Practitioner forum presented at the 16th annual conference of the Society for Industrial-Organizational Psychology, San Diego, CA. April 28, 2001. In M. Kelly and M. Russell (Co-chairs), *Use of Assessment Tools in Leadership Development*.

The Forceful and Enabling Polarity: A Fresh Look at an Old Distinction

Robert E. Kaplan & Robert B. Kaiser
Kaplan DeVries Inc.

A two-factor model of leadership has been around for a long time, and has taken several related forms. Pioneers of this view at Ohio State and Michigan State called these two fundamental dimensions *initiating structure* and *consideration* (Fleishman, 1973; Hemphill, Seigel, & Westie, 1951). Subsequent scholars have looked upon these factors from slightly different perspectives using slightly different labels. We too have given the theory a twist. We have also approached the measurement of it a little differently. Method followed model.

We have conceived of these two sides of leadership as a polarity, two opposing desirable things to do. It is in the nature of any polarity that a person can strike a balance between the two sides or be overbalanced on one side or the other. To be overbalanced, to be lopsided, is to overdo one side and underdo the other. Following the nature of the phenomenon, we[1] adopted a measurement approach that could account for the imbalance, where overdo and underdo could readily be represented. This is a departure from the typical approach.

A New Look at a Familiar Duality

Our work takes its place in a long tradition, not only in relation to leadership but also in relation to human nature. Over and over, roughly the same two sides of human nature and leadership have appeared in the literature. Bakan (1966), for instance, proposed agency and communion as the two basic dimensions of human existence. Agency refers to urges for self-definition and individuation and it involves self-assertion, self-expansion, and self-protection. Communion, on the other hand, refers to the integration of the individual in a larger social entity and includes cooperation, caring, and forming connections with others. More recently, building on interpersonal theory (Leary, 1957; Sullivan, 1953; Wiggins, 1991), Hogan (1983, 1996) has articulated the evolutionary basis of these two basic dimensions with his socio-analytic theory. His formulation points to two overarching drives in human motivation, the needs to get ahead and to get along—in effect, agency and communion. Moreover, the agentic task of getting ahead and the communal task of getting along also apply to the functioning of work teams—to remain viable, a group must achieve its goals and outperform the competition as well as maintain solidarity and a rewarding sense of belongingness among its members.

Given the ubiquity of agency and communion in human affairs, it is no surprise that they have appeared in various forms in the study of leadership. For example, these meta-themes can be identified as the underpinnings of what Bass (1990, p. 416-419), in his exhaustive review of the scholarly work on leadership, defined as the two overarching clusters of active leadership behaviors. One cluster centers on the "autocratic" use of power and a focus on the work to be done, while the other cluster revolves around a "democratic" use of power and concern for people. According to Bass, within each of these distinct clusters are overlapping and empirically related leadership behaviors characterized by the many dualities discussed over the years.

Bass' (1990) autocratic and democratic clusters, respectively, include such familiar dualities as initiation of structure versus consideration (Fleishman, 1973; Hemphill et al., 1951), production-orientation versus people-orientation (Blake & Mouton, 1964, Likert, 1961), concern for task accomplishment versus relationships (Fiedler, 1967), performance versus group maintenance (Misumi, 1985), directive and autonomous decision-making versus consultative and participative decision-making (Bass & Valenzi, 1974; Vroom & Yetton, 1974; Yukl, 1971), power-oriented charisma versus power-sharing consensus (Zaleznik, 1974), and the controlling Theory X versus the trusting Theory Y ideologies (McGregor, 1960).

Our Version of the Two Factors

Our contribution is to construe these two factors as a polarity. We are certainly not the first ones to use the notion of a polarity (c.f., Jung, 1976; Levinson, 1978), nor are we the only ones to call attention to the dynamic

---

Robert E. Kaplan, Co-president, Kaplan DeVries Inc. Robert B. Kaiser, Director of R&E, Kaplan DeVries, Inc. Inquiry about this paper and the measure described in it may be directed to rkaiser@kaplandevries.com.

We gratefully acknowledge thoughtful input for this paper from S. Bart Craig, David DeVries, and Eileen Frances. S. Bart Craig also made significant contributions to the data analyses reported in the appendix.

[1] Bob Kaplan is credited for inventing the instrument and the unique response scale to be described. Rob Kaiser subsequently assisted in psychometric refinement and conceptual revision of the original instrument.

tension between these two factors in leadership (e.g., Quinn, 1988; Quinn, Spreitzer, & Hart, 1991). Our intention is to elucidate the idea of the *forceful and enabling polarity* in a way that is helpful in leadership development efforts as well as in taking a new approach to the measurement of leadership and versatility.

A word on what a polarity is not: it is not a continuum with one element on one end and the other on the opposing end. A polarity consists of two separate entities that can be complements but are routinely treated as mutually exclusive.

By forceful we mean leading off of one's intellect and energy—taking stands, being decisive, making tough calls, holding people accountable, and so on. By enabling we mean creating conditions for other people to make a contribution—granting them autonomy, being receptive to their influence, providing support, helping them feel valued, and so forth.

Construing the two factors as a polarity has certain advantages. One advantage is that it places equal value on both things on the assumption that both have their uses. Together they make a whole. For this reason, in deciding what to call the two factors, Kaplan (1996) carefully chose names with a positive connotation. This has not always been true of the labels applied to previous formulations of the two factors. "Autocratic" and "coercive" are decidedly less desirable terms than **"**democratic" and "participative." And although the terms Theory X and Theory Y are ostensibly neutral, McGregor's (1960) treatment had a decided bias in favor of Theory Y and an equally decided bias against Theory X.

A second advantage of a polarity is that it pairs two good things that are seeming opposites. It puts them in tension with each other and in so doing brings out the dynamic relationship between them. Objectively, forceful leadership and enabling leadership are complements of each other. But many managers don't see it that way. They see one side as being preferable to the other. In extreme cases they see the two as polar opposites. Thinking of leadership in terms of polarity is consistent with the tendency of managers to be one-sided. Fundamentally, the idea of a polarity taps into the universal tendency of human beings to polarize, regularly seen between individuals, between groups, even between nations.

It is this tendency for managers to be lopsided that led Kaplan to the idea of a polarity. It was in the course of consulting to senior managers in a way that doubled as action-research that this phenomenon stood out (Kaplan, Kofodimos, & Drath, 1987). He was led to it empirically. It wasn't just lopsidedness that stood out. It was the connection to effectiveness. When the data on a manager indicated that he or she was out of balance, more often than not that was a manager who received a relatively low rating on overall effectiveness. And by the same token, managers who coworkers described as "balanced" or "complete" tended to get relatively high effectiveness ratings. This effect seemed to make a big difference.

A third advantage is that a polarity-based conception of leadership style suggests a definition of effectiveness as versatility. The root of the word versatility is *to turn*, literally the ability to turn from one side to the other, as the situation requires. It means having full range of motion, as opposed to the restricted range characteristic of one-sided managers. Versatility can also mean combining the opposing qualities in the same act. One executive we work with, for example, was described as "pleasantly demanding." Another executive had the ability to ask his people difficult questions in a non-threatening way that did not disable them.

