

Using the SAS® System to Study the Gender and Level Measurement Equivalence of a Multi-rater Survey

Jim Penny, Center for Creative Leadership, Greensboro, NC

ABSTRACT

This research used logistic regression to model item responses from a popular 360-for-development survey. The survey contained 57 items on 11 scales. The model used gender and rater group to identify items that exhibited differential item functioning (DIF). The rater groups were self, boss, peer, and direct report. The sample consisted of 752 survey families where a survey family consisted of a matched set of four surveys: one self, one boss, one peer, and one direct report. The sample of 3008 surveys contained 76% male and 24% female raters. The procedure to flag items exhibiting differential functioning used effect size computed from Wald chi-square statistics rather than statistical significance, resulting in fewer flagged items.

Three items exhibited rating anomalies due to the gender of the rater or ratee. Twelve items exhibited DIF attributable to rater group. In each instance, the apparent effect of the DIF was small. An examination of the maximum likelihood parameter estimates suggested the rater group DIF was the possible result of hierarchical complexity. The DIF due to gender conformed to expectations of gender-related stereotypical interpretations of item text. This research further suggested that DIF due to environmental complexity could be a naturally occurring phenomenon in some 360-assessment, and that the interpretation of some 360-feedback might need to include the potential for such DIF to exist.

INTRODUCTION

There has been a veritable explosion in the use of 360-assessment, a form of multi-rater assessment for managerial development in organizations. The process of 360-assessment involves providing managers with feedback from four sources: (1) the manager's boss, (2) the manager's subordinates or direct reports, (3) the peers or the customers of the target, and (4) the self. Although the notion of receiving multi-source feedback is not new, at least one premise of multi-rater methodology remains unresolved: Does the 360-process produce a similar measure from different rater groups just as a measuring tape produces a similar measure with different carpenters? Alternatively, does the 360-methodology provide an equivalent measure with each rater group?

In addition to the increase in the use of 360-assessment, there continues to exist the question of whether or not women receive fair assessments of performance in the workplace, or do a variety of psychological and sociological biases influence the results of 360-methodology when applied to women? That is, does 360-methodology provide an equivalent measure for both men and women, or do the gender biases that sometimes accompany performance appraisals function to influence the manner in which some raters interpret some items? Moreover, might there exist a potential interaction between rater group and gender of the ratee or between the gender of the rater and the gender of the ratee?

DIFFERENCES AMONG RATER GROUPS

Discussions of the differences between self and others' ratings sometimes arise during 360-feedback sessions (Van Velsor & Leslie, 1991), making the existence of measurement equivalence important to interpretations of 360-feedback. It is possible, if not expected, that a feedback recipient will receive low ratings in one area from one rater source while receiving high ratings in that same area from another rater group. A manager, for example, may be interpersonally skilled with bosses yet cold and aloof with direct reports. This manager, therefore, could receive high

ratings on interpersonal skills by the boss while receiving low ratings on this dimension by direct reports. However, in the interpretation of the between group difference, there is the underlying assumption that the raters are responding to their perceptions based on their observations of behaviors exhibited by the manager, and that two raters with similar observations will respond similarly to a given item even though the raters may occupy positions of different levels.

CONTINGENCY THEORY

Contingency theories of leadership (Fielder 1978; Fielder & Chemers, 1982) suggest that disparate ratings can be an indication of an effective manager, and that gaps in the perspectives between groups of raters are often a naturally occurring phenomenon of management. Moreover, Yukl (1981, pp. 99-119) suggested that managers often change their behavior to fit particular situations, and, following this line of argument, managers who behave differently toward different groups of co-workers may receive disparate ratings from members of those groups. Hence, between group differences may be an acceptable outcome for some managers.

Concomitant with the interpretation of group differences in 360-feedback is the expectation that a different interpretation of the item by one group of raters does not contribute substantially to the observed difference, and that the observed difference is only the result of behavioral differences produced by the circumstances of contingency. However, it seems reasonable to anticipate that some items may tap into differences produced by organizational contingency to a greater degree than do some other items. The ratings produced by items influenced by contingency, then, become composite scores comprised not only of an estimate of the managers standing on the trait measured by the survey but also of the degree to which contingency influenced the ratings.

