Rating Scale Label Effects on Leniency Bias in 360-degree Feedback

Andrew English, Ph.D., 3D Group
Dale Rose, Ph.D., 3D Group
Jillian McLellan, San Francisco State University

April 2009

Paper presented at the 24th Annual Meeting of the Society for Industrial Organizational Psychologists (2009). New Orleans, LA.

Rating Scale Label Effects on Leniency Bias in 360-degree Feedback

A recent survey found that 90 percent of Fortune 1000 companies currently use 360-degree feedback programs. While the almost universal adoption of 360-degree feedback programs attests to the utility of the process, 360 feedback surveys suffer from score distributions that are highly negatively skewed with very little variability This study found that the utilization of a response scale with a larger number of positive scale anchors resulted in lower mean scores and increased variability for a 360-degree feedback survey.

Introduction

In 1996, 360-degree feedback programs were almost ubiquitous among Fortune 500 companies, and it was estimated that hundreds of millions of dollars were spent annually to support them (Yammarino & Atwater, 1997). A more recent survey found that 90 percent of Fortune 1000 companies currently use 360-degree feedback programs (Carruthers, 2003). Though 360-degree feedback has become a standard within modern organizations, a recent benchmarking study of 360 practices found that 360 feedback practices are far from standardized (Rose & Walsh, 2004). Indeed, the wide range of uses and applications for 360-degree feedback make it difficult to identify a single practice that characterizes all 360 programs. Ironically, though these programs often vary, often the actual survey responses do not. Low variability and inflated scores are challenges common to 360 feedback data. It is unclear, however, whether inflated scores and low variability in 360 ratings reflect some form of rater bias or is driven by the restriction of range intrinsic to incumbent populations.

The draw to 360-degree feedback programs is easy to understand given the restructuring of organizations to cope with a globalized marketplace over the past three decades. Flatter organizational structures are less receptive to the top-down approach of traditional performance evaluation models, and more likely to accept feedback as more accurate when coming from multiple constituencies and organizational members who are most familiar with the participant's work behaviors. While many organizations use 360-degree feedback programs for developmental purposes only, others use them in conjunction with administrative purposes, such as making decisions regarding succession planning, performance appraisal, career management, and promotion. Given the persistent increase in the use of 360-degree degree feedback over the last decade, it appears that 360-degree feedback programs are here to stay. Indeed, organizations seem to continue to find more applications for 360-degree feedback data.

While the almost universal adoption of 360-degree feedback programs attests to the utility of the process, score inflation and generally low variability in ratings are frustrations that practitioners have come to accept and researchers have yet to thoroughly examine. 360 feedback surveys often suffer from score distributions that are highly negatively skewed with very little variability (LeBreton, Burgess, Kaiser, Atchley, and James, 2003). It is unclear whether the typical skewness found in 360 data is caused by the range restriction associated with incumbent populations or if there is some way to improve the measurement characteristics of 360 surveys by altering some aspect of the measures themselves.

Leniency Effects in 360-Degree Feedback Ratings

Guilford, in his classic Psychometric Methods (1936) addressed the issue of errors and leniency. He recommended that "in a similar manner in the numerical type of scale, the strength of the descriptive adjectives may be adjusted so as to counteract the error of central tendency" (1936, p. 272). He suggested that in order to persuade raters to use the entire scale instead of solely focusing on the extreme negative or positive points, it may be necessary to change the scale labels to reflect less extreme points on the scale. Lam and Klockars (1982) looked at the effect of scale labels on leniency and found that the scale with four positive scale labels and one negative scale label had the lowest mean and highest standard deviation compared to the other three scales used in their study. Their findings indicated that when more positive labels are added to a scale respondents use a larger number of scale points. Research has demonstrated that using a rating scale anchored by more positive labels and less extreme endpoints can result in raters

using the scale in a more balanced manner (French-Lazovik & Gibson, 1984; Wyatt & Meyers, 1987). Lam and Klockars (1982) found lower means for a "positive packed" scale ("Poor" - "Fair" - "Good" - "Very Good" - "Excellent") than for a "negative packed" scale ("Poor" - "Moderately Poor" - "Fair" - "Good" - "Excellent") and a "equally spaced" scale ("Poor" - "Need Improvement" - "Satisfactory" - "Quite Good" - "Excellent"). Lam and Klockars (1982) concluded that to obtain finer discriminations between subjects "it may be necessary to pack the scale with response options from one part of the underlying continuum" (p.321).

