

**OVERLOOKING OVERKILL:
ON THE FOLLY OF LINEAR RESPONSE SCALES FOR A NONLINEAR WORLD**

Rob Kaiser & Bob Kaplan

KAPLAN DEVRIES INC.

Over two millennia ago, Aristotle (undated/1982) wrote with exquisite detail in his *Ethics* about how 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 today. 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 these days for measuring performance 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-plus years. Their modern form can be found in the now-ubiquitous 360° survey, routinely used in today's organization to measure the performance of managers. These instruments typically employ a variation on Rensis Likert's (1932) famous solution for measuring attitudes, the summative rating scale. A Likert-type scale is a set of items intended to measure a given construct. Responses to each item are summed to arrive at an overall score—hence, *summative* rating scale. Originally, items were rated for the extent to which the respondent agreed with it. In adapting Likert's method to the measurement of performance, the "agree-disagree" response format has been modified to take one of two general forms.

Two types of response scales. 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 in question 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 performs the described behavior. Or to indicate how characteristic the item 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." This type of scale carries the appearance of objectivity in that it is assumed that raters use it to merely *describe* behavior (Nathan & Alexander, 1988). Moreover, there are empirically derived guidelines for assigning adverbs to anchor response categories that approximate an equal interval measurement scale (e.g., Bass, Cascio, & O'Connor, 1974; Spector, 1976). Perhaps for these reasons, the frequency scale appears to be the response format of choice for measuring performance on questionnaire surveys (Shipper, 1991).

Presented in S.B. Craig (Chair), *360, The Next Generation: Innovations in Multisource Performance Assessment*. Symposium presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL, April 2004.

Rob Kaiser is the director of R&D and Bob Kaplan is a partner at Kaplan DeVries Inc., a firm that specializes in consulting to executives and executive teams on leadership. Inquiries about the research described in this paper may be sent to: Kaplan DeVries Inc., 1903 G Ashwood Court, Greensboro, NC 27455. Electronic mail may be sent to: rkaiser@kaplandevries.com or rkaplan@kaplandevries.com.

The second kind of response scale is the *evaluation* type, where the rater is asked to judge how effectively the manager performs the behavior, role, or function described by the survey item. This response scale is rare in practice. For instance, of the 24 feedback instruments in Leslie and Fleenor's (1998) comprehensive review, only two were equipped with an evaluation scale. There are two general classes of this "how well" variety of rating format: those that evaluate performance in absolute terms and those that evaluate performance in relative terms. Absolute evaluation scales contain response categories with adjective anchors such as "ineffective, adequate, good, effective, 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). However, 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 Rodney Strong, a fictitious senior manager who resembles several executives we've worked with over the years. A keen analyzer of what works and what doesn't, Rodney is clearly in charge of his group and always hits the numbers. Despite consistently making plan and coming in under budget, his team does have some misgivings. In particular, they think Rodney can be very critical, often verging on abusive, when they miss his lofty expectations. Moreover, he's short on praise—you definitely hear about it when you aren't up to snuff, but rarely do you get a "good going" pat on the back. How would you rate Rodney on the items with the frequency and evaluation response scales presented in Figure 1?

Figure 1. Rating Rodney Strong with a frequency and evaluation scale.

	Frequency Scale				Evaluation Scale			
	never	rarely	usually	always	ineffective	adequate	effective	outstanding
Does whatever it takes to get results.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
Zeros in on what isn't working—makes judgments.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
Shows appreciation—helps people feel good about their contribution.	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

¹ We hasten to acknowledge that 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 (Lombardo & Eichinger, 2000; Lombardo & McCauley, 1988; McCall & Lombardo, 1983). The idea that excesses constitute just as important a class of performance issues as deficiencies, however, 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 integral to the design of the measure. For instance, there are instruments that render prescriptions for development by comparing ratings of, on the one hand, "how often" the manager does a particular thing to, on the other hand, an "ideal amount" that is estimated using a statistical formula. And there are a few instruments that have respondents rate how often the manager engages in a number of specific behaviors, and then at the end, ask respondents to indicate whether the manager should do more, less, or the same amount of not each specific behavior, but of the few dimensions those several behaviors comprise. See examples in Leslie and Fleenor (1998).

The frequency scale is flawed because it fails to distinguish between very much and *too* much. There is no question that Rodney "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 gold, there is an unstated assumption here that "more is better." This is unfortunate in developmental feedback, because it's axiomatic that too much of a good thing isn't so good. That's how strengths become weaknesses. But it isn't likely that Rodney will get the message in this case. On the upside, however, the frequency scale does do an adequate job of capturing deficiencies: the low rating on "shows appreciation" effectively indicates an area in which Rodney needs to get better.

The evaluation scale introduces ambiguity at the other end of the register. What are we to make of low scores? Does Rodney conclude he is merely "adequate" at making judgments because he's not discriminating enough? Or because he's hypercritical? And although it is a bit of a stretch, a similar question can arise about showing appreciation: does the low score indicate a lack of it or an issue with doling out praise indiscriminately? Thus, while high scores on effectiveness rating scales may reveal clear strengths, low scores are not prescriptively clear. They confound the distinction between deficiency and excess. Since feedback is all about providing insight to those who lack it, leaving managers to their own devices to interpret low scores is risky business.

Our point with this illustration is that the rating scales commonly used in practice aren't adequate for detecting excess, like when strengths run amok. 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.