COMPLEXITY THEORY

Jacques (1996) and Jacques & Clement (1994) suggested that the degree of environmental complexity and ambiguity seen by a person within an organization generally increases with rank. That is, a supervisor of the manager is likely to see a more complex and a more difficult to comprehend environment than is the direct report of the manager. For example, 360-surveys sometimes contain items that measure the resourcefulness of the manager, and one might argue that the increase in complexity from one level to another could produce different interpretations of what resourcefulness means. Hence, it seems reasonable to anticipate that differences in environment may influence the ratings given on a 360-survey.

As with contingency theory, the interpretation of between-group differences in 360-ratings is posited on the expectation that an observed difference is solely a function of the behavioral differences witnessed by the raters. However, it seems reasonable to anticipate that some items may tap environmental complexity more so than other items. In that event, a rating difference produced by such items may represent a composite of not only the standing of the manager on the trait assessed by the items but of also the degree to which complexity influences the rater's interpretation of that item. One might argue, then, that a manager reviewing 360-feedback could choose to make behavioral changes due not only to the behavioral observations of the raters but also, at least in part, to ratings produced by the anomalous functioning of some items.

DIFFERENCES BETWEEN GENDERS

Although far from conclusive, a condition indicative of the

complex role gender plays in society, many studies have examined the influence of gender on ratings of managerial effectiveness over the past twenty-five years. Some studies have demonstrated statistically significant differences attributable to gender of the ratee (Bartol & Butterfield, 1976; Jacobson & Effertz, 1974; Rosen & Jerdee, 1974; Schmitt & Lapin, 1980). Other studies have failed to produce such differences (Pulakos & Wexley, 1983; Thompson & Thompson, 1983).

In circumstances where a woman functions in a role often associated with men, one might expect to find differences attributable to the interaction of "role gender" and gender of the person filling the role. For example, a woman working as a firefighter might find herself at risk to receive performance reviews that carry not only an assessment of her performance but that also carry the influence of the interaction of her gender with the "gender" of the job. Of course, it stands to reason that men filling roles often associated with women will experience similar bias in performance reviews. Bartol & Butterfield (1976) and Rosen & Jerdee (1974) identified statistically significant interactions between role gender and person gender; however, Jacobsen & Effertz (1974) and Mobley (1982) failed to identify such interactions.

The influence of gender is likely a composite of many factors, some of which may have small effects until they exist in concert with gender. Moreover, one could argue that raters are more likely to remember the gender of the manager long after forgetting particular exemplars of either good or bad behaviors. Such biases attributable to gender may influence ratings more than other factors and behaviors. For instance, Nieva & Gutek (1980) suggested that level of qualification, level of performance, degree of inference resulting from the ratings, and sex-role incongruence may each explain a portion of rating variability. Other explanatory factors also have arisen in the study of gender differences in managerial ratings. Cash, Gillen, & Burns (1977) suggested that some raters attribute a man's success to ability while attributing the success of a woman to effort and luck. Greenhaus & Paurasuraman (1993) confirmed those findings, though only for women in the highest performance levels. At moderate levels of performance, they found that raters were likely to use ability to explain a woman's success.

In addition, one might also suggest that stereotypical behaviors and biases may influence performance ratings. Noe (1988) and Powell (1988) gave evidence to suggest that negative stereotypes against minorities and women can have a substantial impact on ratings of performance and effectiveness. Moreover, Martell (1991) found that if there existed less time to make an assessment of managerial performance, the performance of men was likely to receive higher ratings than comparable performances by women. Maurer & Taylor (1994) rendered this finding even more poignant when they demonstrated that the perceived masculinity of the ratee could produce higher ratings. Lastly, Powell & Butterfield (1989) suggested that the definition of "good manager" still carried connotation of masculinity despite the growing population of female managers.

It seems reasonable to anticipate that some items will tap perceptual differences due to gender to a greater extent than will other items, and one might also suggest that particular items may tap particular gender-related biases and either increment or decrement differentially the resulting 360-ratings. In addition, one may ask if the differential functioning of the item is due to the gender of the rater, the gender of the ratee, the interaction of the two genders, all three or some other combination. Moreover, do either or both genders interact with the rater group?

RESEARCH QUESTIONS

This research sought to establish the degree to which differential item functioning attributable to rater group and gender may influence the ratings of a 360-survey. That is, will a given manager receive similar ratings from the boss, a direct report, and a peer if those three other raters have had similar

experiences with the manager? Are there components in item ratings attributable to the gender of the rater or to the gender of the ratee? Are there items that function differently for particular combinations of rater and ratee gender? Is there evidence to support the existence of an interaction between the gender of ratee and the rater group? Moreover, if such items exist, does an explanatory model exist using extant measurement and psychological theory? Finally, if such items exist and if such explanatory models exist, what, then, may be the subsequent implications for the interpretation of the 360-feedback.