A fourth advantage is that a polarity calls attention to the beliefs that underlie an individual's leadership style. The leadership an individual exercises is not only a function of the skills he possesses but also what he believes to be the right way to lead. Individuals tend to identify more with one of these two sides of leadership than the other. To one degree or another they have an aversion to the other side. The root of the word aversion is the same as for versatility, *to turn*—in this case, to turn away. The side they avert their eyes from becomes their blind side. Talk to a lopsided manager and it won't be long before it emerges that the individual has an "attitude" about the neglected side. An overly forceful leader might come out with a statement like, "I have no use for go-along, get-along types." Or "I'm not a touchy-feely kind of person." Beliefs, or should we say prejudices, like this underpin distortions in a manager's leadership. Managers, like all of us, tend to rationalize what they do, and those self-justifications reinforce their current behavior and stand in the way of stated goals for self-improvement. This is one application of the idea of a polarity: to grow and improve, the manager's job is not just to acquire skills but also to call into question what in truth are distorted ideas about leading.

Thinking of leadership in terms of a polarity throws light on what inside a manager's skin throws off his or her leadership. In addition to distorted beliefs, fears play a part. Fears come into play when managers contemplate changing their approach. Overly forceful managers often worry that if they move in the opposite direction, they will be weak. Overly enabling managers tend to worry that if they move in the other direction, they will become obnoxious and overpowering. In both cases, managers scare themselves with a stereotyped idea of the other side. A polarity-based view can serve as a gate through which managers can discover, in a way they find legitimate, what inside them is in play.

A Different Approach to Measurement

Viewing the two factors as a polarity throws into relief the fact that managers polarize on the two. They are often overbalanced, typically toward the forceful side. This simple fact, plain as day in organizational life, leads inescapably to the conclusion that a measure of the two

factors ought to provide the possibility that managers overdo it.

In conducting action-research on executives, Kaplan (1996) stumbled into this phenomenon when he groped for a way to reflect back to executives what he saw in their leadership. 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 "overly forceful."

The notion that managers go overboard is hardly original with us. This is a widely held idea in organizations. People are fond of saying, for example, a strength taken to an extreme is a weakness. The literature on polarities in human nature also contains this idea. Bakan (1966) suggested that much of the discontent in Western civilization could be traced to an obsessive focus on the individual, where exaggerated agency comes at the expense of communion. Sidney Blatt (Blatt & Blass, 1996; Guisinger & Blatt, 1994), who conceptualized personality development in terms of a dialectic tension between self-definition and relatedness, has identified two broad types of pathology: when people emphasize agency at the expense of communion and when people emphasize communion at the expense of agency (see also, Helgeson, 1994).

Yet the idea that measures of leadership ought to include a provision for "overdoing" has somehow not taken hold. This is despite the fact that many performance problems are cases of managers taking something to an extreme or giving it short shrift.

Design Specs for a Measure of the Forceful and Enabling Polarity

Response scale. The majority of leadership measures involve ratings of questionnaire items: respondents are presented with a list of statements describing behavior and are asked to choose the option on a response scale that best characterizes the manager in question. Response scales typically take one of two forms (Leslie & Fleenor, 1998): evaluation or frequency/magnitude. Evaluation scales ask the respondent to judge *how well* the target manager performs a given task or behavior. Frequency/magnitude scales ask the respondent to evaluate *how often* the manager engages in the behavior, how much the respondent agrees that the statement describes the manager, and so forth. Examples of each scale are presented in Figure 1.

Evaluation scales do a perfectly good job of capturing behaviors that are "underdone," which are represented with a low score. But these scales appear ambiguous, perhaps even misleading, about behaviors taken to the overdo extreme. At worst, low scores confound the overdo/underdo distinction. For instance, does a low rating of "*Underdeveloped*" on the item, "Takes preventative measures to avoid crisis management," (from the Executive Success Profile, Hezlett, Ronnkvist, Holt, & Sloan, 1996) indicate careless recklessness (underdo) or ultra-conservative vigilance (overdo)?

The more common frequency/magnitude scales are purely descriptive of the extent of behavior, where higher ratings simply reflect more of a given behavior. It is often assumed that higher ratings on these scales indicate proficiency or mastery (Shipper, 1991). This is evident in the fascination with "high scores." The assumption is also made in the common practice of relating scores on these scales with criterion variables with the product-moment correlation and other linear modeling techniques (for an exception, see Fleishman & Harris, 1962). These procedures rest on the assumption that more of a given behavior is "better." Again, there is no provision for the overdo extreme.

Conventional response scales simply cannot readily account for the implicitly curvilinear idea that a particular behavior can be done too little, optimally, or too much relative to a given criterion. The only way they can

Figure 1. *Examples of traditional response scales*

Evaluation scale[1]

How would you rate this person's competence in the following areas?

Not developed	**Underdeveloped**	**Competent**	**Very strong**	**Outstanding**
1	2	3	4	5

Frequency/magnitude scale[2]

Judge how frequently each statement fits the person you are describing. Use the following rating scale:

Not at all	**Once in a while**	**Sometimes**	**Fairly often**	**Frequently, if not always**
0	1	2	3	4

Note: [1]Taken from the Executive Success Profile (Hezlett et al., 1996); [2]taken from the Multifactor Leadership Questionnaire (Bass & Avolio, 1997)

measure behaviors taken to an overdo extreme is to write negatively worded items, like "overly forceful." But then a mirror-image item like "not forceful enough" is required to get at the other extreme. Even then, it is not obvious where the optimal amount is.

Given the limitations of traditional rating formats, Kaplan (1996) came up with the format shown in Figure 2. This scale could be described as "evaluation of behavior frequency" because it incorporates both a frequency and an evaluative component. The scale anchors imply that the rating is relative to the manager's context, where it may for example be advantageous to be exceptionally directive or tough (c.f., Schriesheim, House, & Kerr, 1976).

Figure 2. *The implicitly "curvilinear" scale*

Please rate the manager in question on each of the following aspects of executive leadership.

Please note that the scale is probably different from scales that you are accustomed to using. On this scale the best score is "2," smack in the middle of the scale. The premise is that performance problems arise when managers either underdo or overdo something.

Too little		The right amount			Too much	
0.5	**1**	**1.5**	**2**	**2.5**	**3**	**3.5**

WARNING: Some people misread this scale. Please do not mistake it for the usual type where a high score is the best.

Item text: leadership virtues. The next consideration concerned how to cast items that would work with the new response scale. There is really only one specification that is a departure from standard practice: the items must be written so that they admit of a response of "too little" or of "too much." This requirement is consistent with the character of the constructs themselves, which are good things to do, leadership virtues. One vice is not doing it enough; the other is doing it too much—too much of a good thing. This contrasts with the item content in measures of similar constructs. For instance, the SBDQ (Fleishman, 1989a) initiating structure scale includes some pejorative items having to do with coercive, domineering behaviors (e.g., "he rules with an iron hand," "he 'needles' those under him for greater effort," Schriesheim et al., 1976), items that, from our point of view, operationalize the overdo aspect of forceful leadership.

