correlations with inter-rater reliability are in the expected direction, but of generally low magnitude.

Table 2. Pearson correlations between linguistic characteristics and rating similarity.

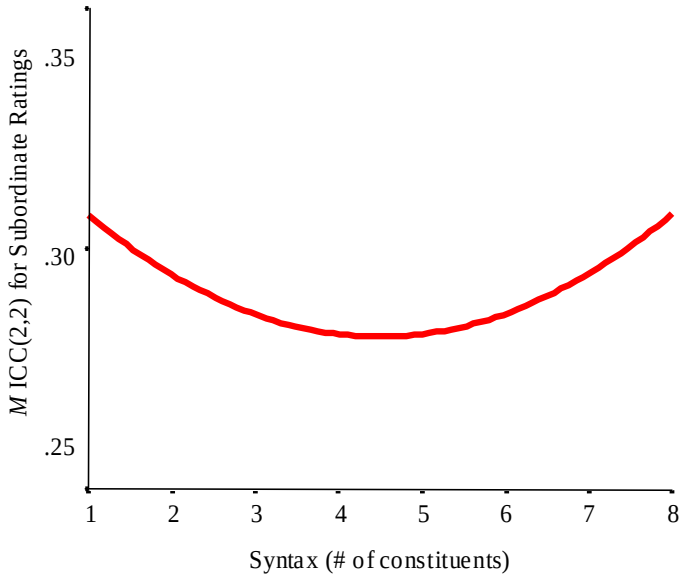
Linguistic Characteristic	Inter-rater Reliability (ICC[2,2])			Inter-rater Agreement (r_{wg})		
	Superiors	Peers	Subordinates	Superiors	Peers	Subordinates
Syntax	-.05	-.09	-.11	.06	.10	.03
Multi-Barreledness	-.04	-.19*	-.20*	.03	.03	.10
Abstraction	-.05	-.12	-.17$^{\downarrow}$	-.08	-.07	-.05

* $p < .05$; $^{\downarrow}$ $p < .10$

Of the three linguistic characteristics under investigation, "multi-barreledness" was the strongest correlate of low inter-rater reliability. Behavioral abstraction had a marginal negative impact on rating similarity for subordinates.

Inspection of the 18 scatterplots relating each predictor with each criterion suggested that there was a nonlinear relationship between syntax and rating similarity. We tested this by computing quadratic and cubic regression models using syntax to predict the three inter-rater reliability coefficients. The results indicated "U-shaped" functions of the following effect sizes: predicting superior ICCs, $R = .13$ ($p > .10$), peers ICCs, $R = .17$ ($p > .10$), and subordinates ICCs, $R = .22$ ($p < .10$). The effect for subordinate ratings is presented in Figure 1.

Figure 1. Curvilinear relationship between syntax and subordinates' inter-rater reliability.



This curvilinear relationship indicates that raters tend to show greater agreement in ratings on the most simple and the most complex items—those of moderate complexity show the least agreement. Content analysis of items at each interval along this continuum provided clues for interpreting this result. It appears that the relatively simple items (low syntax) reflect general personality characteristics, dispositions which raters have been shown to show substantial agreement on in the personality literature (e.g., Funder & Sneed, 1993). For instance, two of

lowest syntax items were “Has a pleasant disposition” and “Has a good sense of humor.” The most complex items (highest syntax) appear to be relatively straight forward, but also include a good deal of clarification and contextualization. These items tended to receive relatively high specificity ratings too. For instance, “Can settle problems with external groups without alienating them” and “Tries to understand what other people think before making judgments about them.” Items in the middle of the continuum of syntax ratings were less clear-cut: in many cases, they were double-barreled (e.g., “Is a visionary able to excite other people to work hard,” “Quickly masters new vocabulary and operating rules needed to understand how the business works”). Several of the items were conceptually complex, describing the interpenetration of seeming polar opposites (e.g., “Is self-confident, but has a healthy humility” and “Can be close enough to others to be empathetic and distant enough to be objective”).