METHODOLOGY

This research used logistic regression to detect DIF. Swaminathan & Rogers (1990) first presented this methodology and demonstrated its relationship to the Mantel-Haenszel procedure (Mantel & Haenszel, 1959; Holland & Thayer, 1988). Swaminathan & Rogers (1990) and Clauser & Mazor (1998) have shown that logistic regression is equal in power to the Mantel-Haenszel procedure for the detection of uniform DIF. Moreover, these same authors with Penny & Johnson (1999) have shown that the Mantel-Haenszel procedure may lack sufficient statistical power to detect some instances of nonuniform DIF. However, Rogers (1989), Rogers & Swaminathan (1993), and Swaminathan & Rogers (1990) found that logistic regression procedures likely to have sufficient power to detect non-uniform DIF.

Much of the initial research in the use of logistic regression for the detection of DIF involved the examination of dichotomous items; that is, items with two possible responses, usually 0 and 1. However, logistic regression is easy to extend to polytomous data where the respondent chooses one of an ordered set of responses. Samejima (1969, 1979) presented the Graded Response Model that describes such item responses which are common on 360-surveys. For example, the Graded Response Model describes a Likert-type item using a 5-point scale of 1=Strongly Disagree to 5=Strongly Agree positing a response function for each point on the scale according to

$$P(x = k) = \frac{1}{1 + e^{(-a(\theta - b_{k-1}))}} - \frac{1}{1 + e^{(-a(\theta - b_k))}} \\ = P^*(k) - P^*(k + 1)$$

in which a is the discrimination parameter and b_{k-1} is the threshold parameter. $P^*(k)$ is the item response function that describes the probability that a response is in category k or higher. In this model, the discrimination parameter is constant for all categories of k , and the threshold parameter, b_{k-1} , is the point on the θ -axis where the probability exceeds 50 percent that the response is in the next category. Researchers sometimes call the threshold parameter the "location" parameter.

For the detection of DIF using logistic regression with polytomous data (Clauser & Mazor, 1998), I can adapt the framework of Samejima (1969, 1979) and write the equation

$$P(x = k) = \frac{1}{1 + e^{-z_{k-1}}} - \frac{1}{1 + e^{-z_k}},$$

where $P(x=k)$ is the probability of a response k to a particular item from a respondent of standing θ , and where k takes the values of the Likert-type response scale (Miller & Spray, 1993; Samejima, 1969, 1979; Swaminathan & Rogers, 1990). From this point forward, I will drop the subscript i on z to make the model easier to read.

MODELLING ITEM RESPONSES

Were the existence of DIF not an issue, I would write z as

$$z = \tau_0 + \tau_1 \theta,$$

where θ represents the standing of the ratee on the attribute that the survey measures. The symbols τ_0 and τ_1 represent the intercept and the slope parameters of the logistic regression model; these symbols also represent forms of the discrimination and location parameters of the Graded Response Model. This

logistic model represents the situation where the rater group membership and gender do not influence the item response, and where the only factor that does influence the response is the standing of the ratee on the attribute that the survey measures.

To expand the model to include components to represent effects due to gender and rater group membership, I would write

$$z = \tau_0 + \tau_1\theta + \tau_2g + \tau_3r,$$

where g and r represent gender and rater group membership, respectively, and τ_2 and τ_3 represent the logistic regression parameters for those two classifications. This model describes the instance where only uniform DIF exists. I can define the values for g in the typical 360-survey as {male, female}. Similarly, I can define the values for r as {self, direct report, peer, boss}. Later, to contrast the functioning of an item across these values, I used dummy codes to represent each value.

I can expand this model to accommodate the potential existence of nonuniform DIF by the addition of two more terms to produce

$$z = \tau_0 + \tau_1\theta + \tau_2g + \tau_3r + \tau_4g\theta + \tau_5r\theta,$$

where the two new terms indicate an interaction, respectively, between (a) gender and standing on the attribute that the survey measures, and (b) rater group membership and standing. The symbols τ_4 and τ_5 represent the logistic regression parameters for these two interaction terms, respectively.