Juxtaposing complementary virtues. Another design requirement we set was to write items in pairs. Each pair was meant to be a sub-polarity of forceful versus enabling leadership. A classic example is talking versus listening. Effective communicators must do both, and they are less effective when they do more of one and less of the other, in either direction. Each item pair was inspired by one of Kaplan's executive clients, who either was a model or a negative role model. A sample item pair, which is an executive version of talking versus listening, appears in Table 1.

Table 1. *Sample item pair*

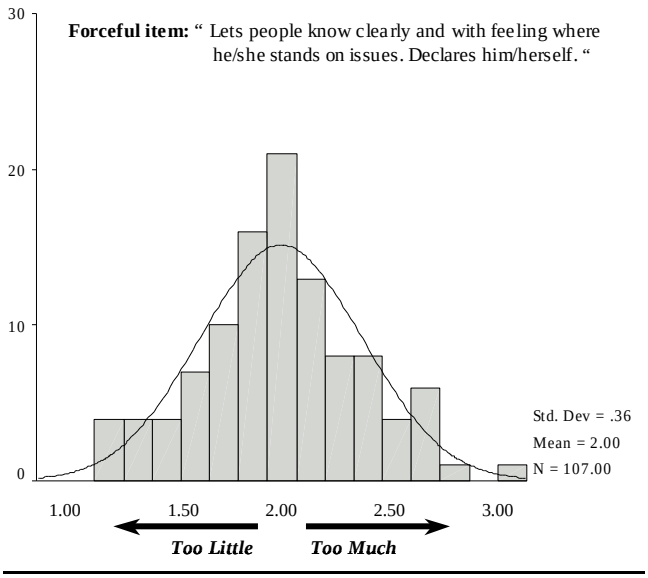
1f. Lets people know clearly and with feeling where he/she stands on issues. Declares him/herself.
1e. Interested in where other people stand on issues. Receptive to their ideas.

The items are presented separately on the survey questionnaire. But on the feedback report, results for item pairs are presented side by side. The reason is that it is more powerful for managers to see these two results juxtaposed than singly. It can be eye-opening, for example, for an overly forceful executive to see how her colleagues rated her in the overdo range for "declares herself" and in the underdo range for "receptive to their ideas."

Does the response scale work?

That some executives tend to both overdo and underdo forceful and enabling leadership is evident in the ratings we observe in the data generated using this new scale. In Figure 3 is the distribution of responses to the forceful item shown in Table 1. The data come from ratings by over 500 subordinates of 107 different executives. It is evident that coworkers can readily distinguish these two types of performance problems—overdoing it and underdoing it—and that both types can be represented using this response format.

Figure 3. *Frequency distributions of subordinates' ratings of 107 executives on a forceful item using the "curvilinear" scale*



Evaluating a New Measure

Do the forceful and enabling scales hold up statistically?

On the basis of our ongoing validation research (Kaiser & Craig, 2001; Kaiser & Kaplan, 2000), which is summarized in detail in the appendix, the answer appears to be yes. Our validation work has culminated in two refined five-item scales—one for forceful and one for enabling—that demonstrate adequate psychometric properties. The behavioral cores of the scale items are presented in Table 2. The scales are factorially distinct according to confirmatory factor analysis; internally consistent (α's for both exceed .75 at the rater level and range from .75 to .90 for data aggregated within various coworker rating sources); demonstrate measurement equivalence across selves, superiors, peers, and subordinates; and show acceptable levels of inter-rater reliability (ranging from .62 to .91 across rating sources) and inter-rater agreement (.88 to .91 across sources). Although we are revising these scales to include more items for better content coverage, we offer the following empirical data to illustrate the utility of this approach.

Table 2. *Behavioral core of forceful and enabling items*

Forceful	*Enabling*
1f. Strong leader	1e. Enables subordinates
2f. Declares self	2e. Receptive to others' ideas
3f. Makes tough calls	3e. Compassionate
4f. Makes judgments	4e. Makes tough calls
5f. Forces issues	5e. Fosters harmony

The results reported below are based on a sample of ratings for 107 executives for whom we did assessments as part of our executive development practice. The target managers are mostly middle-aged white men holding positions ranging from vice president to CEO in a variety of U.S. firms. The sample included ratings from 104 self-raters and 1,036 coworkers—165 superiors, 362 peers, and 509 subordinates.

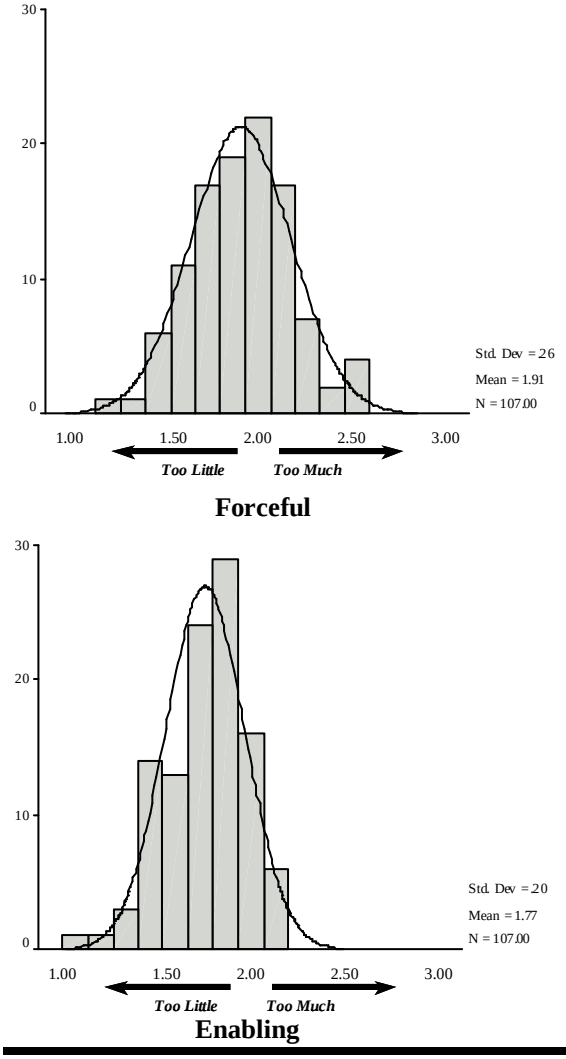
To simplify presentation of results, we created overall scores on both forceful and enabling leadership by computing the average rating on the five items in each scale across all coworkers. Separate analyses conducted within each rating source produced similar results.

First, as with the item shown above, the distribution of scores on the two scales range from the underdo side to the overdo side (see Figure 4). As one might expect, though, the incidence of overdoing enabling leadership is much less common.

Do the scales relate to each other as expected?