LeBreton, et al. (2003) looked at the variance across raters (supervisors, peers and subordinates) for 3,851 managers and found that the results tended to be negatively skewed. They attributed this to the idea of "corporate Darwinism" where individuals in management positions are assumed by raters to have the ability, motivation and experience needed to reach the management level. Consequently, 360 ratings tend to be inflated towards the higher end of the scale. In order to combat this problem, the authors suggested that organizations avoid using rating scales that are insensitive to subtle differences between answer choices, such as having response categories that are too broad. Also, the authors suggested a more normal distribution might be achieved by increasing the number of positive anchor labels and discouraging raters from using the most positive options by creating a more extreme positive anchor label, such as "exceptionally effective."

Another possible reason for inflated ratings on 360 degree feedback surveys is the purpose for which they are being administered. Research has demonstrated that when 360-degree feedback is used for administrative purposes elevated ratings become a significant issue (Farh, Cannella & Bedeian, 1991). Jawahar and Williams (1997) did a meta-analytic review of performance appraisal purpose research that included 22 studies and a total sample size of 57,775. They found that performance evaluations obtained for administrative purposes were, on average, one-third of a standard deviation larger than those obtained for research or employee development purposes. This indicates that raters might inflate their ratings, and not provide honest feedback if they know that information will be used for administrative purposes. McCarthy and Garavan (2001) looked specifically at 360 degree feedback surveys and found that the purpose of the 360 program affected the ratings participants received. Specifically, inflated ratings were found when the feedback was intended for performance management purposes. London and Wohlers (1991) found that 34 percent of employees participating in upward feedback reported they would have rated their boss differently if the feedback was used for administrative purposes. Toegel and Conger (2003) also examined 360 degree feedback programs and suggested that the efficacy of providing honest and constructive feedback would be lost if the ratings were used for administrative purposes. In a study conducted by Robinson, & Mongeon (2006), 360-degree feedback implemented for performance management was associated with higher scores than 360-degree feedback implemented for developmental purposes only.

Inflated ratings on 360-feedback surveys also present an additional issue to those providing feedback facilitation to the participants of such feedback programs. Inflated ratings can make interpreting the 360-degree feedback survey results a difficult process because of the difficulty in identifying development priorities for the participant with data that is negatively skewed.

Rating Scales

The majority of organizations use a behavioral rating scale to measure performance (Kaiser & Kaplan, 2005). The most common type used on 360 degree feedback surveys is the frequency type of response scale, where raters are asked to judge how frequently the feedback recipient exhibits a certain behavior (e.g., "Never" - "Sometimes" - "Usually" - "Always"). The second most common is the evaluation type of response scale, in which raters are asked to evaluate how effectively the feedback recipient performs a certain behavior, role, or function (e.g., "Below Expectations" - "Meets Expectations" - "Exceeds Expectations"). There are several issues that need to be considered when creating the response scale for 360 degree feedback surveys. Three of the most common issues include the number of response categories that should be included, labeling anchors appropriately, and how to reduce inflation in ratings.

Number of Response Categories

There is no definitive agreement in the literature about the optimal number of response categories that should be used in order to get the most reliable data. However, there is a general range provided in the

empirical literature that suggests the number of response categories to use. Preston and Colman (2000) examined response categories ranging from two to eleven and found that test-retest validity was lowest for two to four point scales, highest for seven to ten point scales, and decreased for scales with more than ten response categories. Lozano, Garcia-Cueto, and Muniz (2008) also looked at the reliability and validity of scales ranging from two to nine response options with four different sample sizes and found that the optimum number was between four and seven options. Bandalos and Enders (1996) found similar results in that the reliability was highest for scales having five to seven points. The most common scale found in 360-degree feedback surveys is a five point scale.