To complete the model for this research, I included two additional terms. One term is to indicate the possible interaction between rater group membership and gender; the other term is to indicate the possible three-way interaction of rater group membership, gender, and standing on the attribute that the survey measures. The types of DIF represented by these two terms are uniform and nonuniform, respectively. I can write this model as

$$z = \tau_0 + \tau_1\theta + \tau_2g + \tau_3r + \tau_4g\theta + \tau_5r\theta + \tau_6gr + \tau_7gr\theta,$$

where the symbols τ_6 and τ_7 represent the logistic regression parameters for these two additional interaction terms, respectively.

A CLOSER LOOK AT MODELLING GENDER

Although this model permits the examination of the rater group and rater gender effects, it does not permit the comparison of the gender of the rater to the gender of the ratee. That is, is there evidence to suggest that some items exhibit DIF that is attributable to men rating women, women rating women, and so forth? To examine such an interaction, the model needs to include terms for the two genders in addition to the cross product of those genders. It is possible to add those terms to this model, but to do so leads to a conceptual problem: For the self-rater, do the self-ratings of a male represent a man rating a man or the self rating the self? In a manner of thinking, the self-ratings do represent gender-on-gender ratings, and there may be the occasional self-rater who can step outside of the self and produce gender-on-gender ratings, but it seems far more reasonable that the gender of the self-rater is not as important to the self-ratings as is self-awareness.

Hence, to examine the interaction of rater gender and ratee gender in addition to rater group and gender, I used the model

$$z = \tau_0 + \tau_1\theta + \tau_2r + \tau_3g_s + \tau_4g_r + \tau_5rg_s + \tau_6rg_r + \tau_7g_sg_r + \tau_8rg_sg_r + \tau_9r\theta + \tau_{10}g_sg_r\theta + \tau_{11}g_r\theta + \tau_{12}rg_s\theta + \tau_{13}rg_r\theta + \tau_{14}g_sg_r\theta + \tau_{15}rg_sg_r\theta$$

where θ represents the covariate which is a proxy for standing on the trait measured by the survey, r represents the rater group, g_s represents the gender of the self-rater, and g_r represents the gender of the other rater. The additional terms represent the various interactions that may be important to understanding the functioning of the items. The values of the two gender variables are *male* and *female*. However, the values of the rater group variable no longer include *self*, but only the values *direct report*, *peer*, and *boss*.

STATISTICAL SIGNIFICANCE AND EFFECT SIZE

I used the SAS® System to evaluate this model for the data I collected. The SAS System produced Wald Chi Square statistics to test the null hypotheses that the parameter estimates of τ_0 through τ_{15} were statistically significantly different from 0. It was my anticipation that τ_0 and τ_1 would routinely achieve statistical significance. In addition, it was my anticipation that the parameter estimates of τ_2 through τ_{15} would not routinely achieve statistical significance.

However, I knew that, with number of terms in the model I had chosen, I would be making many statistical tests of significance, and that the experiment-wise Type I error rate could be high. To compensate for the accumulated Type I error rate that could naturally occur in this research and to avoid the complex power analysis (Hsieh, 1989; Whittemore, 1981) of logistic regression that would suggest an appropriate number of subjects to evaluate the model, I decided to use effect size instead of statistical significance. I chose the technique presented in Penny & Johnson (1999) and converted the Wald chi-square statistic to an effect size, w , described in Cohen (1988, ch. 7). The formula that relates the effect size to the sample size is

$$X^2 = nw^2$$

where X^2 is the chi-square statistic, n is the sample size, and w is the effect size. Cohen (1988, ch. 7) used the arbitrary values of .1, .3, and .5 to indicate small, medium, and large effects, respectively. Although these values are arbitrary, Penny & Johnson (1999) found that those values appeared to connote well derived from the Mantel-Haenszel chi-square statistic used to identify DIF, and I used the three values to define four effect ranges to categorize the DIF I discovered with the logistic regression model. These ranges were nil-to-small, small-to-medium, medium-to-large, and large-to-extreme.