Although confirmatory factor analysis established that these are two distinct scales, they are correlated. And they are related in the expected direction. Consistent with the view of the two factors as a polarity, executives in our

Figure 4. *Frequency distributions for forceful and enabling scale scores*



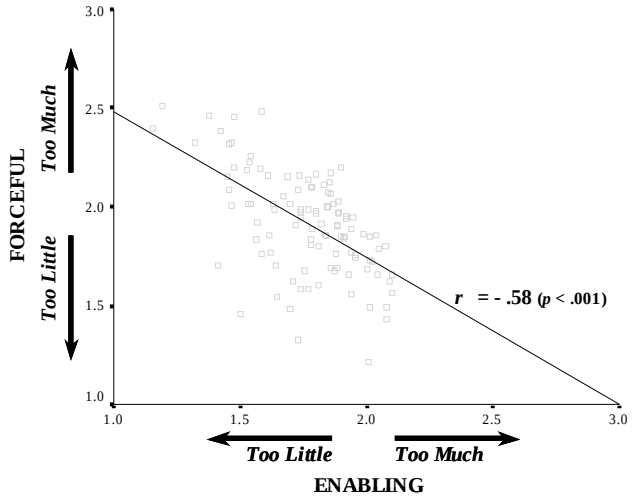
experience who overdo one side tend to underdo the other. In fact, our data reveals a fairly strong negative correlation between forceful and enabling leadership, r = - .58 (p < .001). A scatter plot showing this relationship is depicted in Figure 5. In the self-rating data, this trend was also apparent, though slightly weaker, r = - .31 (p < .01). Thus, when the two sides of leadership are measured using a rating format that allows for respondents to indicate overdoing as well as underdoing, this "polarity effect"—the inverse relationship between forceful and enabling leadership—can be detected.

Although a polarity-based view suggests a negative correlation between these two sides of leadership, that is not what previous research has found. Studies of related constructs like initiating structure and consideration either have found no correlation (Fleishman, 1989b) or, as has more often been the case, a sizeable positive correlation (Schreisheim et al., 1976)!

Bass (1990) and others have noted this troubling finding and offered possible explanations for it. The usual suspect has been halo bias—the idea that general impressions color

ratings of specific dimensions like these. This explanation is quite likely, but it does not preclude the idea that the absence of a negative relationship may be due to the fact that the scales used in previous studies did not allow for the underdo/overdo distinction.

Figure 5. *Inverse relationship between all coworkers' average ratings of 107 executives on forceful and enabling leadership*



Does versatility have the expected relationship with effectiveness?

In a polarity-oriented conception of these two factors, the flip side of the tendency to polarize is to treat them as complements—in other words, to have an adequate capacity on both sides. This is versatility. A critical step in the statistical development of the instrument was to construct a measure of versatility. Kaiser devised what we call a *versatility index*, which considers jointly the extent that a manager uses both forceful and enabling leadership. It is computationally equivalent to the *squared Euclidean distance metric* and reflects the extent to which the individual is rated at, or close to, the midpoint on both scales marked "does the right amount."

For example, in the upper-left-hand corner of Figure 6 is a plot for an executive who scored in the "does too much" region on forceful and in the "does too little" region on enabling. The geometric distance this leader is from being perfectly versatile—that is, distance from a score of the right amount (2.00) on forceful and distance from a score of the right amount (2.00) on enabling—can be derived from the Pythagorean theorem. It is calculated as: $c^2 = a^2 + b^2$ where a = (forceful score – 2), b = (enabling score – 2), and c = *distance from optimal versatility.*

To compute the versatility index, we calculated the ratio of each executive's observed *distance from optimal versatility* to the maximum possible distance from optimal versatility (i.e., scores on the extreme ends of the scale, 0.5 and 3.5). This ratio is an inverse measure of versatility. So that higher values would indicate more versatility, we subtracted this value from 100 per cent. Therefore, versatility indices range from 0 to 1.00, where lower values indicate a greater degree of lopsidedness, higher values indicate more versatility across the two sides.

Figure 6. *Computing the "distance from optimal versatility"*

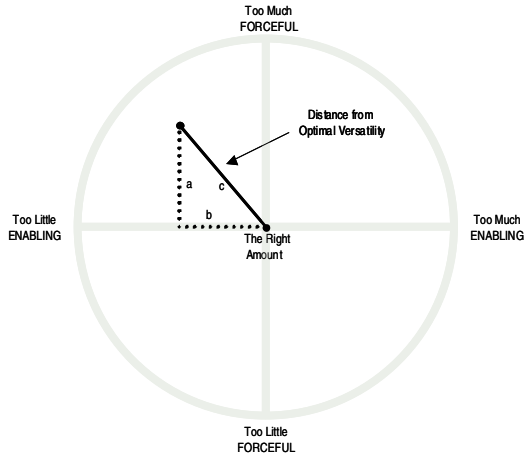
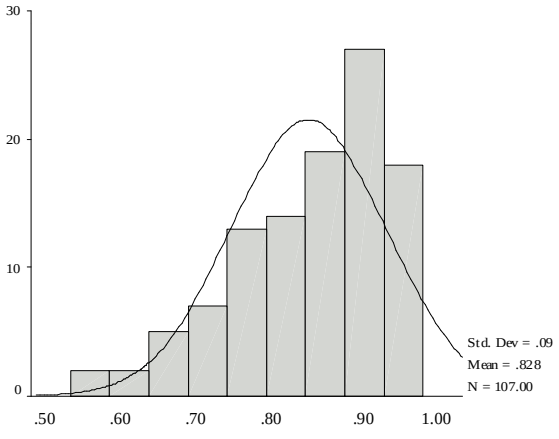


Figure 7 contains a frequency distribution of versatility indices for our sample of 107 executives. The right tail of the distribution is missing, indicating that leaders who draw optimally from both forceful and enabling approaches may be relatively rare. In this sample no executive had a score of 2.00 ("the right amount") on both scales and very few (about 10 per cent) had scores that were not statistically different from 2.00 on both scales.

Figure 7. *Frequency distribution of the Versatility Index for a sample of 107 executives*



With the versatility index in hand, we were able to correlate that number with ratings of overall effectiveness as an executive. The measure of effectiveness we used is a

single-item rating on a scale from 1 to 10, where 5 is adequate and 10 is outstanding. Estimates of inter-rater reliability, inter-rater agreement, and between-source (i.e., superior, peer, and subordinate) convergence for this single-item rating were comparable to meta-analytic estimates of the same statistics for multiple-item scales (Conway & Huffcut, 1997). It appears to be a reasonable estimate of perceived effectiveness.

For the 79 executives in our database for which we have data on both measures, we correlated the versatility index with the all coworkers' average effectiveness rating. The result showed a strong relationship, $r = .53$ ($p < .001$). To control for common source bias, we also ran the correlation between the versatility index based on subordinates' ratings and effectiveness ratings from superiors. Again, the correlation was significant, but somewhat smaller, $r = .31$, ($p < .01$).

Thus, although few executives in our sample were clearly versatile, the results indicated that the more closely a leader's style approximates a balanced mix of *both* forceful *and* enabling leadership, the more effective the leader is.

Interestingly, when we analyzed the executives' self-ratings we did not find a relationship between versatility and effectiveness. Here the correlation was a non-significant $r = .08$. This is consistent with the idea that many managers hold the misguided view that their preferred style is most effective. They believe that their approach with regard to these sides of leadership is the right one. This finding also highlights the importance of developmental feedback: these executives did not see what was so apparent to their coworkers—that their lopsidedness hurt their effectiveness.