Rating Scale Labels

In addition to the number of response options, a second issue for response scales is considering how to best label the anchors (i.e., what are the best verbal qualifiers to use?). Weng (2004) looked at the reliability of Likert-type rating scales and found that the scales with all the response options clearly labeled yielded higher test-retest reliability than those with only the end points labeled. Having all of the points on a scale clearly labeled helps reduce ambiguity. Cools, Hofmans, and Theuns (2006) looked at the number of response categories and the chosen verbal qualifiers and found that a scale with five response options was the least prone to context effects. They also found that the use of extreme answer categories on the left and right ends of the scale did not improve the metric properties of the scale. Viswanathan, Bergen, Dutta and Childres (1996) also looked at the optimal number of response categories and appropriate category descriptors. Their results indicated that finding the right number of categories was important so that there was not a mismatch between participants' natural responses and the response categories. It is important to not have response scales be so fine grained that participants' natural responses are not included, but not so general that there is overlap between categories.

Study Objectives

The purpose of this paper is to test the hypothesis that score inflation and low variability in ratings on 360-degree feedback surveys may be mitigated by changing anchor labels of the rating scale to include more anchors in the positive range. More specifically, the present study examines the effect of two different rating scales on the means and standard deviations of a 360-degree feedback survey.

Method

Participants

Seven thousand five hundred and twenty six individuals provided ratings for a 360-degree feedback survey. Ratings were made by self, direct manager, peers, and direct reports from multiple organizations representing a broad array of industries. For most 360-degree feedback reports, the ratings from the participant's direct manager, peers, and direct reports are aggregated into a score called "overall ratings". For purposes of this study, only ratings from these rater groups were included in the analyses.

Materials

The 360-degree assessment used in the current study was the Leadership Navigator® for Corporate Leaders. This standardized 360-degree feedback survey assesses managerial behaviors relevant to mid-level manager jobs. The feedback reports contain competency norms that provide managers with a reference to compare their competency ratings to other similar-level managers across the United States. Survey items are grouped under eight competencies: Communication Skills and Integrity are the base competencies, Business Focus, Results Orientation, and Customer Focus are grouped under Work Process competencies, and Developing Talent, Inclusiveness, and Team Leadership are grouped under Interpersonal competencies (Healy & Rose, 2003). Using Cronbach's Alpha, the reliability of the eight competencies range from .80 to .88 (Robinson, Rose, & Wilkinson, 2005). The eight competencies listed above are represented

on the survey by 50 behavioral items, with a range of four to eight items per competency. Each item is rated on a five-point frequency scale.

Procedure

Table 1 contains the two scales that were utilized for this study. The "positive scale" is comprised of anchors with a larger number of positive verbal qualifiers (e.g., 2 = sometimes), and the "typical scale" contains anchors that are typically found in 360-feedback surveys (e.g., 2 = infrequently). The mid-point of the positive scale is frequently, while the mid-point of the typical scale is about half the time. Raters participating in a 360-degree feedback process were asked to respond to the items using one of these scale variants. The positive scale was utilized by 5087 raters rating 385 managers from 58 companies, and the typical scale was utilized by 2339 raters rating 153 managers from 17 companies.

Table 1. Scales Compared for the Study

Scale	1	2	3	4	5
Positive Scale	*Almost Never*	*Sometimes*	*Frequently*	*Almost Always*	*Always*
Typical Scale	*Never*	*Infrequently*	*About Half the Time*	*Usually*	*Always*

Results

Descriptive Statistics

All data was analyzed using SPSS v 11.5. The Leadership Navigator® for Corporate Leaders measures eight competencies. The eight competency scores were computed for both samples. Descriptive statistics were computed and competency means and standard deviations are presented in Table 2.

