

Using the SAS[®] System to Demonstrate the equivalence of On-line and On-paper Survey Administration across Levels of Raters

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 model used method of survey delivery and rater group to identify items that exhibited differential item functioning (DIF). The methods of survey delivery were pencil-and-paper and online by web page. The rater groups were self, boss, peer, and direct report. The sample consisted of 374 survey families where a survey family consisted of a matched set of four surveys: one self, one boss, one peer, and one direct report. Half of the survey families were from a pencil-and-paper administration; half were from an online administration. The sample contained 1496 total surveys. 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.

The results indicated little evidence to suggest that rating differences exist due to the method of survey delivery. However, approximately 10% of the survey items exhibited differential item functioning attributable to rater group, though the effect size for each item was small. The examination of the maximum likelihood parameter estimates suggested that some of this differential functioning could be the result of hierarchical complexity. The anomalous functioning of other items was attributable to contingency theory. This research further suggested that such forms of DIF might be naturally occurring phenomena in some 360-assessment, and that the interpretation of 360-feedback may need to include the potential for this DIF to exist.

INTRODUCTION

The past decade has seen a steady increase in the use of 360-assessments for managerial development in organizations, and the use of these assessments appears to continue without abatement. One appeal of the 360-process is that it involves providing managers with feedback from four or more sources. The four primary sources include (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 concept of gathering multi-source information is not new to the field of professional development, at least one premise of multi-rater methodology remains unresolved: Does the 360-methodology provide an equivalent measure with each rater group? That is, does measurement equivalence exist between the rater groups? Without rater equivalence, the comparison of ratings, often by between-group differences, may be problematic.

In addition to the increase in the use of 360-assessment, there has been an unrelenting move from the pencil-and-paper administration of 360-surveys to electronic administration using the Internet and the World Wide Web. However, one can question the influence of delivery mode on the results of a 360-assessment asking, does 360-methodology provide an equivalent measure with each method of delivery, or is one manner of survey delivery likely to produce better, or worse, ratings than the other? Moreover, does there exist a potential interaction between rater group and delivery method? Is there a combination of survey delivery and rater group that is likely to influence ratings, and produce an aberrancy that might threaten rating equivalence? These research questions involving the measurement equivalence of 360-methodology provided the impetus for this research.

DIFFERENCES AMONG RATER GROUPS

Ratings differences among rater groups are central to the 360-assessment process and often lead to much discussion during feedback sessions (Van Velsor & Leslie, 1991). Indeed, the study of such differences informs in part the developmental planning of the individual, and, if such differences did not exist, there would be little motivation to incur the expense of 360-assessment. For example, a manager may be interpersonally skilled with peers yet cold and aloof with direct reports. Such a manager could receive high ratings on interpersonal skills from peers while receiving low ratings on this dimension from direct reports, and learning of these differences in these perceptions provides a central motivation for the 360-assessment process.

Indeed, such rating differences as these often provide rich behavioral feedback for the target manager, enabling that manager to make better choices when planning developmental activities. However, the accurate interpretation of such a rating difference requires that one can assume that each set of ratings uses the same metric. If, for whatever reason, one group of raters interprets the text of an item or a set of items differently than another group, then the resulting differences in the ratings may be the result of not only the observations of the raters but also of the interpretative difference elicited by the item.

CONTINGENCY THEORY

Contingency theories of leadership (Fiedler 1978; Fiedler & Chemers, 1982) suggest that what constitutes an appropriate response may depend on the situation in which the response is to occur, and that two substantially different responses may be appropriate for similar stimuli when the contexts of the stimuli differ. Moreover, Hershey & Blanchard (1969, 1982) and Yukl & van Fleet (1982) suggest that managers often need to change their behavior to fit particular situations, and, following this line of reasoning, managers who behave differently toward different groups of co-workers may receive disparate ratings from members of those groups. Hence, differences between the ratings received from different rater groups may be an acceptable outcome for some managers, though the accurate interpretation of such differences may require some study on the part of the feedback provider. In addition, disparate ratings arising from differential environmental contexts may provide an indication of an effective manager, and, as such, these gaps in the perspectives between groups of raters may exist as a naturally occurring and perfectly acceptable phenomenon of management.

However, if an item taps a particular situation more strongly than other items, the observed difference in ratings produced by the situational behaviors of a manager may become confounded with the propensity of that item to accentuate a situational contingency. In so doing, the observed difference in ratings is not simply the result of observed behavioral differences, but is the combined result of both a behavioral difference and the interaction of the item with the raters. As such, the interpretation of the observed difference becomes problematic because one does not know how much of the rating difference is due to a difference in observed behaviors and how much is due to the anomalous functioning of the item.