As an aside, Quinn and his colleagues have obtained a similar relationship to effectiveness using a measure of "tough-love"—their take on integrating the two sides of leadership (Quinn, Spreitzer, Hart, 1991). That measure contains items on forceful- and enabling-like scales rated with a conventional response format. However, to construct an index of integrating the two, they needed to employ a complex statistical approach (see Bobko & Schwartz, 1984). Our curvilinear scale and versatility index seem more straightforward and easier for both scholars and feedback recipients to understand.

Conclusion

We have made a case, on conceptual grounds, for viewing two long-studied factors in leadership as a polarity. We have also introduced a 360-degree questionnaire built on this conceptual foundation and shown statistically that the tool performs as the theory predicts. We will close by explaining briefly how the assessment tool adds value in practice.

In our consulting practice we take a data-driven approach to helping senior managers to develop. The forceful and enabling measure is used alongside other 360s, personality profiles, and interviews with coworkers. The total set is large and varied; it is a complex process to sift through the data and distill essential themes. It helps managers and consultants grappling with it to have a higher-order theory to integrate it all. When a phalanx of data points can be lined up around the conceptual coordinates of orientation to self and orientation to others, it can be like finding true North on the developmental path.

Managers find it useful to take stock of themselves in terms of the forceful-enabling polarity as well as to define the direction for their development in these terms. "I need to be less forceful and more enabling," for example, is a common, simple consolidating statement of the data points peppered throughout the assessment data. There is not a single manager to whom the distinction between forceful and enabling leadership does not apply.[2]

The forceful-enabling polarity, and in fact any polarity, is useful in management development because it naturally calls attention to the beliefs and values that underlie a manager's posture on the polarity. A lopsided leadership style almost always rests on distorted beliefs about leadership (see Kaplan, 1996). It doesn't take a long conversation to bring this distorted thinking, or should we say prejudicial notions, to the surface. Managers are in a stronger position to grow and improve if they go to work on an internal piece like this and not just the behavioral piece. If a manager's leadership is off, it only makes sense that he or she must examine what throws it off (Kaplan & Kaiser, in press). The idea that management development is personal, however, is not necessarily welcomed by the organizational world or by the professional world for that matter.

Just below distorted beliefs usually lurk fears that also drive the individual's leadership posture and have to be overcome if the individual is to attain versatility on the forceful and enabling polarity. There is the fear, for example, on the part of overly forceful individuals of not being powerful enough that produces the excess. And there is the fear, on the part of overly enabling people, of becoming a caricatured version of forcefulness, as if moving in that direction means being rude, arrogant, and obnoxious. A polarity-based view of leadership can be a door into the inner world of warped beliefs and emotional sensitivities that throw off a manager's form, should the individual be willing and the consultant or boss or HR professional be equipped to go there.

[2] The forceful-enabling polarity is hardly the only important one in management. Another big one is the distinction between strategic leadership and operational leadership.

References

Bakan, D. (1966). The duality of human existence: An essay on psychology and religion. Chicago: Rand McNally.

Bass, B.M. (1990). Bass and Stogdill's handbook of leadership: Theory, research, and managerial applications (3rd Ed.). New York: Free Press.

Bass, B.M. & Avolio, B.J. (1997). Full range leadership development: Manual for the Mutlifactor Leadership Questionnaire. Palo Alto, CA: Mind Garden

Bass. B.M. & Valenzi, E.R. (1974). Contingent aspects of effective management styles. In J.G. Hunt & L.L. Larson (Eds.), Contingency approaches to leadership. Carbondale: Southern Illinois University Press.

Blake, R.R. & Mouton, J.S. (1964). The managerial grid. San Francisco, CA: Gulf.

Blatt, S. J. & Blass, R. B. (1996). Relatedness and self-definition: A dialectical model of personality development. In G.G. Noam & K.W. Fischer (Eds.), Development and vulnerabilities in close relationships. Hillsdale, NJ: Erlbaum.

Bobko, P. & Schwartz, J.P. (1984). A metric for integrating theoretically related but statistically unrelated constructs. Journal of Personality Assessment, 48, 281-320.

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.

Fiedler, F.E. (1967). A theory of leadership effectiveness. New York: McGraw-Hill.

Fleishman, E.A. (1973). Twenty years of consideration and structure. In E.A. Fleishman and J.G. Hunt (Eds.), Current developments in the study of leadership. Carbondale: Southern Illinois University Press.

Fleishman, E.A. (1989a). Examiner's manual for the Supervisor Behavior Description Questionniare (SBDQ) (revised). Chicago: Science Research Associates.

Fleishman, E.A. (1989b). Examiner's manual for the Leadership Opinion Questionniare (LOQ) (revised). Chicago: Science Research Associates.

Fleishman, E.A. & Harris, E.F. (1962). Patterns of leadership behavior related to employee grievances and turnover. Personnel Psychology, 15, 43-56.

Guisinger, S. & Blatt, S.J. (1994). Individuality and relatedness: Evolution of a fundamental dialectic. American Psychologist, 49, 104-111.

Helgeson, V. S. (1994). Relation of agency and communion to well-being: Evidence and potential explanations. Psychological Bulletin, 116, 412-428.

Hemphill, J.K., Seigel, A., & Westie, C.W. (1951). An exploratory study of relations between perceptions of leader behavior, group characteristics, and expectations concerning the behavior of ideal leaders. Columbus: Ohio State University, Personnel Research Board.

Hezlett, S.A., Ronnkvist, A.M., Holt, K.E., & Sloan, E.B. (1996). The Executive Success Profile technical summary. Minneapolis, MN: Personnel Decisions International.

Hogan, R. (1983). A socioanalytic theory of personality. In M.M. Page (Ed.), 1982 Nebraska symposium on motivation (pp. 55-89). Lincoln: University of Nebraska Press.

Hogan, R. (1996). A socioanalytic perspective on the five-factor model. In J. S. Wiggins (Ed.), The five-factor model of personality (pp.163-179). New York: Guilford.

Hooijberg, R. & Choi, J. (2000). Which leadership roles matter to whom? An examination of rater effects on perceptions of effectiveness. Leadership Quarterly, 11, 341-364.

Jung, C.G. (1976). The portable Jung. New York: Penguin Books.

Kaiser, R.B. & Craig, S.B. (2001). [Item selection and validation for forceful and enabling leadership scales: Exploratory and confirmatory analyses of data from 1,140 raters of 107 executives.] Unpublished data analyses. Kaplan DeVries Inc.

Kaiser, R.B. & Kaplan, R.E. (2000, April). Getting at leadership versatility: Theory and measurement of the forceful and enabling polarity. Paper presented at the 15th annual meeting of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Kaplan, R.E. (1996). Forceful leadership and enabling leadership: You can do both. Greensboro, NC: Center for Creative Leadership.

Kaplan, R.E. (1998). Getting at character: The simplicity on the other side of complexity. In R. Jeanneret and R. Silzer (Eds.), Individual psychological assessment: Predicting behavior in organizational settings, (pp. 178-227). San Francisco: Jossey Bass.

