Overlooking Overkill? Beyond the 1-to-5 Rating Scale

Robert B. Kaiser, Partner, and Robert E. Kaplan, Partner, Kaplan DeVries, Inc.

[Editor's Note: An earlier version of this article was presented at the 19th annual meeting of the Society for Industrial and Organizational Psychology in Chicago, Illinois, in April 2004. This article was condensed from a fuller version that has additional supporting statistical and anecdotal analysis.]

Over two millennia ago, Aristotle (trans. 1982) wrote in his *Ethics* that what is good, virtuous, and effective in thought and action is difficult to achieve. He noted that ineffectiveness is characterized either by *deficiency*—too little of the prized behavior—or by *excess*—too much of it. This old and worthy idea, that deficiency and excess constitute two fundamental classes of faulty performance, strikes most people as common sense. Nevertheless, the idea has somehow been overlooked in the design of formal systems and instruments commonly used to assess the performance of managers.

The Problem

The method of choice for measuring performance in organizations is the behavioral rating scale (Murphy & Cleveland, 1995). First applied to the problem of psychological measurement by Francis Galton late in the 19th century (Aiken, 1996), rating scales have evolved considerably over the last hundred years. Their modern form can be found in the now-ubiquitous 360° survey. These instruments typically employ a variation on Rensis Likert's (1932) solution for measuring attitudes, the Likert-type scale. In applying Likert's method to the measurement of performance, the "agree-disagree" response format has been modified to take one of two general forms.

Most common is the *frequency* type of response scale (Leslie & Fleenor, 1998). Rating formats of this "less-to-more" variety require raters to indicate how often the manager exhibits a particular behavior or how characteristic a particular statement is of that manager. Response options are ordered categories anchored by adverbs such as "never, sometimes, usually, often, always" to convey how often the manager engages in the described behavior. Or, to indicate how characteristic the descriptor is of the manager, the anchors might be something like "not at all, to a little extent, to some extent, to a great extent, to a very great extent." These scales carry the appearance of objectivity in that it is assumed that raters use them to merely *describe* the frequency of behavior (Nathan & Alexander, 1988).

The second kind of response scale is the *evaluation* type, in which the rater is asked to judge how effectively the manager performs the behavior, role, or function described by the survey item. There are two general classes of this "how well" variety of rating format: evaluation of performance in absolute terms and evaluation of performance in relative terms. *Absolute* evaluation scales contain response categories with adjective anchors such as "ineffective, adequate, good, effective, and exceptional." *Relative* evaluation scales require the respondent to compare the ratee's performance to some reference group—for example, with instructions and anchors such as "relative to other managers at Acme, this manager's performance is: among the worst, below average, average, above average, among the best."

The key distinction between frequency and evaluation response scales is that the former asks raters to *describe* performance whereas the latter requires raters to *judge* the quality of performance (Stockford & Bissell, 1949). There is another difference between these two types of scales: Each has a unique limitation when it comes to capturing excesses.[1]

An Illustration

Consider Rick Strong, a fictitious senior manager who resembles several executives we've worked with over the years. A keen analyzer of what works and what does not, Rick is extremely results-oriented and consistently achieves his objectives. Despite how productive he and his unit are, his staff has misgivings. In particular, they think Rick can be critical, sometimes verging on abusive, when they do not meet his lofty expectations. Moreover, he is short on praise—you definitely hear about it when you are not up to snuff, but rarely do you get a "good going" pat on the back. How would you rate Rick on the items with the frequency and

EXHIBIT 1

Rating Rick Strong with a Frequency and Evaluation Scale

	Frequency Scale					Evaluation Scale				
	Never	Rarely	Some-times	Often	Always	Ineffective	Adequate	Effective	Very Effective	Outstand-ing
Does whatever it takes to get results.	O	O	O	O	X	O	O	X	O	O
Makes judgments—zeroes in on what is not working.	O	O	O	O	X	O	X	O	O	O
Shows appreciation—helps people feel good about their contribution.	O	X	O	O	O	X	O	O	O	O

evaluation response scales presented in Exhibit 1?