Table 2. Descriptive Statistics across Scales

Competencies	Mean		Standard Deviation		Sample Size	
	Positive Scale	Typical Scale	Positive Scale	Typical Scale	Positive Scale	Typical Scale
Business Focus	4.26	4.45	.70	.49	4654	2157
Results Orientation	4.08	4.29	.70	.52	4632	2169
Customer Focus	4.24	4.46	.74	.53	4592	2081
Communication	4.12	4.32	.67	.52	4672	2175
Developing Talent	3.88	4.16	.82	.64	4493	2105
Inclusiveness	4.07	4.32	.76	.59	4655	2167
Team Leadership	3.95	4.21	.81	.61	4568	2132
Integrity	4.12	4.44	.77	.56	4664	2173

Reliability of Competency Scores

Reliabilities for each competency were examined using Cronbach's Alpha and can be found in Table 3. Excluding the competency of Integrity, reliabilities on the positive scale were larger for all competencies by .02 to .04 compared to the typical scale. The largest difference in reliability was for the Business Focus scale (.04 difference), and while the Integrity competency did not demonstrate a difference in reliability, the reliability estimate was identical across both scales (α = .81).

Table 3. Reliabilities Measured by Cronbach's Alpha

Competencies	Alpha	
	Positive Scale	Typical Scale
Business Focus	.88	.84
Results Orientation	.90	.88
Customer Focus	.88	.84
Communication	.86	.84
Developing Talent	.90	.88
Inclusiveness	.88	.85
Team Leadership	.90	.88
Integrity	.81	.81

Mean Scores across Scales

An independent samples t-test was conducted to determine if mean differences across competencies between scale-types were significant. Table 4 presents the mean score differences across both scales for all eight competencies. Mean differences ranged from .19 to .32 across the competencies. For all eight competencies, the positive scale means were significantly lower ($p<.01$) than the typical scale means. The largest mean difference was found for the Integrity competency (.32) and the lowest mean difference was found for the Business Focus competency.

Table 4. Mean Differences across Scales

Competencies	Mean		Difference
	Positive Scale	Typical Scale	
Business Focus	4.26	4.45	.19*
Results Orientation	4.08	4.29	.21*
Customer Focus	4.24	4.46	.22*
Communication	4.12	4.32	.20*
Developing Talent	3.88	4.16	.28*
Inclusiveness	4.07	4.32	.25*
Team Leadership	3.95	4.21	.26*
Integrity	4.12	4.44	.32*

Note: Mean differences computed using independent samples t-test.
*p<.01

Standard Deviations across Scales

Standard deviations were examined across scales to determine the amount of variance associated with the use of both scales. Table 5 displays the standard deviation differences across scales. Across all eight competencies standard deviations were greater for the positive scale than for the typical scale. Differences ranged from .15 to .21. The greatest difference was found for the Integrity, Business Focus and Customer Focus competencies (.21), and the smallest difference was found for the Communication competency (.15).

Table 5. Standard Deviation Differences across Scales

Competencies	Standard Deviation		Difference
	Positive Scale	Typical Scale	
Business Focus	.70	.49	.21
Results Orientation	.70	.52	.18
Customer Focus	.74	.53	.21
Communication	.67	.52	.15
Developing Talent	.82	.64	.18
Inclusiveness	.76	.59	.17
Team Leadership	.81	.61	.20
Integrity	.77	.56	.21

In addition to reviewing the standard deviations of the competency scores, standard deviations were noted at the item-level. For the typical scale, the lowest standard deviation was .59 and the greatest standard deviation was .96 at the item level. The average standard deviation across all items was .73 using the typical scale. For the positive scale, the lowest standard deviation was .73 and the greatest standard deviation was 1.18 at the item level. The average standard deviation across all items was .93 on the typical scale.

Skewness and Kurtosis

Skewness provides a measure of the extent a distribution of values deviates around the mean (i.e., lack of symmetry). A skewness value of zero represents perfect symmetry, and positive skewness values represent positive skewness (greater number of smaller values), whereas negative skewness values represent negative skewness (greater number of larger values). Table 6 displays the skewness values for both scales. Across scales, skewness values were identical for the competency of Team Leadership. For the positive scale there was a decrease in negative skewness for Integrity, Inclusiveness and Developing Trust, and for the typical scale there was a decrease in negative skewness for Business Focus, Results Orientation, Customer Focus, and Communication. The greatest improvement in skewness was found for the competency of Integrity.

Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. A kurtosis value of zero indicates a normal distribution, and positive kurtosis values indicate a more peaked shape than normal, whereas negative kurtosis values indicate a more flat shape than normal. Table 6 displays the kurtosis values for both scales. For the positive scale kurtosis improved for Results Orientation, Customer Focus, Developing Talent, Inclusiveness, Team Leadership, and Integrity, and for the typical scale kurtosis improved for Business Focus and Communication. The greatest improvement in kurtosis was found for the competency of Integrity.

Table 6. Skewness and Kurtosis across Scales

Competencies	Skewness		Kurtosis	
	Positive Scale	Typical Scale	Positive Scale	Typical Scale
Business Focus	-1.31	-1.02	2.02	1.68
Results Orientation	-.97	-.87	1.04	1.07
Customer Focus	-1.30	-1.15	1.87	1.97
Communication	-1.05	-.96	1.45	1.37
Developing Talent	-.79	-.89	.29	.74
Inclusiveness	-1.06	-1.19	1.06	1.68
Team Leadership	-.94	-.94	.70	.97
Integrity	-1.06	-1.52	.86	3.06

Discussion

The purpose of this study was to examine the effects of two different rating scales on the means and standard deviations of a standardized 360-degree feedback survey. The results of this study extends research to the 360-degree feedback arena that suggests positive anchor labels can be used to counteract leniency effects in scale ratings for-360-degree feedback surveys (Lam and Klockars, 1982). Rating inflation (as measured by mean competency scores) was significantly lower for the positive scale than for the typical scale across all competencies. Variability was also quite a bit better for the positive scale compared to the typical scale as indicated by the standard deviations that were quite a bit larger across all competencies for the positive scale compared to the typical scale.

There are various explanations for leniency effects in 360-degree feedback ratings. One explanation is that individuals in management positions going through the 360-degree feedback process presumably display the appropriate behaviors for their job to some extent or they would not retain their position (i.e. range restriction in true scores expected in an incumbent population). This leads to a negatively skewed distribution as the points on the higher end of the scale are (accurately) utilized to a large degree. Another explanation, specific to circumstances when the 360-degree feedback ratings are tied to administrative decisions, is that raters are more lenient in their ratings because they do not want the individual to receive negative repercussions from the feedback or they fear their individual ratings may be identifiable. Though leniency effects in 360-degree feedback ratings may be caused by a range of factors, the present research indicates one way to mitigate these effects is to improve the response scale by making the anchors more positive.

In the current study, a response scale which provided more positive response options was compared to a more balanced response scale. By using a response scale anchored by more positive labels and less extreme endpoints it appears the raters used the scale in a more balanced manner. If “corporate Darwinism” does in fact play a role in skewed ratings, then a response scale anchored more heavily with positive labels should more accurately represent a rater’s natural response to the items they are rating. In other words, if managers must exhibit, at least to some extent, particular behaviors at work to retain their jobs then including a label such as “never” on the response scale will offer little to no utility. Also, if raters are concerned about the repercussions of their ratings on the participant, then using a more positive response scale should encourage raters to use the scale in a more balanced manner as well because they will not feel as if their feedback is extremely negative in nature. It is important to note that for the current study the competency scale mean which demonstrated the greatest mean shift was for the competency of Integrity. One explanation is that individuals feel uncomfortable providing extremely negative ratings on this dimension and thus only use the positive points of the response scale. When more positive points are used across the scale, they utilized these points, decreasing the mean and increasing the variability of responses.

Limitations and Future Considerations

Limitations of this study are that only two rating scales were utilized. Future research should provide a comparison of a larger number and variety of scale labels. In addition to labels, future research should compare the scales across both development only and administrative purposes contexts. Another consideration for this study is that only the participant’s overall scores were examined. Future research should also look more closely across the various rater groups. It would be important to understand which rater groups are more prone to leniency effects and to understand rater group differences in the utilization of a more positive rating scale. Research should also focus on identifying specific elements of the survey content which are more prone to leniency effects. While the positive scale demonstrated lower mean scores across all competencies, the skewness and kurtosis values provided evidence suggesting that some competencies received even greater benefits from this scale over other competencies.