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 with direct reports seeing a simpler and more easily understood environment than bosses see. This continuity of complexity and ambiguity can produce a reality that creates substantially different experiences for bosses and direct reports and that may require

thinking by bosses that may be difficult for bosses to explain to direct reports. For example, environmental complexity could lead bosses and direct reports to give a manager substantially different ratings in areas such as understanding corporate strategy if the direct reports feel that the manager has a firm grasp of corporate strategy while the boss knows there are areas of strategy that the manager has yet to see. Hence, it seems reasonable to anticipate that environmental differences may produce rating differences given on a 360-survey; however, those rating differences are likely to be a naturally occurring phenomenon in many corporate environments.

However, it also seems reasonable to anticipate that some items on a 360-survey could be more likely than other items to tap environmental complexity. In doing so, a given item could produce ratings from direct reports that are differentially higher than the ratings given by the boss, even though the observations of the both the boss and the direct reports were likely to produce otherwise similar ratings. In such an event, the rating difference between boss and direct report raters is not only the result of behavioral differences observed by the raters but is also the result of an interaction between the item and the environment of the rater.

METHODOLOGY

This research involved the use of logistic regression to detect DIF. I made this choice over other methods such as item response theory (IRT) and factor analytic techniques for several reasons. First, I did not anticipate having a sufficient sample to compute the IRT parameter estimates for each category of rating. As well, I did not anticipate achieving sufficient multivariate normality for use with the factor analytic methods. However, my strongest reason for the choice of logistic regression was the ability to posit a model that I could evaluate in a single analysis and then interpret using much of the explanatory framework developed from linear regression. In addition, the model I built, and will explain anon, was analogous in form and function to the underlying polytomous logistic model of item response theory.

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). 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. I have dropped the subscript k to make the model easier to read. 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 (Samejima (1969, 1979). This model represents the situation where the rater group membership and the survey delivery method do not influence the item response, and where the only factor that does influence the response to the item 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 delivery method and rater group membership, I would write

$$z = \tau_0 + \tau_1\theta + \tau_2g + \tau_3d,$$

where g and d represent rater group membership and delivery method, respectively, and τ_2 and τ_3 represent the logistic regression parameters for those two classifications. I can define the values for d as $\{0,1\}$ where 0 would indicate survey administration by web browser and 1 would indicate administration by pencil-and-paper. Similarly, I can define the values for g as $\{1, 2, 3, 4\}$ indicating *{self, direct report, peer,*

boss} respectively. This model describes the instance where only uniform DIF exists.

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

$$z = \tau_0 + \tau_1\theta + \tau_2g + \tau_3d + \tau_4g\theta + \tau_5d\theta,$$

where the two new terms indicate an interaction, respectively, between (a) rater group membership and standing on the attribute that the survey measures, and (b) survey delivery method 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 survey delivery method; the other term is to indicate the possible three-way interaction of rater group membership, survey delivery method, 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

$$z = \tau_0 + \tau_1\theta + \tau_2g + \tau_3d + \tau_4g\theta + \tau_5d\theta + \tau_6gd + \tau_7gd\theta,$$

where the symbols τ_6 and τ_7 represent the logistic regression parameters for these two additional interaction terms, respectively. It is this later model that I tested in this research.

I used the SAS® System to evaluate this model for the data that I collected. The SAS System produced Wald Chi Square statistics to test the null hypotheses that the parameter estimates of τ_1 through τ_7 were statistically significantly different from 0. It was my anticipation that τ_1 and τ_2 would routinely achieve statistical significance. It was also my anticipation that the parameter estimates of τ_3 through τ_7 would not routinely achieve statistical significance.

However, I knew that, with number of terms in the model I had chosen, I would be making quite a few 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 incur 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 to follow the lead of Penny & Johnson (1999) and to convert 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, and, although these values are indeed arbitrary, Penny & Johnson (1999) found that those values appeared to connote well when applied to the Mantel-Haenszel chi-square statistic. I used these three values to define four effect ranges that I would use to categorize the DIF that I discovered with the logistic regression model. These ranges were nil-to-small, small-to-medium, medium-to-large, and large-to-extreme.