CLASSIFICATION OF TYPE OF DIF

After I classified the items by the type of DIF, either uniform or nonuniform, and the source of DIF, I classified them further according to the apparent explanation of the differential functioning. An item influenced by complexity theory should produce maximum likelihood parameter estimates suggestive of a continuum from direct report to boss. For example, a boss may interpret an item that assesses the resourcefulness of a manager differently than would a direct report because the boss, by virtue of working in a more complex and, perhaps, more ambiguous environment, may have a different idea of what actions, and quantities of actions, constitute resourcefulness in a manager. If a manager, boss, peer and direct report have similar assessments of the manager's performance, and that item functions to augment the ratings of the direct report over those of the boss while not altering the ratings from the self and peers, then the item is exhibiting differential item functioning of a type that produces evidence of a continuum.

In this research, I denoted this type of differential functioning as "DR(+) to S/P to B(-)." To save space, I often dropped the "S/P" part of the notation to produce "DR(+) to B(-)." Hence, an item that functioned to produce differentially lower ratings from direct reports would receive the notation "DR(-) to B(+)."

An item influenced by contingency theory should produce maximum likelihood parameter estimates that suggest a contrast of one rater group to all the others. For example, an item that assesses the propensity of a manager to learn the work performed by direct reports may function differently for direct reports than it does for any other rater group. If the item functions to differentially decrement the ratings from the direct report raters, I can describe the functioning using the notation "DR(-) vs. all others." Of course, another item could function to isolate any of the other rater groups, and the ratings from the isolated group could be incremented or decremented resulting in notation such as "P(+) vs. all others" or "B(-) vs. all others."

I classified items influenced by gender into groups suggestive of the gender stereotype tapped by the item by reviewing the text of

the item after reviewing the regression results. After the final categorization, I presented my findings to a review panel of 360-process and 360-content experts who proceeded to challenge my classifications.

DATA

I used the 1996-1999 Prospector[®] database from the Center for Creative Leadership (CCL) for this research. Prospector is a 360-assessment-for-development feedback instrument developed to give managers and executives insight into their strengths and development needs. The instrument consists of 57 items designed to assess eleven domains related to managerial effectiveness. Table 1 presents the names and descriptions of the 11 scales. Each item used a 7-point, Likert-type response format. The textual anchors for the response scale were 1="very strongly disagree" to 7="very strongly agree" with 4="neutral" providing the anchor for the middle of the scale.

I selected a random sample of 752 survey families for use in this research. Each survey family included matched ratings from a manager (the self or ratee), a peer, a boss, and a direct report resulting in 3008 total surveys. Tables 2 and 3 give the breakdown of the gender, race, and rater group information for this sample. Because each family contained exactly one survey from each rater group, there were 752 each of self, peer, boss, and direct report surveys.

RESULTS

THE INFLUENCE OF STANDING ON RESPONSES

The main effect of standing on the attribute measured by the survey had by far the greatest influence on the item ratings of all the terms in the logistic model. The average effect due to the covariate was in the large-to-extreme category with only a very few items exhibiting a main effect due to covariate in the medium-to-large category. The effect of the covariate was never smaller than medium.

DIF ATTRIBUTABLE TO RATER GENDER

One item exhibited uniform DIF attributable to the gender of the rater. The effect was .12, placing the item in the lower end of the small-to-medium category. The product of the differential functioning was to elevate differentially the ratings given by male raters over those given by female raters. The text of the item was "Is willing to make substantial personal sacrifices for the sake of the business." The item occurred on the scale called "Committed to making a difference."

DIF ATTRIBUTABLE TO RATEE GENDER

Two items exhibited uniform DIF attributable to the gender of the ratee. The first item occurred on the scale called "Open to criticism." The text of the item was "Is not threatened by criticism." The second item occurred on the scale called "Seeks broad business knowledge." This item also exhibited DIF with respect to the rater type, and was the only item to exhibit DIF of two types. The text of this item was "Understands the financial side of the business." The effect associated with both items was .12, which placed the items in the lower end of the small-to-medium category. Both items functioned to differentially increment the ratings given to men.

DIF ATTRIBUTABLE TO RATER SOURCE

Twelve items on the survey exhibited uniform DIF attributable to the rater group. Table 4 presents these items along with the effect size and the impact of the DIF on the ratings. In each case, these items functioned to produce evidence of continuity from direct report to boss. Half of the items functioned to differentially increment the ratings given by bosses while the other half functioned to increment the ratings given by direct reports. Within particular scales, the direction of this functioning was consistently the same.