Solutions

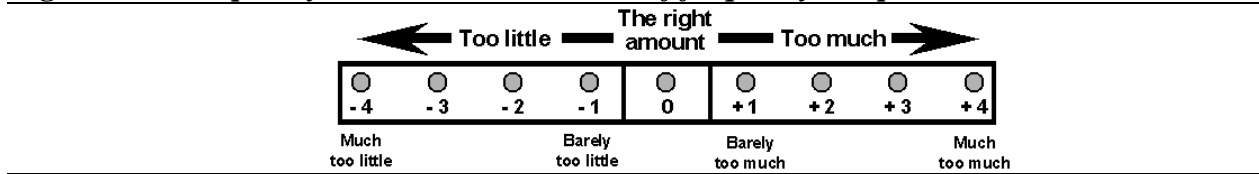
The limitations of traditional rating scales dawned on one of us, Bob Kaplan, in the early 1990s. The insight came out of comprehensive assessments of executives that involved extensive interviews with coworkers past and present as well as a battery of psychological tests and a 360° instrument. In the course of helping his clients make sense of their data, Kaplan stumbled over the problem (see Kaplan, 1996). He found himself remarking to many of them, "You are a force to be reckoned with." And then it followed logically 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, whether peers were complaining about never getting a word in edgewise, or whether superiors were concerned about an intense drive. But there was something that just didn't add up: none of the 360° ratings directly indicated overkill like this.

Frustrated by the limits 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 "the right amount." Rob Kaiser has joined Kaplan in conducting ongoing research and refining the prototype 360° questionnaire, which we now call the *Leadership Versatility Index*®.

In its present form, the new response scale looks like the one in Figure 2. Minus scores on the deficiency side and plus scores on the excess side help cue raters that this scale is not simply less-to-more where "more is better," but, in fact, implicitly "curvilinear." And according to recent developments in the study of cognitive 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). We have found the scale to be a powerful way to tease apart the two types of ineffective performance in describing behavior for developmental feedback.

Figure 2. The implicitly curvilinear "evaluation of frequency" response scale.



This response format could be described as an *evaluation of frequency* scale because it contains both descriptive (how much?) and judgmental (how well?) components. Also, it would appear that the scale takes context into account: it seems to imply a judgment of frequency relative to *this job in this organization at this time*.

After the development of this scale, we were commissioned by Motorola Inc. to help develop a leadership model and attendant performance measures to be used with their top 1,000 or so executives around the world (Kaiser, Craig, Kaplan, & McArthur, 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. Results would be used for both development and administrative purposes, and they needed to directly compare performance between individuals. On the instrument we created together, raters are first asked to judge effectiveness on each item in absolute terms and then to *prescribe* changes for improving, using the scale below, an adaptation of our scale for this purpose.

Figure 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

We'll present data below to demonstrate how this scale complements an *evaluation* response scale. As we demonstrate below, it clarifies the interpretation of lower ratings.

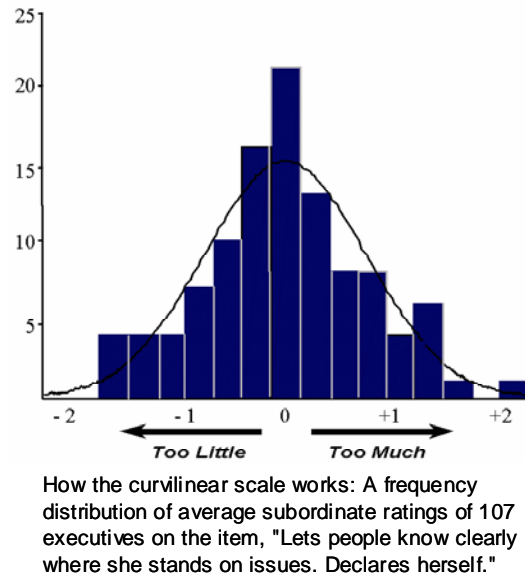
Utility

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

Benefits to raters. In workshops we run introducing this new approach, we ask respondents to complete a short version of an instrument that employs the curvilinear "evaluation of frequency" response format. Afterwards, we ask people to describe what it was like. We have also conducted protocol analyses to understand the mental processes involved in filling out a 360 with this scale, asking raters to "think aloud" while completing the form (Kaiser et al., 2002). In both scenarios, we find two striking results. First, the too little/too much distinction is not hard to grasp—people intuitively seem to get it. And we find a corresponding phenomenon empirically:

in large samples, ratings are nicely distributed across the continuum from "too little" to a peak near "the right amount" and a tailing off on the "too much" side. Below is a near perfect example, based on ratings from over 500 subordinates of 107 executives to an item on our instrument, the *Leadership Versatility Index*[®] (Kaplan & Kaiser, 2002).

Figure 4. Sample frequency distribution of ratings on the new curvilinear response scale.



The second thing we have found about the rater's experience is that people report feeling like they were less constrained in their assessments using the new scale. Common are statements like "I felt like I could better express what I was thinking," or "It let me say what I wanted to" or "I felt like he (the person being rated) would get it." One telling thought regarding the use of the too much side of the scale: "This [item] is clearly a strength. But I think sometimes he leads with it and doesn't have to. He could tone it down a bit and be even more effective. I'd say a +1, on the too much side."

Other raters can see that the response scale is new and opens up 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," or "I feel like I'm being forced to squeeze my perception into a narrow box [with traditional scales]."

To the extent that traditional response scales impose artificial constraints or cause different raters to interpret response categories differently, we would expect inter-rater reliability to be adversely affected. Or, to state it positively, if the new scale removes some of these barriers, we would expect inter-rater reliability to be higher.² In fact, this is what we find when we compare inter-rater reliability estimates on our instruments to Conway and Huffcut's (1997) meta-analytic estimates of the inter-rater reliability of ratings of managerial performance that use traditional response scales. After correcting for unreliability, they reported an average correlation between two raters in the same source of .44 for superiors, .36 for peers, and .30 for subordinates. In contrast, and without any corrections, we found comparable intra-class correlations (Shrout &

² We are grateful to an anonymous reviewer for pointing this out.

Fleiss, 1979) that were somewhat higher. Specifically, these correlations (averaged across dimensions) were .50 for superiors, .44 for peers, and .47 for subordinates (Kaiser & Craig, 2001). Of course, we're not strictly comparing apples to apples here—there are other differences than just the response scales (e.g., items). A definitive test would involve having the same respondents rate a set of identical dimensions twice—once with the new scale, once with a traditional scale. Nonetheless, this preliminary indication is consistent with the idea that the new response scale affords more opportunity to express perceptions of performance.

