The Secret to The "Best" Ratings from Any Evaluation Scale

Ronald A. Berk
The Johns Hopkins University

Do you ever complete, select, write, or administer rating scales to provide data for decision making? "NO?" Then you must be working in a storm drain. Most faculty developers I know have a wide variety of rating scales that fly across their desk tops as their incremental program activities unfold during the academic year. The primary issue for this column is: What is the quality of those ratings used for decisions about people and programs?

When students, faculty, and administrators rate a program or someone's performance, we assume that they will read each item carefully and make their honest assessment with scrupulous impartiality (LOL). Are we dreaming in 3D IMAX like Avatar? The problem is that human tendencies may contaminate their responses, rendering them less than honest and impartial. In fact, these tendencies may be driven by demonic forces, such as those seen on *Fringe* or *The Vampire Diaries* or appearing in Stephen King's books. However, they may be conscious and intentional or unconscious. These tendencies are known as *response sets* or *biases*.

More than 10 types of bias can affect ratings, such as halo effect; end-aversion bias; extreme-response bias; acquiescence bias; and gender, racial/ethnic, and sexual orientation bias. There are four that particularly afflict faculty or administrator ratings: leniency-stringency bias, incompetence bias, buddy bias, and back-scratching bias.

Your awareness of these biases and the techniques to minimize or eliminate them can produce more accurate ratings and foster more perceptive decisions than those currently being made. These biases are briefly described next along with suggestions for minimizing their effects. For a more detailed, rollicking treatment of the topic with research citations, see Berk (2006), Nunnally and Bernstein (1994), and Streiner and Norman (1995).

1. Halo Effect

This effect is the extent to which a rater's *overall impression of a person will affect his or her rating on each item*. For example, if the global impression is positive and a student really likes her instructor, she may simply mark "Strongly Agree" to all positive statements. However, despite the positive image of a halo on an angel (*Remember:* The TV series "Touched by a Halo"?), this effect can also be negative, such as when a student hates the course.

Possible solutions: Since this is an individual bias, the larger the group of raters, the greater the chance that positive and negative halo effects will cancel each other out. Unfortunately, there is *no simple solution to control or minimize halo effect*.

2. End-Aversion/Central Tendency Bias

This is the tendency to *ignore the extreme response options or anchors on the scale* because they may be viewed as too strong. When the extremes are not selected, the ratings may be squished into the middle of the scale, which restricts the range of responses.

Possible solutions: This bias should not be a problem on most rating scales in the *agree–disagree* format. However, scales consisting of several evaluation anchors, such as *excellent–poor*, or frequency anchors with absolutes, such as *always–never*, can produce this type of bias. Either *soften the extreme anchors* with "Almost Always" and "Almost Never," or *extend the number of anchors* with the expectation that the extremes will rarely be chosen.

3. Extreme-Response Bias

This is the opposite of end-aversion bias. In this case, the respondents may *mark the extreme anchors rather than those in between*. This bias is difficult to detect because the reason for the choice of the extremes may also be due to their honest ratings or the halo effect.

Possible solution: This type of bias may occur less

often than the other sources of bias. However, there is no direct solution.

4. Acquiescence/Affirmation/ Agreement/Yea-Saying Bias

This is the tendency to *agree or give positive responses to statements irrespective of their content*. Most of us are socialized to be agreeable, to say "yes" instead of "no," and when asked, "How are you?" we answer "~~yucky~~" "fine," whether we honestly mean it or not. For example, faculty peer evaluators may select "Excellent," "Very Good," and "Good" more often than negative anchors or "Needs Improvement" on an observation checklist. This response set tends to inflate the ratings so that an instructor's performance appears much better than it really is.

Possible solutions*:* One remedy is *to word half of the statements on the scale positively and the other half negatively*. This 50-50 distribution does not eliminate or reduce the bias; it simply cancels out its effect. This solution is workable in practice with most rating scales designed to measure educational, psychological, business, management, and healthcare constructs. However, this practice may not be appropriate for student rating scales and other measures of teaching effectiveness.

5. Gender, Racial/Ethnic, and Sexual Orientation Bias

Given the issues of salary inequity, differential hiring and promotion rates, and available benefits/privileges at the different ranks, *cross-gender, cross-race, or straight–gay ratings can exhibit conscious or unconscious bias*.

Possible solutions: *Sensitivity or diversity training* is the most common strategy to minimize this type of bias. Such bias, or prejudice, may manifest itself in so many insidious forms that it is frequently difficult to detect unless the ratings are systematically lower for certain persons than for others. Even then, those ratings may be justified on nonprejudicial grounds. *Requiring multiple ratings* by a diverse band of raters may tend to counterbalance the bias by any single rater.