The frequency scale fails to distinguish between very much and *too* much. There is no question that Rick "always" does whatever it takes and makes judgments, so he gets the highest rating on these items. And because "high" scores are taken to be ideal, there is an unstated assumption here that "more is better." This is unfortunate because it is widely understood that too much of a good thing is not so good. That is how strengths become weaknesses. But it is not likely Rick will get the message in this case. On the upside, the frequency scale does an adequate job of capturing deficiencies: The low rating on "shows appreciation" effectively indicates something Rick needs to do more often.

The evaluation scale introduces ambiguity at the other end of the register. What does Rick conclude from his merely "adequate" score on "Zeroes in on what isn't working"? Is he not discriminating enough or is he hypercritical? And a similar question can arise about his score on "Shows appreciation." Does the low score indicate he does not give enough praise or that he doles it out indiscriminately? Thus, although high scores on evaluation rating scales may reveal clear strengths, low scores are unclear. They muddle the distinction between deficiency and excess. Our point with this illustration is that the rating scales commonly used in practice are not adequate for detecting excess—when strengths are overused. This despite the widespread recognition that managers, the intense and driven lot that they are, can get into trouble by going overboard just as well as they can by being deficient (Kaplan & Kaiser, 2003a, 2003b; Lombardo & Eichinger, 2000; McCall, 1998; McCall & Lombardo, 1983).

A Solution

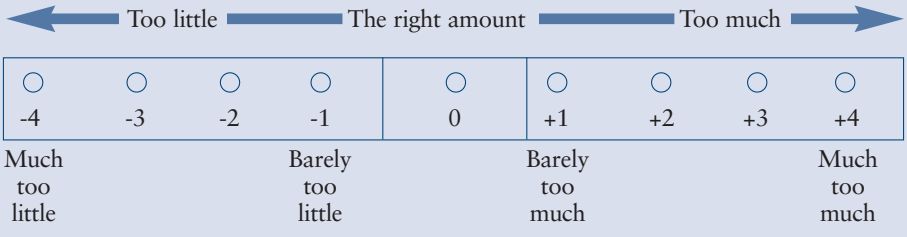
The limitations of traditional rating scales dawned on us in the early 1990s. The insight came out of comprehensive assessments of executives that involved extensive interviews with coworkers past and present as well as a battery of psychological tests and 360° ratings. In the course of helping his clients make sense of their data, Bob Kaplan stumbled on the oversight (see Kaplan, 1996). He found himself remarking, "You are a force to be reckoned with." It followed that he would sum up their shortcomings with the phrase, "too forceful." It was plain as day in the interview data, whether direct reports were bemoaning an autocratic style, peers were complaining about never getting a word in edgewise, or superiors were concerned about an intense drive. Something just did not add up: None of the 360° ratings directly indicated overkill.

Looking for a way to correct for this limitation of existing 360° instruments (including his own, *SKILLSCOPE® for Managers* (Kaplan, 1988)), Kaplan (1996) devised what he called a "curvilinear" rating scale. Low ratings were anchored with "too little," high ratings were anchored with "too much." And like Goldilocks' favorite porridge, the optimal rating, in the middle, was anchored with "the right amount." Rob Kaiser has joined Kaplan in conducting ongoing research and refining the new rating scale and a prototype 360° questionnaire, now called the *Leadership Versatility Index®*.

In its present form, the new response scale looks like the one in Exhibit 2. Raters are alerted that scale is not simply less-to-more where "more is better." For instance, minus scores on the deficiency side and plus scores on the excess side call attention to these two different types of performance problems. According to recent developments in the study of mental processes involved in making ratings, the negative and positive numbers (and the arrows) also convey to raters that each side of the scale is distinct: Low is not a lack of high, it is the opposite of it (Schwartz, 1999). The scale is be a powerful way to tease apart the two types of ineffective performance in developmental feedback.

EXHIBIT 2

The Implicitly Curvilinear, "Too Little/Too Much" Response Scale



The "too little/too much" response scale combines elements of both the frequency and evaluation format because it contains descriptive (how much?) as well as judgmental (how well?) components. Also, this scale appears to takes context into account: It implies a judgment of frequency relative to *this job in this organization at this time*.