Benefits to feedback recipients. The primary benefits of the new scale for those being rated revolve around clarity in interpreting results. Whereas low scores on evaluation scales leave it open as to the nature of the ineffectiveness and high scores on frequency scales don't draw the line between plenty and too much, there is a straightforward diagnosis when results are cast in terms of "too little," "the right amount," and "too much." As one director for corporate talent development at a firm that has adopted our framework and tool said, "There is a confidence to interpreting results—you know right away what to do about it, whether it's step up, tone down, or keep it up with more of the same."

Additional clarity comes from the fact that the new response format takes context into consideration. When coworkers say a manager does too much of something or that she should do it less, the message is clear. These raters, of this person, in this job, in this organization, at this time say that this behavior is taken to a counterproductive extreme. No matter how central the given behavior is to performance, there is such a thing as *too much*. And having raters indicate this directly is far more persuasive than trying to explain how "research indicates that 'usually' doing this is more effective than 'always' doing it."³ Especially when we are so used to thinking of higher scores as "better."

In recent years, we've heard much grumbling from senior managers over competency models and 360° degree feedback as practiced in organizations. One common complaint is that "everyone receives high scores on everything." In other words, ratings don't appear to discriminate within people (across dimensions) or between people (separate higher from lower performers). No doubt one reason is that raters mostly use only a portion of the typical five-point scale. For instance, in one data base involving 360° ratings from over 5,000 raters using an evaluation type response format, we found that 73% of the item ratings were either a 3 or a 4. About 10% were a 5 and the remaining 17% were either a 1 or a 2. Once you average across raters and items within each dimension, the result is that about 85% of scale scores fall between 3.25 and 4.25. We found a similar pattern in a large 360° data set (ratings from over 100,000 raters across a range of industries) based on a frequency response scale. Specifically, about 80% of scale scores fell between 3.25 and 4.20. With most scores clumped together at the high end of

³ For instance, some instruments reviewed by Leslie and Fleenor (1998; e.g., Acumen's *Leadership Skills*, Clark Wilson Group's *Survey of Executive Leadership*, and *SYMLOG*) do provide prescriptive feedback around the do less/do more principles. On these instruments, individual behavior items are rated on a traditional frequency scale. Then scale scores are compared to an "ideal amount" based on an "expert system" that amounts to statistically estimating what the optimal level is for that individual in that job. When scores are higher (lower) than this "ideal point," the prescription is to do less (do more). Again, the raters are not asked to render such a prescription—the statistical model does. This statistical approach suffers the obvious limitation that, to use an extreme example, what is an optimal degree of "holding people accountable" at the American Red Cross may be vastly different from that at General Electric or the Santa Barbara Department of Corrections. Moreover, any statistical model based on Bayesian principles is fallible to some degree. One reason is that the statistical model lacks the contextual richness that is naturally a part of a rater's frame of reference.

the distribution, it is no mean feat for the naked eye to detect peaks and valleys worth discussing. In other words, relative strengths and developmental needs are hard to discern.

LeBreton and colleagues (LeBreton, Burgess, Kaiser, Atchley, & James, 2003) have argued that this is to be expected. In sum, their position is that "corporate Darwinism" selects individuals into management positions who have the ability, motivation, and experience to do the job. Those who cannot and/or will not perform the kinds of behaviors typically described on 360° instruments tend to not be in those roles—hence, rating distributions tend to be heavily skewed.

The new *evaluation of frequency* type of scale appears to help spread scores out. First, because the optimal score is in the middle of the scale, frequency distributions tend to 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 9 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.⁴

One of our clients recently remarked about how the constriction in typical scales is troubling. He told us, "Our company uses a five-point scale in the 360° that is part of our annual performance appraisals. What happens is almost all the respondents give you a 3 or 4. They don't use the full scale. So what you get is this ridiculous thing where your highest score is 3.9 and your lowest is 3.3. So the 360° doesn't differentiate. It doesn't identify the problem players. And it doesn't help you sort out your strengths and weaknesses."

An example to demonstrate how the new response format helps: An executive we recently did an assessment for had had one done internally just a year earlier using his company's own 360° survey. Although he wanted to improve, he felt that the questionnaire results gave him little to go on. The data were based on ratings using the standard frequency response scale that ranged from 1, "to a very small extent," to 5, "to a very great extent." The problem was that only 14 of 110 items lay outside the range of 3.25 and 4.25, and even then only slightly. Nothing much stood out. And as a result nothing much came of that assessment. On the in-house questionnaire he got ratings in the low 3s on "delegates effectively" and "listens well." The results using our instrument and the curvilinear scale painted a much more complete and vivid picture. He received scores in the *too much* range on taking charge (+1.27) and stepping in when problems come up (+1.73) and scores deeper into the *too little* zone on empowering people (-2.13) and trusting people to handle problems (-1.80).

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 a concrete instantiation of that idea in the form of a performance-

⁴ We should point out that while there is more variance in an absolute sense with our new scales, this is something of a methodological artifact due to the fact that our scale has nine intervals and typical scales have only five intervals. The usual observed *SD* on a traditional scale is around .50 (in two large data sets, we found average *SD*s of .51 for an evaluation response scale and .56 for a frequency scale), which is about .10 to .11 units on the native scale (.51/5 ~ .10; .56/5 ~ .11). The average *SD* on our scale is .82, which is about .09 units on the native scale (.820/9). Thus, there is *relatively* less variance on our scale when you control for number of response options. But 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).

appraisal tool can impact 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 more/do less" 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, evaluating individuals in terms of too much/too little as well as absolute effectiveness packs a powerful one-two information 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 energizing people while at the same time facing up to uncomfortable realities. 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 balance and the daunting trade-offs senior managers must contend with.

One senior HR person remarked a few years after launching the model and assessment tools, "It's those cases where the person gets relatively high effectiveness ratings but also prescriptions to do less that are most fascinating. These tend to be the fast-trackers that risk derailing because their intensity can become enough already. The level of dialogue in these sessions is amazing. You can see the light bulb go on at the flash of insight."