References

Bandalos, D. L., & Enders, C. K. (1996). The effects of nonnormality and number of response categories on reliability. Applied Measurement in Education, 9, 151-160.

Carruthers, F. (2003). Nothing but the truth. Australian Financial Review, 78.

Cools, W., Hofmans, J., & Theuns, P. (2006). Context in category scales: Is “fully agree” equal to twice agree? European Review of Applied Psychology, 56, 223-229.

Dalessio, A.T. (1998). Using multi-source feedback for employee development and personnel decisions. In Smither J.W. (Ed.), Performance appraisal: State of the art in practice (278-330). San Francisco: Jossey-Bass.

Farh, Jiing-Lih, Albert A. Cannella Jr., and Arthur Bedeian. 1991. The effects of purpose of appraisal on the peer evaluation process. Group and Organization Studies, 16: 367-386.

French-Lazovik, G., & Gibson, C. L. (1984). Effects of verbally labeled anchor points on the distributional parameters of rating measures. Applied Psychological Measurement, 8, 49-57.

Guilford, J. P. (1936). Psychometric Methods. New York: McGraw-Hill.

Jawahar, I. M., & Williams, C. R. (1997). Where all the children are above average: The performance appraisal purpose effect. Personnel Psychology, 50, 905-926.

Kaiser, R. B., & Kaplan, R. E. (2005). Overlooking overkill? Beyond the 1-to-5 scale. Human Resource Planning, 28, 7-11.

Klockars, A. J., & Yamagishi, M. (1988). The influence of labels and positions in rating scales. Journal of Educational Measurement, 25, 85-96.

Lam, T. C. & Klockars, A. J. (1982). Anchor point effects on the equivalence of questionnaire items. Journal of Educational Measurement, 19, 317-322.

LeBreton, J. M., Burgess, J. R. D., Kaiser, R. B., Atchley, K. E., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? Organizational Research Methods, 6, 80-129.

London, M., Wohlers, A.J. (1991), "Agreement between Subordinates and Self-ratings in Upward Feedback", Personnel Psychology, 44, 375-90.

Lozano, L. M., Garcia-Cueto, E., Muniz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 4, 73-79.

McCarthy, A. M., & Garavan, T. N. (2001). 360[degrees] feedback process: Performance, improvement and employee career development. Journal of European Industrial Training, 25, 5-28.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. Acta Psychologica, 104, 1-15.

Robinson, G. N., & Mongeon, K.M. (2006). Relationship Between the Purpose of 360-Degree Feedback Program and Feedback Scores Over Time. Technical Report #8301. Berkeley CA: 3D Group.

Rose, D.S. & Walsh, A.B. (2004). Current Trends in 360 Degree Feedback. Technical Report #8285. Berkeley CA: 3D Group.

Toegel, G., & Conger, J. A. (2003). 360-degree assessment: Time for reinvention. Academy of Management Learning & Education, 2, 297-311.

Viswanathan, M., Bergen, M., Dutta, S., & Childres, T. (1996). Does a single response category in a scale completely capture a response? Psychology & Marketing, 13, 457-479.

Weng, L. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. Educational and Psychological Measurement, 64, 956-972.

Wyatt, R. C., & Meyers, L. S. (1987). Psychometric properties of four 5-point Likert-type response scales. Educational and Psychological Measurement, 47, 27-35.

Yammarino, F.J., Atwater, L.E. (1997), "Do managers see themselves as others see them? Implications of self-other rating agreement for human resources management", Organizational Dynamics, 25, 35-44.

About 3D Group

3D Group is dedicated to helping businesses of all sizes use human resource assessments to increase the effectiveness of their people and programs. Experts in 360º feedback, personnel selection, training assessment, and program evaluation, 3D Group's assessment tools provide in-depth information about how well employees and organizational initiatives are performing. 3D Group researchers are widely published and regularly present studies to national audiences on a variety of assessment related topics.

For more information about 3D Group contact us at (510) 525-4830 or visit our website: www.3dgroup.net