DATA

I used the 1999 Benchmarks® database from the Center for Creative Leadership (CCL) for this research. Benchmarks is a 360-degree assessment-for-development feedback instrument developed by CCL researchers who conducted extensive research exploring the experiences that promote managers' development (Lindsey, Homes, & McCall, 1987; McCall & Lombardo, 1983; McCall, Lombardo, & Morrison, 1988). CCL researchers initially developed Benchmarks as a pencil-and-paper survey; it became available over the web in 1999. The instrument consists of three sections designed to assess skills and perspectives related to managerial effectiveness. I examined only one section of scales in this research. Table 1 presents the

names and descriptions of the 16 scales. All 106 items used a 5-point, Likert-type response format. The textual anchors for the scale were 1="Not at all," 2="To a little extent," 3="To some extent," 4="To a great extent," and 5="To a very great extent."

I selected a random sample of 374 survey families from the 1999 Benchmarks® database of the Center for Creative Leadership 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 1496 total surveys. Half of the families (N=187) were pencil-and paper-surveys; the other half was online surveys. Because each family contained exactly one survey from each rater group, there were 187 each of self, peer, boss, and direct report surveys.

I made no selection using any other demographic variables, so the sample reflected the demographic composition of the parent population, and was comprised of approximately 61% male ratees. Moreover, the sample contained approximately 90% white and 8% black ratees. In addition, about 49% held bachelor's degrees, 31% held master's degrees, and about 15% held doctoral or professional degrees. These data represented a broad range of organizations such as governmental agencies, manufacturing companies, and educational institutions. About 91% of the ratees held middle to upper level positions as managers.

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 was .48 with a standard deviation of .08; the range was from .25 to .61. In most instances, the effect was in the medium-to-large and large-to-extreme categories.

DIF ATTRIBUTABLE TO DELIVERY METHOD

These data provided no evidence of any substantial uniform or nonuniform DIF attributable to the manner in which the raters completed the survey. The mean effect of delivery mode was .02 with a standard deviation of .02 and a range from .00 to .07. In every instance, the effect was in the nil-to-small category.

DIF ATTRIBUTABLE TO RATER SOURCE

Ten items on the survey exhibited uniform DIF attributable to the rater group. Table 2 presents these items along with the effect size, the apparent explanatory theory, and the impact of the DIF on the ratings. These ten items produced effects that barely made the cut-off, .1, for a small effect. Complexity theory could explain the DIF found in six of the ten items. The differential functioning produced lower ratings from the direct reports in half of these six items.

Contingency theory explained the differential functioning found in the other four items exhibiting uniform DIF. The effect associated with these items was also small. Three of these items produced differential functioning that contrasted the ratings of direct reports with those from all other groups. With these three items, the differential functioning produced higher ratings from the direct reports. The other item that appeared to tap into contingency theory contrasted the peer ratings with those from all other groups. For this single item, the differential functioning produced reduced ratings from the peers.

Only one item on the survey exhibited nonuniform DIF. The effect associated with this item is .11, which is still a small value. This item appeared to tap into complexity theory, and the resultant DIF produced lower ratings from direct reports and higher ratings from bosses; however, the nature of nonuniform DIF suggests that the direction of the differential functioning may reverse for some range of the covariate.

DIF ATTRIBUTABLE TO DELIVERY METHOD BY RATER

SOURCE INTERACTION

This research uncovered no items that exhibited DIF attributable to an interaction between rater source and delivery method.

DISCUSSION

In most studies of differential item functioning, the existence of DIF is not something that is good. Rather, the occurrence of DIF may threaten the validity of an assessment, and reduce the usefulness of the survey (Lord, 1980; Penny & Johnson, 1999). I tend to concur with those who suggest that DIF is a quantity to remove from an assessment. 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 the type and degree of DIF found might be the result of two naturally occurring phenomena in the workplace. First, the differential environmental complexity that can exist in the workplace may produce unavoidable, if not expectable, differential functioning. Second, DIF could also be the natural result of the different behaviors that a manager can present when interacting with different groups of people. 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.

For example, the first scale contained 17 items, 3 of which exhibited DIF. I computed scale scores for the complete scale, and compared those values with the scale scores computed using the 14 items that did not exhibit DIF. The mean difference (and standard deviation of the difference) of those scale scores was {.12(.19), .04(.17), .10(.18), .03(.16)} for direct report, self, peer, and boss raters respectively. These differences, though statistically significant, do not appear to be of practical significance, conforming to the prior finding of small effect sizes.

One could argue that a manager receiving 360-feedback could benefit from knowing the degree to which environmental complexity influences particular feedback. With such a breakdown, the manager could better understand how much of a given rating arose from the observations of the boss, and how much arose from the different environment from which the boss interpreted not only the behaviors of the manager but also the text of the item. It is conceivable, then, that from such feedback a manager could at once choose the developmental areas on which to work for improved performance while learning more about the world as seen by the boss, a world in which the manager may soon work. Unfortunately, measurement theory does not yet permit the reliable partitioning of ratings that would allow the manager to understand the rating component due to behavior and the component due to environment, and additional theoretical research will be necessary to produce the methodology that will permit a breakdown of ratings into components.