A project for a client led to the development of another version of our rating scale. Motorola Inc. commissioned us to help develop a leadership model and attendant performance measures to be used with its top 1,000 executives (Kaiser, et al., 2002). Motorola approached us because senior management was taken by our "too little/too much" scale and wanted to employ it in their tool. But there was also a need for a traditional effectiveness scale because the results would be used both for development and for administrative purposes and because the company needed to compare scores directly among individuals. Motorola therefore decided to use two rating scales, an *evaluation* scale and an adaptation of our new scale designed to complement an evaluation scale, the "do less/do more" scale shown in Exhibit 3. We describe later how this scale complements an evaluation response scale by clarifying the meaning of "less effective" ratings.

Benefits

Through our consulting practice and program of basic research, we have found several advantages of this new design for response scales. These benefits accrue to raters, feedback recipients, organizations, and researchers. We also have some concerns and questions that need to be addressed. First, the benefits.

Benefits to Raters

In introducing this new approach to groups of managers, we find two striking results. First, the "too little/too much" distinction is not hard to grasp: People intuitively seem to understand it. Second, some people report feeling less constrained in making assessments using the new scale. Others can see that the response scale adds

EXHIBIT 3

The Prescriptive "Do Less/Do More" Scale for Supplementing Evaluation Scales

Do a lot less	Do less	Do a little less	Do the same	Do a little more	Do more	Do a lot more
○	○	○	○	○	○	○
-3	-2	-1	0	+1	+2	+3

new possibilities—but tend to be at a loss for fully explaining how. When we ask them to contrast this experience to their experience with traditional scales, we hear things like: "Well, I'm not always sure what a '3' is supposed to mean," or "I usually use the middle two values, but on this scale I couldn't because they weren't always a strength—it forced me to use more of the options." Sometimes we also hear: "This scale allowed me to indicate, 'yes, you are strong in that area, but sometimes a little too strong.'"

Benefits to Feedback Recipients

There are two benefits of the new scale to feedback recipients. One, what the results mean is much clearer. Low scores on evaluation scales are ambiguous, and high scores on frequency scales do not draw the line between plenty and too much, but the "curvilinear" scale leaves little doubt what the results mean when they are cast in terms of "too little," "the right amount," and "too much." As one director of talent management whose firm has adopted our model of leadership and tool said: "There is a confidence in interpreting results—you know right away what to do about it, whether it's step up, tone down, or do more of the same."

The second benefit is a better spread using this response scale than using standard response scales. In recent years, a common complaint heard in organizations is that "everyone gets high scores on everything." In other words, ratings do not appear to discriminate within a person (that is, distinguish between his or her strengths and weaknesses) or between people (that is, distinguish between higher and lower performers). No doubt, one reason is that raters mostly use only a portion of the typical five-point scale. This is to be expected: Through "corporate Darwinism" individuals selected into management positions are the ones who have the ability, motivation, and experience to do the job (LeBreton, et al., 2003). Rating distributions get heavily skewed toward the top end, especially over time as junior managers get better through experience.

The "too little/too much" scale also helps spread scores out. First, because the optimal score is in the middle of the scale, frequency distributions tend be relatively normal and centered. Second, because deficiency and excess are teased apart, there is a generous spread in both directions surrounding optimal. Finally, because the response scale is effectively nine points (-4 to +4), nearly double the typical scale (1 to 5), scores are distributed over a wider range and differences are more readily apparent to the naked eye.[2] Thus, when it comes to making sense of feedback results, the curvilinear scale provides an advantage by spreading scores out and by distinguishing between too little and too much.

Benefits to Organizations

In our work with Motorola, we learned firsthand how the idea of accounting for overkill and an application of that idea in the form of a performance-appraisal tool can have an impact on an organization (Kaplan & Kaiser, 2003a). Recall that we designed a leadership model and tool for them that involved two ratings for each item—an absolute *evaluation* rating and a *prescriptive* "do less/do more" rating. The first thing we learned was how the basic idea of excess can expand the language an organization uses to discuss leadership and development. Second, assessing individuals in terms of "too little and too much" as well as absolute effectiveness with an evaluation scale packs a powerful one-two informational punch for decision makers.