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 licensed the content in Eichinger and Lombardo's (2000) *For Your Improvement (FYI)* development guide to incorporate in their e-learning systems that support training and development. *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 in their model and 360° onto the dimensions in *FYI* so after individuals receive feedback they have, literally at their fingertips, access to a host of concrete suggestions and tips for what to do about behaviors and skills they lack—as well as those they've developed to hypertrophy.

In addition to these developmental applications, actually measuring behavior in terms of too little/too much opens up new possibilities to organizational decision makers. Specifically, we found that the *prescriptive* "Do More/Do Less" scale yields substantial supplemental information to augment that provided by the traditional evaluation scale. For four years running, we've analyzed the data generated by the Motorola tools. One of the things we've studied is the value of the supplemental *prescriptive* scale. In 2001, the scale included seven points, ranging from -3 (Do a lot less) to +3 (Do a lot more). In an effort to reduce the cognitive load on raters, the scale was shortened to three points in 2002 (-1 = Do less, 0 = Do the Same, +1 = Do more).

Each year we use leadership ratings to predict *calibration rankings*. Calibration is the process by which managers convene annually to discuss their direct reports among one another. The outcome and purpose is to use a forced distribution to assign every manager in one of three categories—least effective, solidly effective, and most effective. It's a rigorous process that requires all of the superiors discussing personnel to reach consensus. This criterion represents where each manager falls out in the pecking order, relative to his or her peers.

To determine the value-added of the "Do More/Do Less" scale, we conduct a hierarchical regression analysis to predict calibration ranking. Superior and subordinate ratings of leadership made with the evaluation scale are entered into the equation in step one. In step two, we test for the change in variance accounted for by entering superior and subordinate ratings on the "Do More/Do Less" scale (both main effect terms and their cross-products to model the quadratic curvilinear relationship; Cohen & Cohen, 1983). Every year, we've found that the additional information provided by the Do More/Do Less scale increases prediction of calibration by at least 25%; in one year, it enhanced prediction by 55%.⁵

These results suggest that the Do More/Do Less ratings are furnishing information that is both unique from the effectiveness ratings and important to how senior managers are regarded by higher-ranking executives. To further investigate how this works, we calculated the cross-tabulation representing how often the Do less, Do the same, and Do more ratings were made in conjunction with each of the five levels on the effectiveness response scale across each of the 33 items on the instrument. The results for the data in 2002 are shown in Table 1.

Table 1. Cross-tabulation of Do More/Do Less Ratings by Evaluation Ratings

<i>DM/DL Scale</i>	<i>Evaluation (Effectiveness) Scale</i>					Total	Frequency
	1	2	3	4	5		
Do less	154	376	642	1064	315	2551	1.2%
Do the same	54	4489	43429	83753	21148	152873	69.8%
Do more	5025	22082	28403	7552	364	63426	29.0%
Total	5233	26947	72474	92369	21827	218,850	
Frequency	2.4%	12.3%	33.1%	42.2%	10.0%		

These results suggest the Do More/Do Less ratings help primarily by clarifying the interpretation of lower to middling ratings on the effectiveness evaluation scale: for instance, an effectiveness rating of 3 coupled with "Do the same" would seem to be a better result than an effectiveness rating of 3 paired with "Do more." Supporting this conclusion, running five separate curvilinear regression analyses—one for each level on the effectiveness rating scale—using only the Do More/Do Less ratings (and their cross-products) to predict calibration yields significant effects, with more variance accounted for when the effectiveness rating of 3 is used.

The advantage that these statistical results represent is about how to deal with the fact that most ratings are clumped together on traditional rating scales. In the example in Table 2, you can see that about 74% of ratings are either a 3 or 4 on the evaluation scale. The Do More/Do Less ratings effectively add shades of gray to aid interpretation. A 4 with a "do less" is less optimal than a 4 with "do the same." This can be very helpful, for instance, when comparing two individuals who have roughly equivalent scores on the evaluation scale (as many do): the one with the fewest ratings of Do More or Do Less is likely the more effective performer. Furthermore, this can help an individual struggling to decide which lower scoring items to take action on: those with more do more or do less ratings seem to warrant a higher priority.

Benefits to researchers. Finally, we have discovered at least two benefits the new response scale lends to students of management. First, it helps in detecting curvilinear relationships

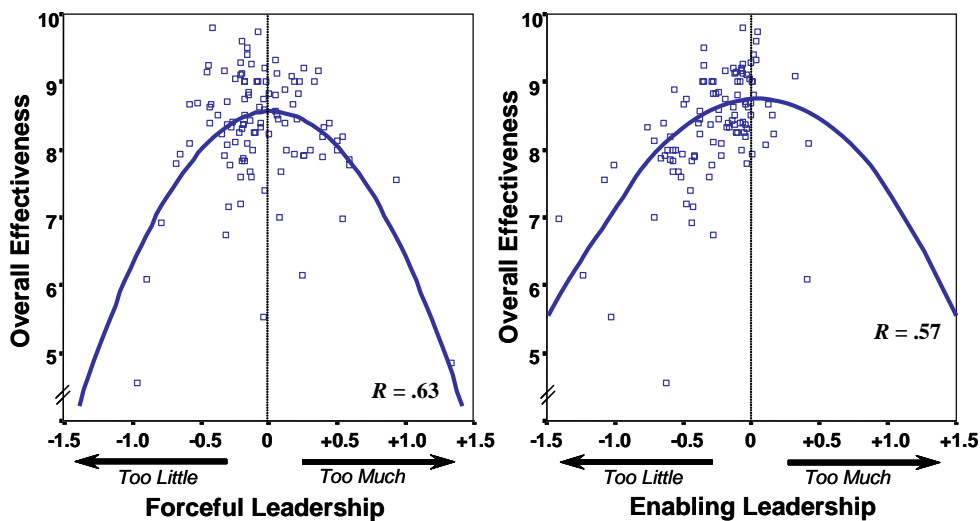
⁵ We are unable to report exact validity coefficients due to their proprietary nature. However, we can say that the magnitude of the multiple *R*s compare favorably to those typically reported in the leadership literature.

between managerial behavior and various criteria. Second, relationships between supposed opposites (e.g., task-oriented versus people-oriented) are consistent with theoretical expectations when measured with an evaluation of frequency of scale—and controversial when measured with a more traditional response scale.