Kaplan, R.E. & Kaiser, R.B. (in press). How sensitivities throw off performance in executives. Greensboro, NC: Center for Creative Leadership.

Kaplan, R.E, Kofodimos, J.R., & Drath, W.H. (1987). Development at the top: A review and a prospect. In R.W. Woodman & W.A. Pasmore (Eds.) Research in Organizational Change and Development, Vol. I, (pp. 229-273). Greenwich, CT: JAI Press.

Leary, T. (1957). Interpersonal diagnosis of personality. New York: Ronald Press.

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.

Levinson, D.J. (1978). The seasons of a man's life. New York: Ballantine Books.

Likert, R. (1961). New patterns in management. New York: McGraw-Hill.

Misumi, J. (1985). The behavioral science of leadership. Ann Arbor: University of Michigan Press.

McGregor, D. (1960). The human side of enterprise. New York: McGraw-Hill.

Quinn, R.E. (1988). Beyond rational management. San Francisco: Jossey-Bass.

Quinn, R.E., Spreitzer, G.M., & Hart, S. (1991). Challenging the assumptions of bipolarity: Interpenetration and managerial effectiveness. In S. Srivatva & R. Fry (Eds.) Executive and organizational continuity (pp. 222-252). San Francisco: Jossey-Bass.

Schriesheim, C.A., House, R.J., & Kerr, S. (1976). Leader initiating structure: A reconciliation of discrepant research results and some empirical tests. Organizational Behavior and Human Performance, 15, 297-321.

Shipper, F. (1991). Mastery and frequency of managerial behaviors relative to sub-unit effectiveness. Human Relations, 44, 371-388.

Sullivan, H.S. (1953). The interpersonal theory of psychiatry. New York: Norton.

Vroom, V.H. & Yetton, P.W. (1974). Leadership and decision-making. New York: Wiley.

Wiggins, J.S. (1991). Agency and communion as conceptual coordinates for the understanding and measurement of interpersonal behavior. In W. Grove & D. Cicchetti (Eds.), Thinking clearly about psychology: Essays in honor of Paul E. Meehl Vol. 2, (pp.89-113). Minneapolis: University of Minnesota Press.

Yukl, G.A. (1971). Towards a behavioral theory of leadership. Organizational Behavior and Human Performance, 6, 414-440.

Zaleznik, A. (1974). Charismatic and consensual leaders: A psychological comparison. Bulletin of the Menninger Clinic, 38, 222-238.

APPENDIX

Psychometric Properties of the Forceful and Enabling Measure

Source: Kaiser, R.B. & Craig, S.B. (2001). [Item selection and validation for forceful and enabling leadership scales: Exploratory and confirmatory analyses of data from 1,140 raters of 107 executives.] Unpublished data analyses. Kaplan DeVries Inc.

Preliminary validation of the psychometric properties of our measure of forceful and enabling leadership has been reported previously (Kaiser & Kaplan, 2000). With an increased sample including the data in the preliminary study, we have conducted further construct validation research (Kaiser & Craig, 2001) using a measure including an initial set of 11 forceful and 11 enabling items rated on the "curvilinear" response scale shown below. This appendix includes a summary of those analyses, which resulted in refined five-item forceful and five-item enabling scales with adequate measurement characteristics. These scales were used in all quantitative references in the main text of this paper.

Figure 1. *The implicitly "curvilinear" scale*

Please rate the manager in question on each of the following aspects of executive leadership.

Please note that the scale is probably different from scales that you are accustomed to using. On this scale the best score is "2," smack in the middle of the scale. The premise is that performance problems arise when managers either underdo or overdo something.

Too little		The right amount			Too much	
0.5	**1**	**1.5**	**2**	**2.5**	**3**	**3.5**

Sample

The sample used in the validation study included 360° ratings for 107 senior executives (vice presidents on up to CEOs) from a variety of U.S. firms. The executives tended to be white men between the ages of 40 and 60. Included are ratings from 104 self-raters and 1,036 coworkers—165 superiors, 362 peers, and 509 subordinates. We randomly split the sample of coworker ratings in half—the first (n = 519) to be used as a development sample and the second (n = 517) as a holdout validation sample—to create forceful and enabling scales with adequate measurement properties. The full sample was then used to assess test functioning with item response theory, inter-rater agreement and reliability, relationships between forceful and enabling leadership as rated by multiple sources, and relationships with effectiveness.

Measurement Equivalence Across Rating Sources

Since recent research has indicated that rating source has only a trivial effect on the latent structure of 360° data (e.g., Mount, Judge, Scullen, Systma, & Hezlett, 1998; Scullen, Mount, & Goff, 2000), we combined data from selves, superiors, peers, and subordinates in all structural analyses. As a check on the appropriateness of this procedure, we first assessed measurement equivalence (differential item functioning) on the 11 item pairs across the rating sources with item response theory (IRT; Hambleton, Swaminathan, & Rogers; 1991) using the full sample of self- and coworker ratings

Measurement equivalence was investigated using the IRT-based "differential functioning of items and tests" (DFIT) framework (Raju, van der Linden, & Fleer, 1995). All possible pair-wise comparisons between the four rating sources were examined for the degree to which each of the 22 original items (DIF) and the two 11-item a priori scales (DTF) were equivalently related to the underlying forceful and enabling constructs (c.f., Facteau & Craig, 2001). No significant differences were found for any of the parameters in any of the six comparisons, indicating that scores from each rating source are on equivalent metrics and are thus directly comparable.

Selection of Items and Factor Structure

In the developmental stage, we conducted an iterative series of exploratory factor analyses using maximum likelihood procedures and oblique rotation methods to extract two factors from the original set of 11 forceful and 11 enabling items. After each factor analysis, we dropped the poorest performing item—defined on the basis of significant cross loadings on the unintended factor—and repeated the process. We also conducted IRT-based analyses to ensure that items being dropped for cross-loading problems weren't worth reconsidering on the basis of relative contribution of information about respondents' standing on the underlying construct.

The analyses with the development sample indicated that six of the original 11 item pairs didn't function adequately, but five item pairs had promise. Thus, we used confirmatory factor analysis (CFA) to evaluate a forceful and enabling measurement model with ten indicators, the remaining five item pairs. The behavioral core of these items is presented in Table 1.

Table 1. *Behavioral core of forceful and enabling items*

Forceful	*Enabling*
1f. Strong leader	1e. Enables subordinates
2f. Declares self	2e. Receptive to others' ideas
3f. Makes tough calls	3e. Compassionate
4f. Makes judgments	4e. Makes tough calls
5f. Forces issues	5e. Fosters harmony

Three alternative structural models of forceful and enabling leadership were compared using CFA on the holdout sample of 621 individual raters—including 104 self-raters, 82 superiors, 181 peers, and 254 subordinates. A one-factor model was tested to determine how well the data fit a structure corresponding to a continuum with forceful leadership on one end and enabling on the other. A two non-correlated factors model was evaluated to test how well the data fit a structural model in which forceful and enabling factors are unrelated. The final model tested was the one we hypothesized by the underlying polarity theory—a two correlated factors structural model where the forceful and enabling constructs are *inversely related*.