Senior leaders at Motorola wanted to reflect the tensions and trade-offs inherent in the business world in their model and measures of leadership. They were talking about a kind of leadership that navigated the straits and avoided crashing on one side or the other: for instance, balancing vision with execution and balancing "edge," the tough side of leadership, with empowering and supporting people. The idea that problems come in both flavors, deficiency and excess, played naturally to this view: Out-of-balance leadership could easily be described as too much focus on execution, not enough vision; too much pushing for results, not enough support; and so on. By recognizing overkill explicitly in their model, tools, and conversations, senior leaders at Motorola created a leadership culture that was wary of excesses. They also provided a new way to appreciate agility and the daunting trade-offs with which senior managers must contend.

One senior HR person remarked a few years after launching the model and assessment tools: "What's most fascinating are those cases where the person gets a relatively high effectiveness rating on an item like 'Expects a lot,' but several coworkers also indicate 'do less.' These tend to be the fast-trackers who risk derailing because their intensity can become too much. The level of dialogue in these sessions is amazing. You can see the light bulb go on."

On a broader scale, weaving the idea of overkill directly into the fabric of their leadership model and 360° tools has opened the door to capitalizing on other developments in the field. For instance, Motorola has incorporated Eichinger and Lombardo's (2000) *For Your Improvement* (FYI) development guide in their e-learning system. Not coincidentally, *FYI* is one of the few resources that explicitly address how strengths become weaknesses through overuse. The HR/OD team at Motorola has mapped the behaviors assessed by the 360° onto the dimensions in FYI so feedback recipients have, literally at their fingertips, tips on what to do about skills they lack as well as those they have overdeveloped.

In addition to developmental applications, measuring behavior in terms of too little and too much adds to the tool's predictive power. The "do less/do more" ratings furnish information that is distinct from that provided by the effectiveness ratings. "Calibration" is an annual process by which managers at Motorola get together and decide where each of their subordinates falls out in a forced distribution—least effective, solidly effective, or most effective. To determine the value-added of the "do less/do more" scale, we first used ratings on the eval-

uation scale to predict calibration ranks and then tested whether the "do less/do more" ratings add to the tool's ability to predict. We've been doing this analysis every year since 2000 and have found that the "do less/do more" ratings increase how well scores on the 360° predict calibration rankings by at least 25 percent; one year, it enhanced predictive power by 55 percent.

Our statistical analyses also revealed that the "do less/do more" ratings help primarily by clarifying the low-to-middling ratings on the evaluation scale. Perhaps an example will illustrate this best: On the item "Holds people accountable," one manager received an effectiveness rating of 3, and no one indicated do more or do less; another manager also received an effectiveness rating of 3, but five coworkers indicated "do more." Clearly, the former manager is in better shape than the latter. In this way the "do less/do more" scale helps the supervisor as well as the manager receiving feedback determine what to work on.

Benefits to Researchers

Finally, we have discovered at least two benefits of the new response scale for students of management. First, precisely because the new response format was designed on a curvilinear principle, it helps in detecting curvilinear relationships between managerial behavior and various criteria. Not surprisingly, we routinely detect curvilinear relationships between measures of effectiveness and leadership dimensions measured with our evaluation of frequency scale.

A second benefit is that the new response scale clears up an anomaly in the body of research on opposites in leadership (e.g., task-oriented versus people-oriented). In recent years interest has increased in the paradoxes that confront modern managers and, by extension, in the notion of managerial flexibility or versatility (Kaiser, et al., 2005). One would expect a negative correlation between opposites like short-term orientation versus long-term orientation, competition versus collaboration, autocratic versus participative, and so forth. That is, we would expect that doing too much on one side in each pair of opposites would correspond to doing too little of the other side or that being more skilled at one would correspond with being less skilled at the other (Kaplan, 1996; Kaplan & Kaiser, 2003b). The research literature is clear on this point: When measured with a traditional response scale, correlations between ratings on these theoretical opposites are actually positive, often on the order of .50 or so. When opposites are measured using the "too little/too much" scale, a very different pattern emerges: we find negative correlations around -.50. How to account for the wildly discrepant results? We think the difference comes from the type of response scale employed: Traditional scales only cover half the story by stopping short of excess; by not allowing for the possibility of overkill, they therefore cannot detect lopsidedness.