6. Leniency-Stringency Bias

Some *raters tend to be more lenient or forgiving, while others are more stringent or unforgiving*. The motives or reasons behind the bias may be unknown.

Consider the recent 2010 Winter Olympics, which rolls around every 16 years. The most obvious display of leniency-stringency bias is the pattern of judges' ratings of figure skaters. This provides a prime example of an untainted sport, like boxing, where judges carefully evaluate both the technical merit and artistic impression of each skater's performance and then vote for whomever they were going to vote for anyway. Since each judge represents a particular country, as the competition progresses, a visible trend develops in the judges' ratings. For example, when the ratings for each skater are posted, everyone can see that the judge from France consistently rates lower than all of the other judges. That is *stringency bias*. At the other end of the rating spectrum, the judge from Jamaica is so excited just to be in the Olympics with his country's ice hockey team, although the team is made up of Canadians who trained in Frostbite Falls, Minnesota (home of Rocky and Bullwinkle), that he seems to be giving the skaters consistently higher ratings than the other judges. This is *leniency bias*.

Possible solutions: These biases have implications for peer ratings. The biggest problems are that lenient and stringent raters and the reasons for their particular biases are nearly impossible to identify. Peer raters should have *adequate training with the scale and the observational procedures*. That might help minimize the bias. Also, u*se multiple peer raters* for each instructor to balance the possible biases.

7. Incompetence Bias

This is the tendency to *assign high ratings because of a lack of competence and/or confidence in rating teaching or other behaviors*. When raters are incompetent on the characteristics being rated, they tend to give more positive ratings, rather than penalize the person being rated for his or her own shortcomings.

Possible solutions: *Proper training* in the particular teaching behaviors or performance being rated may minimize incompetence bias. Alternatively, observers shouldn't be asked to rate the items on the scale that they are unqualified to rate.

8. Buddy Bias

Friendship and degree of acquaintance can inflate peer, administrator, and employer ratings.

Possible solutions: This bias can be eliminated if the *peer rater is chosen by someone other than the peer's "buddy,"* such as an administrator, or if another administrator or employer/supervisor conducts the ratings.

9. Back-Scratching Bias

This occurs when a *faculty member gives high ratings to peers on the exchange assumption that he or she will then receive high ratings*, kind of a "mutual admiration society" mentality. This mutual back-scratching is most common when faculty select their own peer raters.

Possible solutions: If these observers are *selected by an administrator*, such as a department chair, associate dean, or dean, and they are *trained in teaching observation*, back-scratching bias can be minimized and even eliminated.

10. Other Sources of Bias

There are other types of bias that pertain to self-report rating scales, where the respondent may have the tendency to give the socially desirable answer or the one that society regards as positive (*social desirability bias*), to intentionally attempt to create a false positive impression (*faking good*), or to respond with deviant answers to present a negative image (*faking bad/deviation bias*). The latter two types of bias, in fact, can occur in combination, known as the *hello-goodbye effect*.

Epilogue

Now, what are you supposed to do? You knew there were problems with ratings before you read this list. That list gave off the smell of aged cheese. Maybe your angst is kicked up another notch. At least I've given you putrid cheese to substantiate your angst. But all's not lost.

Let's connect some dots among those 10 types of bias to reboot your thinking about rating scale data. Knowing the specific weaknesses of a scale you are using to measure student or faculty performance or the outcomes of your faculty development program is important for four reasons: (1) you can take specific actions to minimize the sources of bias, (2) you can evaluate the types and degrees of bias that can affect the results, (3) you can consider the biases in the interpretation of the results, and (4) you can add other sources of evidence or scales for decision making. Collectively, this information can markedly improve the validity of the ratings from any single scale and the accuracy of your decisions based on multiple sources of evidence.

References

Berk , R. A. (2006). *Thirteen strategies to measure college teaching: A consumer's guide to rating scale construction, assessment, and decision making for faculty,administrators, and clinicians.* Sterling, VA: Stylus.

Nunnally, J., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Streiner, D. L., & Norman, G. R. (1995). *Health measurement scales: A practical guide to their development and use* (2nd ed.). New York: Oxford University Press.

Ronald A. Berk, Ph.D., is professor emeritus, biostatistics and measurement, and former assistant dean for teaching at The Johns Hopkins University. He can be contacted at rberk@son.jhmi.edu and www.ronberk.com.