Finally, we assessed the multivariate effects of these three item characteristics on inter-rater reliability. Three separate regression models were constructed—one for each rating source. Each model reflected linear relationships for “multi-barreledness” and behavioral specificity and quadratic relationships for syntax. The results revealed effects that were practically noteworthy, yet not “statistically significant” by conventional standards because of the low statistical power afforded by an *n* of 106 items. Specifically, the total amount of variance in inter-rater reliability accounted for by the set of item characteristics was $R^2 = .02$ for superiors, $R^2 = .07$ for peers, and $R^2 = .09$ for subordinates.

Discussion

The growing body of research on MSA ratings suggests that there is ample room to improve the quality of the data provided by these instruments. In particular, the low level of convergence routinely observed between multiple raters of a common target is alarming. The current state-of-the-art in 360-degree assessments in organizations provides more information about raters than about the individuals who are receiving the feedback (e.g., Scullen et al., 2000), feedback that is ostensibly about their performance. In this study, we sought to determine if one key to improving the quality of the data provided by the MSA process might be in the design and development of items contained in the rating instruments. Our results offer a cautious glimmer of hope, at least so far as how these tools can be created to minimize idiosyncratic rater effects and maximize valid information about the job performance of the person being rated.

In interpreting the results of this study, it is important to bear in mind two facts. First, we only included data from one instrument containing a total of 106 items. Thus, most statistical tests had only limited power to detect meaningful effects. It is probably more instructive to interpret effect sizes and directions than to go by the traditional “tabular asterisks” denoting magical *p* values (e.g., Cohen, 1994). A second fact is that the instrument used in the current study was developed under relatively rigorous conditions. Our experience with some commercially available instruments and many of those created in-house for organizational use suggests that the degree of psychometric craftsmanship that went into the present instrument is far above that typically found in practice. To the extent that this is true, the present data likely suffer from range restriction in both predictors (item characteristics) and criteria (inter-rater agreement and reliability). To that end, our results probably underestimate the deleterious effects of poorly written items. We intend to extend this study by including several other instruments—both commercial and proprietary—to test our logic.

To the extent that our interpretation of the present data is valid, there are important practical

implications for those who create MSA tools. First, it is important when crafting items to think not as a student of behavior but rather as a typical employee who is asked to evaluate a coworker. This includes consideration of the ways coworkers—not scientists—think about each other, the degree of sophistication in the mental maps coworkers use to encode perceptions of performance, and limits to the amount of cognitive resources busy managers have available to devote to a rating task that is far removed from "the bottom line."

The present research indicates that there are several kinds of "good" items. One type is the short and simple description of general tendencies. Raters tend to show higher levels of agreement on these dispositional statements (e.g., "Has a warm personality"). One caveat here, though, it that this kind of item may not be specific enough to be of prescriptive value to the feedback recipient. It is probably best to use this type of item sparingly, in favor of the more behaviorally precise kinds of items.

This leads to another type of useful item, the one that is very behaviorally precise and also includes enough context or additional information to delimit the domain of interest. These are the items in our study that had very high syntax ratings and very high specificity ratings. For instance, rather than refer to how the manager behaves towards people in general, specify which people. Rather than describe the behavior by itself, include contextual parameters (e.g., follow "Tries to understand what other people think" with the modifier "before making judgments about them"). And rather than just describe what is done, consider elaborating how it is done (e.g., "Can settle problems with external groups without alienating them").

Finally, avoid writing items that refer to more than one behavior or activity. If an item contains multiple discrete behaviors, skills, or actions, then it is likely that different raters will attend to different aspects of the sentence. Moreover, the feedback recipient is hamstrung when confronted by low ratings on an item like "Is a visionary able to excite other people to work hard": is the performance issue in not having a clear view of the future or in underdeveloped influencing skills? In a similar vein, complex items that refer to the integration of opposing behaviors (e.g, "Is confident yet humble") appear to confuse raters. Consider creating two distinct items from these components or choose the one that is most relevant to effective performance.