Central to the design of the new response format was the idea that there is a curvilinear relationship between effectiveness and the frequency and intensity of certain behaviors, where the golden mean lies between deficiency and excess. It is interesting to note in this light how Fleishman (1998), reflecting on his research on consideration and initiating structure (e.g., Fleishman & Harris, 1962), regarded the impact of his work. He lamented that one of his least influential findings was the curvilinear relationship between the two sides of leader behavior and such criteria as turnover, grievances, and unit productivity. His original research showed that there were increases up to a point, then a leveling off, followed by a nosedive in the relationships between initiation of structure and unit performance and between consideration and various indices of unit morale. Yet, curiously, these very clear and provocative demonstrations of overdoing neither garnered much attention nor became part of the cumulative knowledge base in the leadership literature.

It will no doubt be of little surprise to hear that we routinely detect curvilinear relationships between outcome criteria and leadership dimensions measured with our evaluation of frequency scale. For instance, Figure 5 shows the relationship between Forceful leadership (an assertive, task-oriented style measured with 16 items) and Enabling leadership (an empowering, people-oriented style also measured with 16 items) and ratings of overall effectiveness. In this case, the data reflect the average rating across all coworkers (typically about 8 raters) for 107 executives and middle managers (see Kaiser & Kaplan, 2002). The criterion measure is a single-item rating in response to the question, "Please rate this individual's overall effectiveness as a manager on a 10-point scale where 5 is adequate and 10 is outstanding."

Figure 5. Relationships between forceful and enabling leadership and overall effectiveness.



These figures show curvilinear regression lines between Forceful and overall effectiveness of $R = .63$; for Enabling the strength of the relationship is $R = .57$. These relationships indicate that both approaches are indeed central to how executives size each other up overall, and about equally so. More importantly, they show how overdoing and underdoing are *both* related to

lower ratings of overall effectiveness. An advantage of working with ratings made on the implicitly curvilinear response scale is that it reminds researchers to seek out these kinds of non-linear relationships.

A second benefit to researchers from the new response scale is how supposed opposites are related to one another. In recent years, a movement has begun to identify the kinds of paradoxes that confront modern managers and, by extension, to define managerial flexibility or versatility in terms of being able to cope with these oppositions equally well (Lindberg & Kaiser, 2004). For instance, Quinn's (1988) Competing Values Framework identifies an inherent opposition between pushing for productivity while *also* building cohesion, and between establishing a stable environment *and* introducing change. Similarly, Sloan (1994) listed several balances to be struck, like competition and collaboration, vision and pragmatism, and change and continuity.

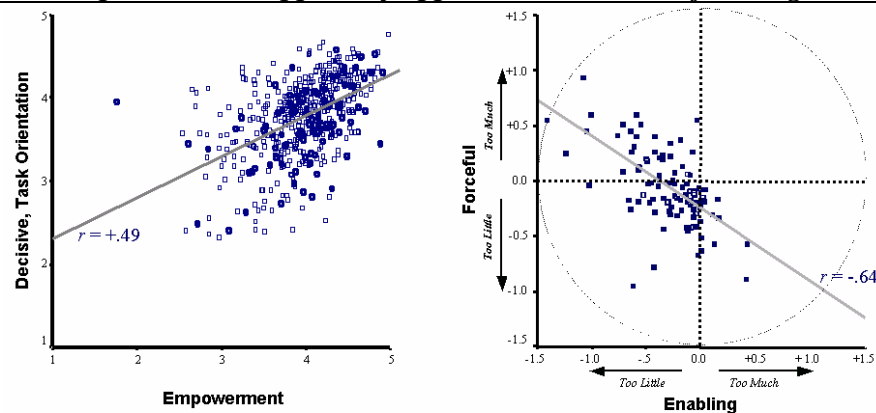
These theories all speak to a tension between opposites—hence, the frustrating paradoxes that make the manager's job a balancing act. By logical extension, one would expect there to be a negative correlation between such opposites as short-term orientation versus long-term orientation, competition versus collaboration, autocratic versus participative, and so forth. That is, we'd expect that doing one side in each pair of opposites would correspond to doing less of the other side or that being more skilled at one would covary with being less developed at the other (Kaplan & Kaiser, 2003b). However, the literature is clear on this point: when measured with a traditional response scale, correlations between these putative opposites are actually positive, and often it is a sizable positive correlation on the magnitude of .50 or so. For instance, Judge's recent meta-analysis of the twin pillars of the leader behavior paradigm, consideration and initiating structure, found that the true correlation between these two constructs on the most valid measure of them, the LBDQ XII, was $\rho = .46$ (Judge, Piccolo, & Illies, 2004). Even measures based on an explicitly oppositional theory such as Quinn's show sizable positive correlations like this when measured with a traditional response scale (see review in Lindberg & Kaiser, 2004).

However, when oppositions like this are measured using the "evaluation of frequency" (i.e., too little/too much scale), a very different pattern of relationships emerge. Figure 6 contrasts the relationship between our forceful and enabling measures and the relationship between similar constructs, Decisive, task-oriented and Empowerment. The Forceful and Enabling scales were rated using the too little/too much response scale; the decisive and empowerment scales were measured using a frequency scale (to what extent?).⁶ Both samples are based on average subordinate ratings of senior executives.

How to account for the wildly discrepant results? We think the difference is due to the type of response scale. For instance, if you cover up the two portions of the plot representing too much in the figure to the right (> 0 on forceful and > 0 on enabling), you effectively have a scale ranging from "too little" to "the right amount" and remove the regions representing "too much." This is similar to a traditional frequency scale. The result: a scatter plot that suggests a positive relationship on the order of $+.50$. This raises a question: by overlooking overkill, has the last 100 years of research on managerial behavior been looking at only half of the story?