This statistical finding is not just a researcher's concern: It is also relevant to practice. The positive correlation found using traditional response scales means that most managers get feedback that says: "The more skilled you are at this, the more skilled you are at its opposite too." The negative correlation for ratings on the new scale means these managers hear: "The more you overuse this skill, the more likely you under-use the complementary skill," thus pulling the lopsidedness of their leadership into sharp relief.

Concerns and Further Development

Here are the major concerns that have occurred to us or that have been raised by our colleagues.

Some Things Cannot Be Overdone

This is something we frequently hear, particularly from scholarly researchers. For instance, some people claim that you cannot be too smart. And in an age in which visionary leadership is all the rage, some have argued that today's leaders cannot be too strategic. We disagree with these claims on the grounds of research. For example, after studying three different samples of managers, Ghiselli (1963, p. 898) concluded: "…the relationship between intelligence and managerial success is curvilinear with those individuals earning both low and very high scores being less likely to achieve success in managerial positions." Similarly, in the 360° data we collect, leaders do get faulted by their coworkers for being too strategic: Too much time on strategic planning, grandiose visions that defy implementation, pushing growth too far and too fast, and so on. With regard to the larger claim that some things simply cannot be taken too far, that may be true. Some experts question even this moderate stance. For instance, McCall (1997; pp.35-29) took the opposite view in a section of *High Flyers* titled: "Every Strength Can Be a Weakness."

We do not know for sure that all leadership behaviors can be overdone, but clearly many can. A key lesson we have learned in using the new response format is that items must be phrased in a way that helps the respondent easily see what "too much" of that behavior might look like. Using items that are value-laden will not work. For instance, "Effectively makes her point to a resistant audience" will not work because one cannot be *too effective*. But "Persists in trying to persuade people" does admit to overdoing.

Difficulty Creating Scale Scores

Another limitation involves the computation of scale scores across several items rated on the -4 to +4 scale. The problem occurs when some items are in the negative, "too little" region, but others are in the positive, "too much" region. The net effect is for the scores to cancel each other out and to dilute the average, bringing it closer to zero, optimal, than ought to be the case. We have yet to discover a satisfying solution to this arithmetic problem. We simply suggest caution with scale scores, recognizing that no measure is perfect. For now we regard the dilution that occurs from the way that positive ratings and negative ratings cancel each other out as a cost of making room to detect overkill.

Sometimes a Linear, Absolute Measure Is Needed

One of the strengths of the new response format is that it takes context into account to some degree. This is especially helpful in development: The focus on using the data is specific to one person. But in other applications, particularly administrative uses of ratings where data is used to compare people, this can be a drawback. For instance, some academics have questioned whether it makes sense to compare ratings for two different people on the new scale. As the argument goes, if the scale does assume a great deal of context, then scores between people in different contexts (e.g., different jobs, different organizations) are not comparable.

We take these concerns seriously and have begun a study aimed at investigating them;

however, at this point we are relatively confident that comparing ratings for two or more people on the new scale makes some sense. Our confidence comes from a simple empirical fact: Our cross-sectional research consistently yields sizable correlations between behaviors measured on the new scale and external criteria (e.g., leader effectiveness, subordinate satisfaction). If between-person comparisons were invalid, these correlations would equal zero.

No Direct Comparisons Between Alternative Response Scales

Astute methodologists will note we have made several direct conceptual comparisons between the new response format and traditional response formats, yet have only made indirect empirical comparisons. Many of our claims remain hypotheses about how the two methods would compare directly. Specifically, what is needed is an experimental study with a controlled design that involves having the same respondents rate the same target manager on a set of dimensions, once with the new scale and once with a traditional scale. A study like this could provide control adequate to ruling out competing explanations for the observed results, and could isolate the effects of each type of response scale. We currently have such a study under way.

Concluding Thought

We are optimistic about this innovation in response scale technology, but only cautiously so. There is still much to learn about how best to apply the new scale in practice and in research. We encourage other independent research teams to conduct their own studies of the strengths and limitations of this new format. To that end, we would gladly share whatever materials and thoughts interested parties may need to get started.