The fit of each model was tested using the CALIS procedure of SAS (SAS, 1996). Parameters were estimated with the maximum likelihood method. Following recommendations in the literature, multiple indices were used to assess model fit (Hu & Bentler, 1995). We adopted the conventional wisdom that adequate fit is indicated when CFI, GFI, AGFI, and NNFI values exceed .90; RMSR values fall below .06; and RMSEA values are less than .08 (Hu & Bentler, 1999).

Following Fornell and Larcker (1981), we next examined the magnitude of factor loadings in the confirmed two correlated factors model. They recommend as a stringent test for reliability that factor loadings should approximate .70, which would suggest that less than half of the item's variance is due to unmeasured sources. Third, we calculated the average variance extracted by each construct from the items, which is recommended to be .50 or higher (Fornell & Larcker, 1981; Hooijberg & Choi, 2000).

Table 2. *Fit indices for three alternative models of forceful and enabling leadership*

			Fit Indices					
Model	χ^2	*df*	CFI	GFI	AGFI	NNFI	RMSR	RMSEA
One factor	404.03	35	.73	.82	.72	.65	.02	.15
Two non-correlated factors	211.13	35	.87	.93	.88	.83	.03	.10
Two correlated factors	129.22	34	.93	.95	.92	.91	.01	.07

Note: N = 621 total raters. *df* = Degrees of Freedom, CFI = Comparative Fit Index, GFI = Goodness of Fit Index, AGFI = Adjusted Goodness of Fit Index, NNI = Non-Normed Fit Index, RMSR = Root Mean Square Residual, RMSEA = Root Mean Square Error of Approximation.

As shown by the fit statistics in Table 2, the hypothesized model provided the best fit to the data. The five forceful items cohere in one factor that is distinct from the factor formed by the five enabling items. Further, the significantly improved fit in the correlated factors model over the non-correlated one indicates that the two factors are inversely related. The results also indicate that this polarity effect—the tendency for managers to overdo one and under do the other—is fairly strong at the individual rater level of analysis: the estimated true correlation between the forceful and enabling constructs, corrected for measurement error, was r = - .50 (t = -10.91, p < .001). Table 3 presents the factor loadings for the 10 items based on the confirmed two correlated factors model.

Table 3. *Factor loadings for forceful and enabling items*

	Factor Loading	
Item (behavioral core of text)	I	II
1f. Strong leader	.56	
2f. Declares self	.67	
3f. Makes tough calls	.68	
4f. Makes judgments	.67	
5f. Forces issues	.71	
1e. Enables subordinates		.52
2e. Receptive to other's ideas		.67
3e. Compassionate		.67
4e. Shows appreciation		.62
5e. Fosters harmony		.65

Note: All loadings significant (p < .001)

Reliability

Reliability for the two scales was assessed using a variety of methods. First, Cronbach's coefficient alpha was computed as one index of internal consistency. Finally, we used IRT to estimate the marginal reliability of each scale, which roughly corresponds to the average reliability of scores across all possible levels on the underlying construct (Thissen, 1995). These reliability estimates are presented in Table 4.

Table 4. *Reliability estimates for forceful and enabling scales*

Reliability statistic	*Forceful*	*Enabling*
Coefficient Alpha	.80	.76
Average factor loading of items	.66	.63
Average variance extracted	.43	.40
Marginal reliability	.81	.77

Note: For alpha and marginal reliability, N = 1,140 raters, including 104 selves, 165 superiors, 362 peers, and 509 subordinates. For average factor loadings and variance extracted, N = 621 raters, including 104 selves, 82 superiors, 181 peers, and 254 subordinates.

These results paint a mixed picture. Both scales exceed the traditional rule of thumb that alpha reliability coefficients should exceed .70 (e.g., Nunnally, 1978). The marginal reliability estimates also exceed this recommended value. However, the estimates fall short of the high bar set by requiring average factor loadings to approximate or exceed .70. Also, the average amount of variance extracted from the items by the underlying constructs fell bellow .50, suggesting that over half of the variance on scale scores is due to unmeasured sources. These latter two shortcomings are likely due to construct under-representation (Messick, 1995): forceful and enabling leadership are broad constructs and the current measure represents them with only five items each. Including more items that tap under-represented elements of each will likely increase the amount of common variance and thus reduce the proportion of error variance.

Recognizing that the current instrument does contain a non-trivial degree of measurement error, we still believe that it is a useful tool. To the extent that the scales contain error variance, they will underestimate the true relationships shown in correlations with other variables, including each other. Nonetheless, future work on instrument development will include the generation and validation of additional items to more adequately map the respective construct domains.

Test Functioning as Assessed by Item Response Theory

We analyzed the measurement properties of the two five-item scales with IRT to further understand how they function. Because IRT generates parameters that are not sample specific and does require large sample sizes to estimate parameters, we used the total sample of 1,140 sets of ratings—including 104 selves, 165 superiors, 362 peers, and 509 subordinates. A major difference between IRT and classical test theory approaches to measurement is the IRT acknowledgement that items/scales demonstrate different measurement properties at different levels on the underlying construct. IRT allows one to identify how precisely the items and the scale reflect performance across the hypothetical distribution of performance in the population on the latent trait (referred to as theta) measured by the items/scale. This is done by computing the Standard Error for each interval along the standardized theta distribution.