⁶ The decisive and empowerment scales are from an extensively researched and validated commercial instrument. We use them in this example because we have the data at our disposal and they illustrate the larger point. Note that the relationship between the two is about what was reported in the Judge et al. (2004) meta-analysis of initiating structure and consideration.

Figure 6. Relationships between supposedly opposite dimensions of managerial behavior.



Relationship between supposedly "opposite" dimensions of leadership behavior when measured with a traditional *frequency* scale (left) and the new "curvilinear" scale (right). Both figures represent average subordinate ratings for a sample of executives ($N = 493$ [left] and $N = 107$ [right]).

There is also a more basic question concerning construct validity, in the sense of Cronbach and Meehl's (1955) seminal definition of the concept as concerning how accurately we can draw inferences about how various constructs are related in a nomological system. If the paradox-based theories of managerial versatility are correct, then how should the supposed opposites relate to one another? We have argued that the relationship should be negative, representing how individuals tend to be better at one side than the other and only the most effective ones, a small minority, would tend to be well-developed on both (Kaplan & Kaiser, 2003a). If we are correct, then it raises a certain shadow of suspicion on measures of opposites that evidence positive relationships.

Of course, construct validity isn't just a researcher's concern: it is essential in practice, where data is used to make decisions that affect the careers and livelihood of real people, the fate of real organizations, and ultimately have a ripple mark cascading to society at large.

Concerns and Further Development

The previous section considered what we believe to be several advantages of the new response scale format. At the same time, let us be clear that we are not without our doubts or concerns about the limitations of this new method. Below we list the major concerns we have and some of the more severe criticisms we've received from colleagues.

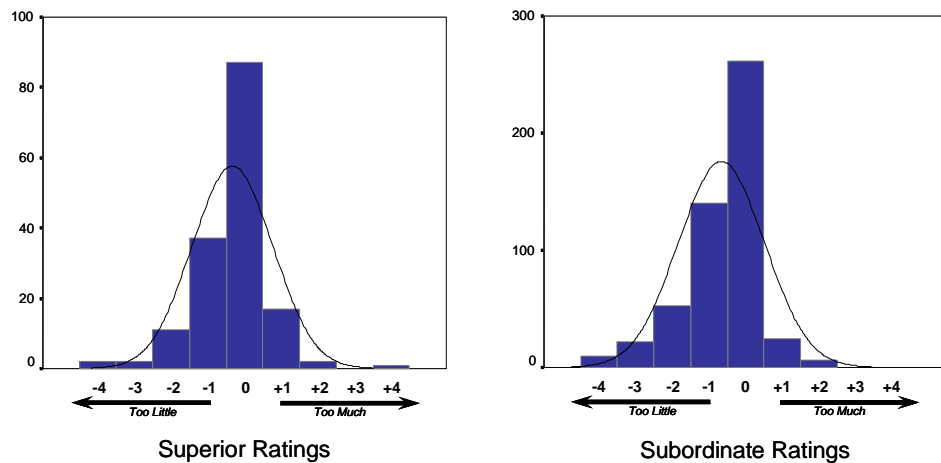
Some things can't be overdone. This is something we hear frequently, particularly from scholarly researchers. For instance, it's been claimed that you can't be too smart. One anonymous reviewer has taken us to task by stating "it doesn't appear to make much sense to say that a manager is ineffective because he needs to be less sensitive to the needs of others." And in an age where strategic leadership is all the rage, some have argued that today's leaders can't be too visionary. We disagree with each of these particular claims as well as the larger one that some things simply can't be taken too far.

Barnard, as far back as 1938, observed that an executive needs to be very intelligent, but not so smart as to intimidate others. And the U.S. District Court of Connecticut found the New London Police Department to be justified in its decision to deny employment to Robert Jordan

on the grounds that he scored *too high* on the Wonderlic, a test of cognitive ability (Jordan v. City of New London, 1999). The Police Department effectively argued that individuals with relatively high IQ tend to get bored by the lack of intellectual challenge and turnover more rapidly than their less gifted counterparts. So much for no such thing as too smart.

In terms of the idea that a manager can't be too considerate of other people, consider the frequency distributions in Figure 7. On the left is the distribution of 159 superiors' ratings on the item, "Compassionate. Responsive to people's needs and feelings." And on the right is the distribution for 518 subordinates' ratings. These ratings were made for 107 executives.

Figure 7. Ratings on the item, "Compassionate. Responsive to people's needs and feelings."



Clearly, it is possible to overdo the virtue of sensitivity (albeit also true than underdoing is more rampant than overdoing). And it helps to consider the source: superiors see this as an overkill issue more readily than do the presumed beneficiaries of compassionate behavior.

A similar empirical defense can be presented regarding strategic and visionary leadership. Although perhaps a visible public example is more poignant. The Chairman of General Motors during the 1980s, Roger Smith, was regarded as an anomaly at traditionally conservative GM for his interest in and formalization of strategic planning processes. Mr. Smith had a vision of transforming GM to regain the automobile market from the Japanese by leapfrogging state-of-the-art technologies and creating "fourth wave" and "lights out" factories—locations that were entirely computerized and automated, requiring no human workforce. His persistence on this bold, but ultimately unrealistic and ungrounded, futuristic vision is widely regarded as the key reason why GM continued to lose as much as 40% market share and depleted nearly all of its cash reserves by the early 1990s (see Hunt & Ropo, 1995).

Each of these examples demonstrates how certain skills, abilities, and behaviors, no matter how valued, can indeed be taken to a counterproductive extreme. A key lesson that we have learned in using the new response format is that you must phrase items in a way that helps the respondent easily see what "too much" of that behavior might look like. Related to this, it won't work to write items that are value-laden—for instance, "Effectively makes her point to a resistance audience" won't work because you can't be *too effective*. But "Persists in trying to persuade people despite resistance" does admit of overdoing.