DISCUSSION

I used effect size instead of statistical significance to flag anomalous items in this study. I did this to avoid the complex power analysis that would suggest an appropriate sample size for use with logistic regression and these data. By using effect size, I failed to flag several items that I would have flagged were I using the typical p-value of .05, though the use of a Bonferroni correction may have ameliorated that Type I error rate. This result suggests to me that the sample size was sufficiently large for the analysis of the logistic model, and that the incidence of false positives is lower than it might otherwise be using statistical significance.

In most studies of differential item functioning, the existence of DIF is not something that is good, and I tend to concur with those who suggest that DIF is a quantity to remove from an assessment. The occurrence of DIF may threaten the validity of an assessment, and reduce the usefulness of the survey (Lord, 1980; Penny & Johnson, 1999). Psychological measures that assess the learning, the performance, or perhaps even the potential, of a candidate for a higher appointment are not the places where one generally wants to find measures influenced, even in part, by demographic quantities such as gender, race, nationality, or native language.

However, this study, suggests that the existence of differential item functioning may not be a completely bad thing, and further suggests that DIF produced by environmental complexity might be a naturally occurring phenomenon in the workplace. Hence, one could argue that the interpretation of 360-feedback reports should take into account the possibility of anomalous item functioning produced by such environmental and experiential differences. Moreover, this anomalous functioning may function to ameliorate, or exacerbate, some observed rating differences, and the degree of either may depend, at least in part, on the standing of the ratee on the trait measured by the survey.

The DIF related to the gender of the ratee and the raters seems another matter. The development of a useful and competitive 360-assessment is a long, involved, and arduous process where items receive repeated critical review from many groups of people. That this research uncovered (a) no DIF related to the cross of rater and ratee gender, (b) no DIF related to the cross of rater group and ratee gender, and (c) very little DIF attributable to the gender of ratees or raters is an important and heartening finding. However, that some items still tap, if ever so lightly, particular instances of stereotypical gender associations reminds us of the long road it has been towards gender equality and of the longer road that remains before us.

REFERENCES

- Bartol, K. M., & Butterfield, D. A. (1976). Sex effects in evaluating leaders. *Journal of Applied Psychology*, 61, 446-454.
- Cash, T. F., Gillen, B., & Burns, D. S. (1977). Sexism and "beautyism" in personnel consultant decision making. *Journal of Applied Psychology*, 62, 301-310.
- Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practices*, 2, 31-44.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Fiedler, F.E. (1978) The contingency model and the dynamics of the leadership process, In *Advances in experimental social psychology*, ed. L. Berowitz. New York: Academic Press.
- Fiedler, F.E. & Chemers, M.M. (1982). *Improving leadership effectiveness: The leader match concept*. 2nd Ed. New York: Wiley.
- Greenhaus, J. H., Parasuraman, S. (1993). Job performance attributions and career advancement prospects: An examination of gender and race effects. *Organizational Behavior and Human Decision Processes*, 55, 273-297.
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I.

- Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.
- Hsieh, F. Y. (1989). Sample size tables for logistic regression. *Statistics in Medicine*, 8, 795-802.
- Jacobson, M. B., & Effertz, J. (1974). Sex roles and leadership perceptions on the leaders and the led. *Organizational Behavior and Human Performance*, 12, 383-397.
- Jacques, E. (1996). *Requisite organization: A total system for effective managerial organization and managerial leadership for the 21st century*. Cambridge, MA: Cason Hall & Co.
- Jacques, E., & Clement, S. D. (1994). *Executive leadership: A practical guide to managing complexity*. Cambridge, MA: Basil Blackwell.
- Lord, F. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.
- Martell, R. F. (1991). Sex bias at work: The effects of attention and memory demands on performance ratings of men and women. *Journal of Applied Social Psychology*, 21, 1939-1960.
- Maurer, T. J., & Taylor, M. A. (1994). Is sex by itself enough? An explanation of gender bias issues in performance appraisals. *Organizational Behavior and Human Decision Processes*, 60, 231-251.
- Miller, T. R., & Spray, J. A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement*, 30, 107-122.
- Mobley, W. H. (1982). Supervisor and employee race and sex effects on performance appraisals: A field study of adverse impact and generalization. *Academy of Management Journal*, 25, 598-606.
- Nieva, V. F., & Gutek, B. A. (1980). *Women and work: A psychological perspective*. New York: Praeger.
- Noe, R. A. (1988). Women and mentoring: A review and research agenda. *Academy of Management Review*, 13, 65-78.
- Penny, J., & Johnson, R. L. (1999). How group differences in matching criterion distribution and IRT item difficulty can influence the magnitude of the Mantel-Haenszel chi-square DIF index. *The Journal of Experimental Education*, 67, 343-366.
- Powell, G. N. (1988). *Women and men in management*. Newbury Park, CA: Sage.
- Powell, G. N., & Butterfield, D. A. (1989). The "good manager": Did androgyny fare better in the 1980's? *Group and Organizational Studies*, 14, 216-233.
- Pulakos, E. D., & Wexley, K. N. (1983). The relationship among perceptual similarity, sex, and performance ratings in manager-subordinate dyads. *Academy of management Journal*, 26, 129-139.
- Rogers, H. J. (1989). A logistic regression procedure for detecting item bias. *Dissertation Abstracts International*, 50, 3928A. (University Microfilms No. 90-11,788).
- Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, 17, 105-116.
- Rosen, B., & Jerdee, T. H. (1974). The influence of sex-role stereotypes on evaluations of male and female supervisory behavior. *Journal of Applied Psychology*, 57, 44-48.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, 17.
- Samejima, F. (1979). *A new family of models for the multiple choice item*. Office of Naval Research Report 79-4. Knoxville, TN: University of Tennessee.
- Schmitt, N., & Lippin, M. (1980). Race and sex as determinants of the means and variance of performance ratings. *Journal of Applied Psychology*, 65, 428-435.
- Swaminathan, H., & Rogers, H. J. (1990). *Detecting differential item functioning using logistic regression procedures*, 27, 361-370.
- Thompson, D. E., & Thompson, T. A. (1983). Task-based performance appraisal for blue-collar jobs: Evaluation of race and sex effects. *Journal of Applied Psychology*, 70, 747-753.
- Van Velsor, E., & Leslie, J. B. (1991). *Feedback to managers: Vol 1. A guide to evaluating multi-rater feedback instruments*. Greensboro, NC: Center for Creative Leadership.
- Whittemore, A. (1981). Sample size for logistic regression with small response probability. *Journal of the American Statistical Association*, 42, 415-427.
- Yukl, G. A., (1981). *Leadership in Organizations*. Englewood Cliffs, New Jersey: Prentice Hall.

Table 1: Scales on the Prospector survey

Scale	Items
1. Seeks opportunities to learn	5
2. Seeks and uses feedback	5
3. Learns from mistakes	5
4. Open to criticism	3
5. Committed to making a difference	4
6. Insightful: Sees things from new angles	4
7. Has the courage to take risks	4
8. Brings out the best in people	5
9. Acts with integrity	4
10. Seeks broad business knowledge	4
11. Adapts to cultural differences	5

Table 2: Breakdown of gender and race by rater group

Rater Group	Gender		Race		
	Male	Female	White	Black	Other
Direct Report	495	257	638	62	52
Self	576	176	645	56	51
Peer	569	183	658	49	45
Boss	657	95	695	30	27

Table 3: Breakdown of gender by race

Race	Gender	
	Male	Female
White	2047	592
Black	136	61
Other	114	58

Note: N=3008

Table 4: Items that exhibited DIF attributable to rater type

Scale and Item	Text of item	Effect Size	Influence of DIF on responses
1.1	Has grown over time.	.15	B(+) to DR(-)
1.4	Has developed significant new skills over time.	.10	B(+) to DR(-)
2.4	Responds effectively when given feedback.	.13	B(+) to DR(-)
2.5	Has changed as a result of feedback.	.22	B(+) to DR(-)
6.1	Is good at identifying the most important part of a complex problem or issue.	.12	B(-) to DR(+)
6.4	Is good at asking insightful questions.	.11	B(-) to DR(+)
9.1	Can be depended on to tell the truth regardless of the circumstances.	.18	B(+) to DR(-)
9.3	Is seen by others as an honest person.	.12	B(+) to DR(-)
10.1	Has a solid understanding of our products and services.	.13	B(-) to DR(+)
10.2	Knows how the various parts of the organization fit together.	.18	B(-) to DR(+)
10.3	Knows the business.	.17	B(-) to DR(+)
10.4	Understands the financial side of the business.	.22	B(-) to DR(+)

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jim Penny
Center for Creative Leadership
One Leadership Place
Greensboro, NC 27438-06300

Work Phone: 336-286-4442
Fax: 336-286-4434
Email: pennyj@leaders.ccl.org
Web: www.ccl.org