NOTES

[1] Modern approaches to leadership development usually recognize how strengths can become weaknesses when overused. This idea has been widely disseminated in the work of M. Lombardo and M. McCall (Lombardo & Eichinger, 2000; McCall, 1998; McCall & Lombardo, 1983). The idea that excesses constitute just as important a class of performance issues as deficiencies is rarely reflected in the design of standard assessment tools. When it is taken into account, it tends to be treated as an afterthought or as a supplemental feature rather than as integral to the design of the measure. See examples in Leslie and Fleenor (1998).

[2] Although there is more variance in an absolute sense with our new scales, this is something of a methodological artifact because our scale has nine intervals and typical scales have only five intervals. The average SD on our scale is .82, which is about .09 units on the native scale (.820/9). Typically, performance ratings on five-point scales have an SD around .50 (.10 units on the native scale). Thus, there is *relatively* less variance on our scale, controlling for number of response options. There is more variance in absolute terms, which may be more important given the near-universal practice of providing 360° results as raw scores, on the original metric established by the response scale (Leslie & Fleenor, 1998).

REFERENCES

Aiken, L.R. (1996). *Rating Scales and Checklists: Evaluating Behaviors, Personality, and Attitudes*. New York: John Wiley & Sons.

Aristotle (undated). *Nicomachean Ethics*. Translated by H. Rackham (1982). Cambridge, MA: Harvard University Press.

Eichinger, R.W. & Lombardo, M.M. (2000). *For Your Improvement*. Minneapolis, MN: Lominger Limited, Inc.

Ghiselli, E.E. (1963). "The Validity of Management Traits in Relation to Occupational Level." *Personnel Psychology*, 16, 109-113.

Kaiser, R.B., Craig, S.B., Kaplan, R.E., & McArthur (2002). "Practical Science and the Development of Motorola's Leadership Standards." In K.B. Brookhouse (Chair) *Transforming Leadership at Motorola*. Practitioner Forum presented at the 17th Annual Conference of the Society for Industrial and Organizational Psychology, Toronto, Ontario.

Kaiser, R.B., Lindberg, J.T., & Kaplan, R.E. (2005). "Assessing the Flexibility of Managers with Coworker Ratings: A Comparison of Methods." Manuscript under review.

Kaplan, R.E. (1996). *Forceful Leadership and Enabling Leadership: You Can Do Both*. Greensboro, NC: Center for Creative Leadership.

Kaplan, R.E. (1988). *SKILLSCOPE®* for Managers. Greensboro, NC: Center for Creative Leadership.

Kaplan, R.E. & Kaiser, R.B. (2003a). "Developing Versatile Leadership." *MIT Sloan Management Review*, 44, 19-26.

Kaplan, R.E. & Kaiser, R.B. (2003b). "Rethinking a Classic Distinction in Leadership: Implications for the Assessment and Development of Executives." *Consulting Psychology Journal: Research and Practice*, 55, 15-25.

LeBreton, J.M., Burgess, J.R.D., Kaiser, R.B., Atchley, E.K., & 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, 78-126.

Leslie, J.B., & Fleenor, J.W. (1998). *Feedback to Managers: A Review and Comparison of Multi-Rater Instruments for Management Development*. Greensboro, NC: Center for Creative Leadership.

Likert, R. (1932). "A Technique for the Measurement of Attitude Scales." *Archives of Psychology*, 140, 44-53.

Lombardo, M.M. & Eichinger, R.W. (2000). *The Leadership Machine*. Minneapolis, MN: Lominger Limited, Inc.

McCall, M.W. Jr. (1998). *High Flyers: Developing the Next Generation of Leaders*. Boston, MA: Harvard Business School Press.

McCall, W.M. Jr. & Lombardo, M.M. (1983). *Off the Track: Why and How Successful Executives Get Derailed*. Greensboro, NC: Center for Creative Leadership.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding Performance Appraisal: Social, Organizational, and Goal-Based Perspectives*. Thousand Oaks, CA: Sage.

Nathan, B.R., & Alexander, R.A. (1988). "A Comparison of Criteria for Test Validation." *Personnel Psychology*, 41, 517-535.

Schwartz, N. (1999). "Self Reports: How the Questions Shape the Answers." *American Psychologist*, 54, 93-105.

Stockford, L. & Bissell, H.W. (1949). "Factors Involved in Establishing a Merit-Rating Scale." *Personnel*, 26, 94-116.