An extension of the basic idea with the new response scale may make it easier for respondents to understand how to make full use of the "too much" side of the scale. It would involve developing BARS (Behavioral Anchored Rating Scale) type descriptions at the "too little" and "too much" extremes as well as for what the ideal looks like. Of course this would complicate the item generation process—each item would require three carefully crafted statements based on critical incidents. And the fact that BARS-type rating scales do not appear to confer significant advantages over traditional response scales (Landy & Farr, 1980) raises the possibility that all may be for naught. It remains an empirical question whether BARS-type anchors would improve the psychometric properties of ratings made on the new response format.

Difficulty with creating scale scores. Another limitation involves the problem of creating scale scores by computing the average rating across several items rated on the -4 to +4 scale. The problem occurs when some items are in the negative, "too little" region, while others are in the positive, too much region. The net effect is for the scores to cancel each other out and dilute the mean, bringing it closer to 0, optimal, than ought to be the case. That is, this kind of scoring procedure assumes a compensatory model, where too much of this behavior makes up for "too much" of that behavior and so on. We don't agree with this assumption.

We have yet to discover a satisfying solution to this arithmetic problem. Even with scales with high internal consistency (e.g., $\alpha > .90$), anywhere from 3% to 30% of ratings for a given target may be on opposite sides of optimal. One cumbersome way around miscommunication is to present the scale score as a mean and also present the percentage of ratings in the "too little" and "too much" region. Admittedly, this is an inelegant solution.

Another way to look at this issue is to realize that no psychological measure is perfect and to regard the dilution that occurs from positive and negative ratings canceling as a necessary cost for the gain in detecting excess separately from deficiency. To this point, the criterion-related validity coefficients we have found in our program of research are of sufficiently high magnitudes as to argue against the idea that the measurement error caused by the dilution renders the methodology invalid. Specifically, we find correlations with criteria as high as .81 (Kaiser & Kaplan, 2002) and generally in the .4 to .6 range (Kaplan & Kaiser, 2003a; 2003b). These validity coefficients suggest that the proportion of error variance due to dilution is relatively small, perhaps even trivial.

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, where the focus on using the data is more idiographic. But in other applications, particularly administrative uses of ratings where the data is used in a more nomothetic way, this can be a drawback. For instance, in some organizations, overall performance is measured by computing the average rating across a number of performance dimensions. When traditional rating scales are used, the dimensions all tend to be positively correlated and this partly justifies such practice. But with the new response scale, opposing dimensions tend to be negatively correlated. Thus, coming up with an overall score by computing a straight average runs into another form of the "dilution" problem discussed above. Moreover, it is somewhat complicated to rank-order individuals on the "curvilinear" scale, with ideal scores in the middle, and use this ordering as a basis for comparing individuals against one another.

Related to this set of concerns is the question of whether or not it even makes sense to compare ratings for two different people on the new, evaluation of frequency scale. The

argument goes that 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 in the first place. In other words, normative comparisons don't make sense.

Or do they? We're not so sure that this argument is sound. And we're not completely sure that it isn't either. It seems to assume that raters use traditional scales consistently, with the same frame of reference. This implies that raters are objective describers or judges of performance, applying the exact same standards across items and ratees. This may not be tenable. For instance, recent research on 360° ratings suggests that as much as 55% of the variance in their latent structure can be attributed to idiosyncratic *rater* effects—that is, something unique to each individual rater (Scullen, Mount, & Goff, 2000). This suggests that raters, even multiple raters of the same manager, bring their own somewhat unique "context" to the rating task anyway.

Also, one could turn this argument on its head and claim varied contexts (e.g., job, organization, organizational level, industry, culture) introduce error variance in normative comparisons and nomothetic analyses involving ostensibly "objective" traditional response scales. This would suggest that the new response scale provides a kind of quasi-experimental control for these variations.

We are relatively confident that normative comparisons of ratings on the new response format do make some degree of sense. Our bullishness comes from a simple empirical fact combined with a bit of methodological rationalism: our nomothetic, cross-sectional research analyses consistently turn up sizable positive relationships between behaviors measured on the new scale and various criteria. If our scale was flawed in that the different contexts render cross-individual comparisons meaningless, then we should find virtually no reliable relationships with criteria in samples with a variety of subjects from a variety of organizations. The fact that we do regularly find reliable correlations suggests that these between-subjects comparisons are at least somewhat valid.

We leave it to future work to further elucidate the boundary conditions of applicability of the new scale. And to identify its limitations and the severity and implications of those limits.

No direct, falsifiable comparisons between response scales. The astute methodologist will note that we have made several direct conceptual comparisons between our 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 (or twice, once each for the magnitude and evaluation scales). This could provide for adequate control to rule out competing explanations for the observed results and isolate the effects of type of response scale.

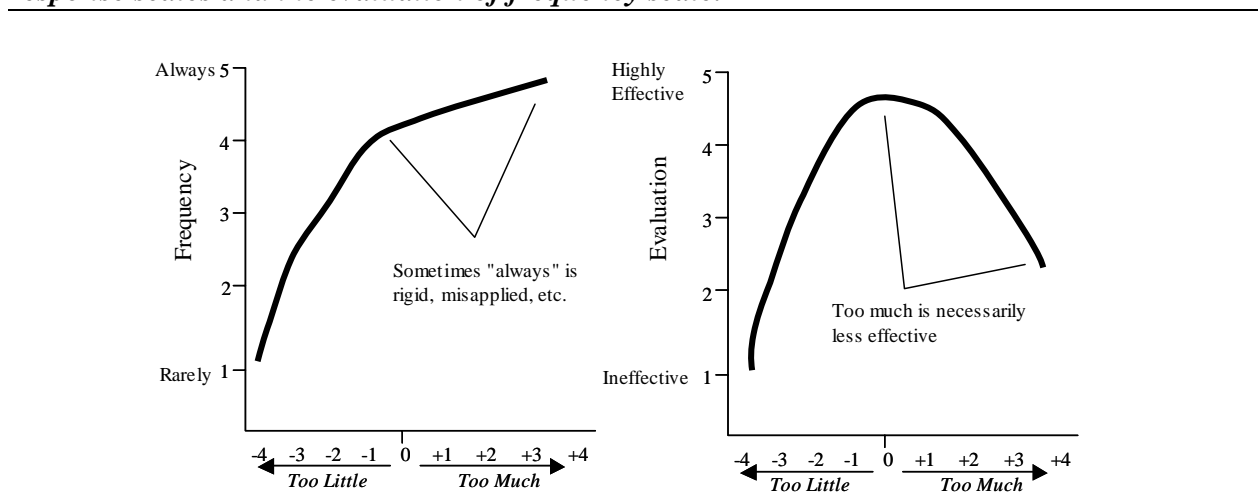
In fact, that is a study we are preparing to conduct. To keep it simple, we are using four dimensions for the behavioral item content: the LBDQ XII Consideration and Initiating Structure scales and our own Forceful and Enabling scales. Key hypotheses we will test include:

1. Scaling the same constructs measured with a frequency scale on the x -axis against the curvilinear scale on the y -axis will result in a positive curvilinear quadratic relationship

that reaches asymptote near "always"—where sometimes always corresponds to "the right amount" and other times it corresponds to "too much." (See Figure 8, left panel.)

2. Scaling the same constructs measured with an evaluation scale on the x -axis against the curvilinear scale on the y -axis will result in an inverted "U" shape, such that evaluation ratings of "ineffective" will correspond to both "too little" and "too much." (See Figure 8, right panel.)
3. The relationship between consideration and structure (and forceful and enabling) will be positive when measured with a traditional scale and negative when measured with an evaluation of frequency scale.
4. Because the new scale removes the confound in traditional scales between deficiency and excess, ratings on it will account for more variance in a range of criteria such as job satisfaction, psychological empowerment, team efficacy, and unit performance. Additionally, the new scale will provide incremental validity over traditional scales, but the reverse will not be true.
5. Inter-rater reliability will be higher for ratings made on the new scale compared to traditional scales.

Figure 8. Hypothesized relationships between ratings of the same construct on traditional response scales and the evaluation of frequency scale.



Obviously we are optimistic about our innovation in response scale technology. At the same time, we wish to be cautiously optimistic. There is still much to learn about how best to apply the new scale in practice and research. We encourage other independent research teams to conduct their own studies of the strengths and limits of this new format. To that end, we'll gladly share whatever materials and thoughts interested parties may need to get started.

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. Cambridge, MA: Harvard University Press, 1982.
- Barnard, C.I. (1938). *The functions of the executive*. Cambridge, MA: Harvard University Press.
- Bass, B.M., Cascio, W.F., & O'Connor, E. (1974). Magnitude of estimations of frequency and amount. *Journal of Applied Psychology*, 59, 313-320.
- Cohen, J. & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd edition). Hillsdale, New Jersey: Lawrence Erlbaum.
- Conway, J.M. & Huffcut, A.I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of supervisor, peer, subordinate, and self-ratings. *Human Performance*, 19, 331-360.
- Cronbach, L.J. & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.
- Eichinger, R.W. & Lombardo, M.M. (2000). *For your improvement*. Minneapolis, MN: Lominger Limited, Inc.
- Fleishman, E.A. (1998). *Patterns of leadership behavior related to employee grievances and turnover: Some post hoc reflections*, 51, 825-834.
- Fleishman, E.A., & Harris, E.F. (1962). Patterns of leadership behavior related to employee grievances and turnover. *Personnel Psychology*, 15, 43-56.
- Hunt, J. G. & Ropo, A. (1995). Multi-level leadership: Grounded theory and mainstream theory applied to the case of General Motors. *Leadership Quarterly*, 6, 379-412.
- Jordan v. City of New London, 3:97CV1012 (1999).
- Judge, T.A., Piccolo, R.F. & Ilies, R. (2004). The forgotten ones? The validity of consideration and initiating structure in leadership research. *Journal of Applied Psychology*, 89, 36-51.
- Kaiser, R.B. & Craig, S.B. (2001). *Leadership Versatility Index technical report: Item selection and validation*. Greensboro, NC: Kaplan DeVries Inc.
- 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. & Kaplan, R.E. (2002). *Leadership Versatility Index[®]: User's Guide*. Greensboro, NC: Kaplan DeVries Inc.

- 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.
- Kaplan, R. E., & Kaiser, R. B. (2002). *Leadership Versatility Index[®]*. Greensboro, NC: Kaplan DeVries Inc.
- Landy, F.J. & Farr, J.L (1980) Performance rating. *Psychological Bulletin*, 87, 72-107.
- 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.
- Lindberg, J.T. & Kaiser, R.B. (2004, April). *Assessing the behavioral flexibility of managers: A comparison of methods*. Poster session presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL.
- Lombardo, M.M. & Eichinger, R.W. (2000). *The leadership machine*. Minneapolis, MN: Lominger Limited, Inc.
- Lombardo, M.M. & McCauley, C. (1988). *The dynamics of management derailment*. Greensboro, NC: Center for Creative Leadership.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 5-53.
- McCall, W.M. & 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.
- Quinn, R.E. (1988). *Beyond rational management*. San Francisco: Jossey-Bass.

- Schwartz, N. (1999). Self reports: How the questions shape the answers. *American Psychologist*, 54, 93-105.
- Scullen, S.E., Mount, M.K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, 85, 956-970.
- Shipper, F. (1991). Mastery and frequency of managerial behaviors relative to sub-unit effectiveness. *Human Relations*, 44, 371-388.
- Shrout, P. & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420-428.
- Sloan, E. B. (1994). Assessing and developing versatility: Executive survival skill for the brave new world. *Consulting Psychology Journal: Practice and Research*, 46, 24-31.
- Spector, P. E. (1976). Choosing response categories for summated rating scales. *Journal of Applied Psychology*, 61, 374-375.
- Stockford, L. & Bissell, H.W. (1949). Factors involved in establishing a merit-rating scale. *Personnel*, 26, 94-116.