REFERENCES

- 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.

Hershey, P., & Blanchard, K. H. (1969). Life cycle theory of leadership. *Training and Development Journal*, 23, 26-34.

Hershey, P., & Blanchard, K. H. (1982). *Management of Organizational Behavior*, (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Hsieh, F. Y. (1989). Sample size tables for logistic regression. *Statistics in Medicine*, 8, 795-802.

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.

Lindsey, E., Homes, V., & McCall, M.W., Jr. (1987). *Key events in executives' lives* (Tech. Rep. No. 32). Greensboro, NC: Center for Creative Leadership.

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

McCall, M.W., Jr., & Lombardo, M.M. (1983, February). What makes a top executive? *Psychology Today*, 26-31.

McCall, M.W., Jr., Lombardo, M.M., & Morrison, A.M. (1988). *The lessons of experience: How successful executives develop on the job*. Lexington, MA: Lexington Books.

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.

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.

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.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedure. *Journal of Educational Measurement*, 27, 361-370.

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., & van Fleet, D. (1982). Cross-situational, multi-method research on military leader effectiveness. *Organizational Behavior and Human Performance*, 30, 87-108.

Table 1. Name and description of scales on the Benchmarks survey

Scale	Items	Description
1. Resourcefulness	17	Can think strategically, engage in flexible problem-solving behavior, and work effectively with higher management.
2. Doing Whatever It Takes	14	Has perseverance and focus in the face of obstacles.
3. Being a Quick Study	4	Quickly masters new technical and business knowledge.
4. Decisiveness	4	Prefers quick and approximate actions to slow and precise one in many management situations.
5. Leading Employees	13	Delegates to employees effectively, broadens their opportunities, and acts with fairness towards them.
6. Setting a Developmental Climate	5	Provides a challenging climate to encourage employees' development.
7. Confronting Problem Employees	4	Acts decisively and with fairness when dealing with problem employees.
8. Work Team Orientation	4	Accomplishes tasks though managing others.
9. Hiring Talented Staff	3	Hires talented people for his or her team.
10. Building and Mending Relationships	11	Knows how to build and maintain working relationships with coworkers and external parties.
11. Compassion and Sensitivity	4	Shows genuine interest in others and sensitivity to subordinates' needs.
12. Straightforwardness and Composure	6	Is honorable and steadfast.
13. Balance between Personal Life and Work	4	Balances work priorities with personal life so that neither is neglected.
14. Self-awareness	4	Has an accurate picture of strengths and weaknesses and is willing to improve.
15. Putting People at Ease	4	Displays warmth and a good sense of humor.
16. Acting with Flexibility	5	Can behave in ways that are often seen as opposites.

Table 2. Description of the items that exhibited DIF and the type of DIF observed

Item text with scale and item number	Type of DIF	Effect	Explanatory Theory	Impact of DIF
1.6 Understands higher management values, how Higher management operates, and how they see things.	Uniform	.10	Complexity	Dr (-) to Boss (+)
1.10 Does his/her homework before making a proposal to top management	Uniform	.10	Contingency	DR (+) vs. all others
1.13 Establishes effective management practices for directing employees he/she sees only twice a month.	Uniform	.11	Contingency	DR (+) vs. all others
8.4 Focuses more on managing other people to accomplish a task than on personally finishing everything the work group does.	Uniform	.11	Complexity	DR (+) to Boss(-)
9.3 Surrounds him/herself with the best people	Uniform	.11	Complexity	DR (+) to Boss (-)
10.6 When working with peers from other functions or units, gains their cooperation and support	Uniform	.11	Contingency	DR (+) vs all others
11.3 Is sensitive to signs of overwork in others.	Uniform	.12	Complexity	DR(-) to Boss (+)
12.2 (Does not) rely on style more than substance in dealing with top management.	Uniform	.11	Contingency	Peer (-) vs. all others
2.11 Accepts conflicts as inevitable and does not shy away from them.	Uniform	.13	Complexity	DR (+) to Boss (-)
4.1 Displays a real bias for action, calculated risks, and quick decisions.	Uniform	.11	Complexity	DR (-) to Boss (+)
4.4 Does not hesitate when making decisions.	Nonuniform	.11	Complexity	DR (-) to Boss (+)

Note: An effect between .10 and .30 is a small-to-medium effect. The sign beside the rater group name indicates the direction in which the DIF changed the ratings.

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