One may claim that this last piece of advice isn't anything new because, "we've known for years not to write multi-barreled items." To that, we'd point out that despite this long-standing admonition, in the present commercially successful (and favorably peer-reviewed instrument; Zedeck, 1995), nearly 7% of the items were unanimously deemed by nine expert judges to refer to more than one discrete behavior. Moreover, we have seen several instances where instrument authors, in an effort to reduce the length of a survey, have combined two or more "critical concepts" into one item. It seems worth saying again, then, "don't use rating items that describe more than one behavior!"

References

Bartko, J.J. (1976). On various intraclass correlation reliability coefficients. Psychological Bulletin, 83, 762-765.

Borman, W.C. (1997). 360° ratings: An analysis of assumptions and a research agenda for evaluating their validity. Human Resource Management Review, 7, 299-315.

Brutus, S. & Facteau, J.D. (2002). Short, simple, and specific: The influence of item design characteristics in multi-source assessment contexts. Manuscript under review.

Chomsky, N. (1965). Aspects of the theory of syntax. Cambridge. MIT Press.

Chomsky, N. (1986). Knowledge of language: Its nature, origin, and use. New York: Praeger.

Cohen, J. (1994). The Earth is round ($p < .05$). American Psychologist, 49, 997-1003.

Conway, J., & Huffcutt, A. (1997). Psychometric properties of multi-source performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. Human Performance, 10, 331-360.

Crocker, L. & Algina , J. (1986). Introduction to classical and modern test theory. Orlando, FL: Harcourt Brace Jovanovich.

Fleenor, J. W., Fleenor, J. B., & Grossnickle, W. F. (1996). Inter-rater reliability and agreement of performance ratings: A methodological comparison. Journal of Business and Psychology, 10, 367-380.

Fleenor, J.W., McCauley, C.D., & Brutus, S. (1996). Self-other rating agreement and leader effectiveness. Leadership Quarterly, 7, 487-506.

Funder, D.C. & Sneed, C.D. (1993). Behavioral manifestations of personality: An ecological approach to judgmental accuracy. Journal of Personality & Social Psychology, 64, 479-490.

Greguras, G. J., & Robie, C. (1998). A new look at within-source inter-rater reliability of 360-degree feedback ratings. Journal of Applied Psychology, 83, 960-968.

James, L. R., Demaree, R., G. & Wolf, G. (1984). Estimating within-group inter-rater reliability with and without response bias. Journal of Applied Psychology, 69, 85-98.

James, L. J., Demaree, R. G., & Wolf, G. (1993). r_{WG}: An Assessment of within-group inter-rater agreement. Journal of Applied Psychology, 78, 306-309.

Kozlowski, S. W. J., & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. Journal of Applied Psychology, 77, 161-167.

LeBreton, J.M., Burgess, J.R.D., Atchley, E.K., Kaiser, R.B., & James, L.R. (2001). True or false? Different sources of performance ratings don't agree. In R.B. Kaiser, & S.B.Craig,

Modern Analytic Techniques in the Study of 360° Performance Ratings, presented at the 16th annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

McCauley, C. D. & Lombardo, M. (1990). Benchmarks: An instrument for diagnosing managerial strengths and weaknesses. In K. E. Clark & M. B. Clark (Eds.), Measures of Leadership. West Orange, NJ: Leadership Library of America.

Murphy, K. R., & Cleveland, J. N. (1995). Understanding performance appraisal: Social, organizational, and goal based perspectives. Thousand Oaks, CA: Sage Publications.

Schwartz, N. (1999). Self reports: How the questions shape the answers. American Psychologist, 54, 93-105.

Scullen, S.E., Mount, M.K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. Journal of Applied Psychology, 85, 956-970.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86, 420-428.

Tsui, A.S., & Ohlott, P. (1988). Multiple assessment of managerial effectiveness: Inter-rater agreement and consensus in effectiveness models. Personnel Psychology, 41, 779-803.

Zedeck, S. (1995). [Review of Benchmarks]. In J. Conoley & J. Impara (Eds.), The twelfth mental measurements yearbook (Vol. 1, pp 128-129). Lincoln, NE: Buros Institute of Mental Measurements.