Figure 2. *Standard error functions for the forceful and enabling scales*

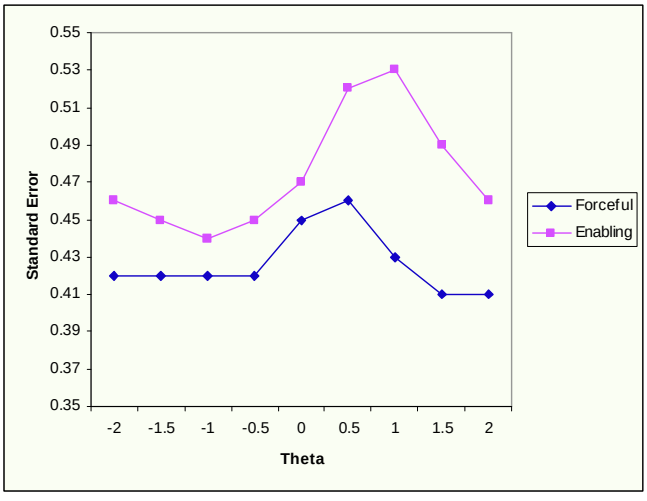
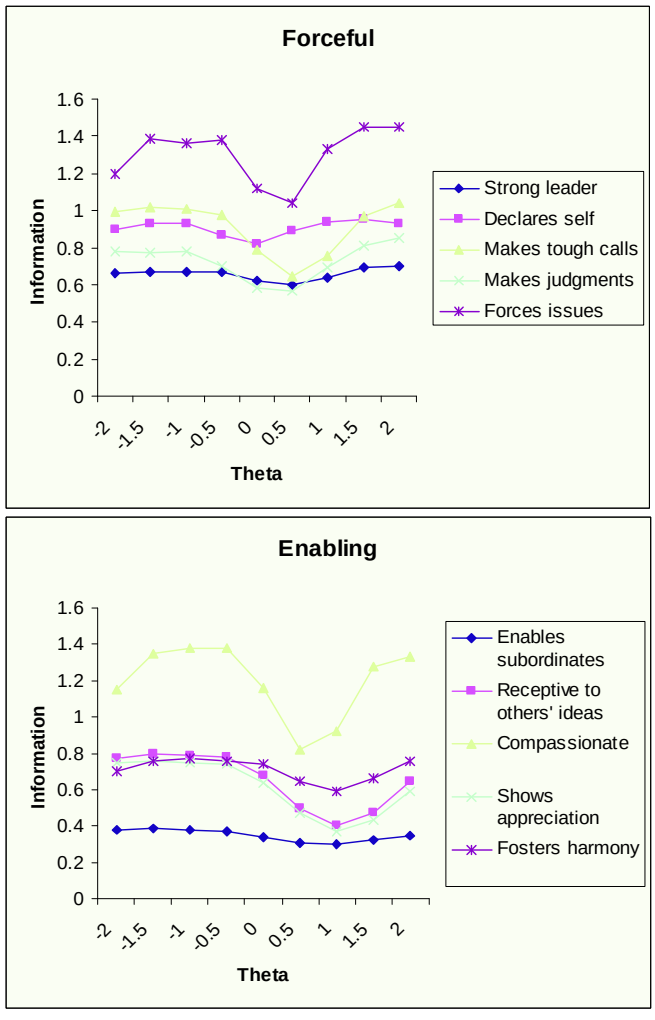


Figure 2 provides a plot of the Standard Errors across theta for the forceful and enabling scales. Two observations are worth pointing out. First, the enabling scale shows higher Standard Errors across the performance distribution compared to the forceful scale. It is the less precise, which is consistent with the reliability analyses above. Precision, in this case, refers to the level of confidence we can have in scale scores' correspondence to true standing on the latent construct (i.e., to the width of confidence intervals). Second, the peak of the Standard Error function occurs around .5 SD above the mean for forceful and 1 SD above the mean for enabling. These points roughly correspond to 2.00 ("the right amount") on the original response scale. Thus, the scales are more precise at the underdo and overdo extremes, and less precise at the optimal response level.

IRT is also a powerful analytic tool for understanding how much information each item in a scale provides relative to the other items. This is particularly useful when it comes time to generate additional items: items can be written to target ranges on theta at which the current test items are less precise. Finally, it also provides a sense of which items do a good job of estimating different levels of performance.

Figure 3 provides the item information functions for the two scales. Item "information" is simply the inverse of Standard Error. The item "forces issues" provides the most relative information (has the lowest Standard Errors) on the forceful construct; "strong leader" the least. The item "compassionate" is the most relatively informative of the enabling indicators; "enables subordinates" is the least.

Figure 3. *Relative information provided by the forceful and enabling items*



These results help to clarify the conceptual meaning of our operationalization of forceful and enabling leadership.

Forceful scores are mainly driven by perceptions of the leaders' agentic assertion of self in raising difficult issues, stating his or her positions on matters, and stepping up to hard decisions. Enabling scores are largely a function of the degree to which the leader is seen as being responsive and showing consideration for people's feelings, maintaining relations within the group, and being open to influence. This can be taken as construct validity evidence that the two scales are conceptually similar to the twin pillars of initiating structure and consideration as well as the performance versus group maintenance distinction as articulated in the leadership literature (e.g., see reviews in Bass, 1990).

Inter-rater Agreement and Reliability

Since the measure is to be used as a multi-rater feedback instrument, we assessed the extent to which it is appropriate to aggregate ratings within superior, peer, and subordinate rating sources. Recall that the response scale created for this measure is new; it combines features of traditional response formats and could be called an "*evaluation of behavior frequency*" scale. It requires raters to make a value judgment as to what constitutes "too much" or "too little" for a particular item. Since managers vary in the implicit mental models of effective leadership used to guide their judgments (Lord & Maher, 1993; Sivasubramaniam, Kroeck, & Lowe, 1997), it is possible that these value-laden assessments are purely "in the eye of the beholder." If this were the case, then our instrument would be of little value as a research tool or feedback instrument because the ratings would say more about the raters than the focal leader. Thus, our analysis included close attention to inter-rater agreement.

Whereas inter-rater *reliability*—the most commonly used index of similarity of ratings within a group—assesses rating congruence in rank-ordering or correlational terms, inter-rater *agreement* provides information on how similar ratings are in terms of overall level (Fleenor, Fleenor, & Grossnickle, 1996; James, Demaree, & Wolf, 1984; 1993). To illustrate the difference, consider two hypothetical raters who evaluated the same manager on three items. Suppose the first rater gave scores of 1, 1.5, and 2, and the second rater gave scores of 2, 2.5, and 3. The inter-rater reliability of these two sets of ratings (i.e., the correlation between them) would achieve unity, 1.00. However, inter-rater agreement (i.e., level of agreement) would tell a different story: the mean of the first rater's evaluations would be 1.5 whereas that for the second rater would be 2.5. On our scale, the first set of ratings would indicate underdo whereas the second set would show overdo.

We reasoned that it is critical to demonstrate that raters of the same manager give ratings that are roughly equivalent in overall level (high inter-rater agreement) as well as ratings that are reasonably correlated (inter-rater reliability). This could be taken as evidence that, despite differences among raters in their personal theories of effective leadership, they are able to reach a reasonable degree of consensus about whether the focal leader does too much, optimally, and too little with respect to forceful and enabling leadership.

The two five-item scales were evaluated separately in terms of inter-rater agreement to determine whether raters agreed on what the focal leader does too little, just right, and too much. James' $r_{wg(j)}$ statistic was used for this purpose (James et al., 1984; 1993). Separate $r_{wg(5)}$ values were computed for each focal leader for each rating source where two or more coworkers provided ratings. Specifically, $r_{wg(5)}$ statistics were calculated for superior ratings of 36 targets, peer ratings of 88 targets, and subordinate ratings of 106 targets. The mean $r_{wg(5)}$ for both scales computed across rating targets is presented in Table 5. These values exceed standards (e.g., James et al. 1984; 1993) for an acceptable level of agreement within rating groups.

Inter-rater reliability was estimated with intraclass correlations (ICCs; Shrout & Fleiss, 1979). ICCs were calculated for the average number of raters per source. ICCs were computed as the reliability of the mean rating for a random sample of two superiors per target (where possible) and random samples of three peers and three subordinates—for superiors ICC[2,2] (n = 36), for peers ICC[3,3] (n = 83) and for subordinates ICC[3,3] (n = 100) (see Shrout & Fleiss, 1979). As reported in Table 5, ratings on these scales generally meet recommended standards (e.g., .70; Nunally, 1978). Moreover, these inter-rater reliabilities compare favorably to meta-analytic estimates of 360° ratings of middle managers across a variety of performance dimensions (Conway & Huffcut, 1997).

Table 5. *Inter-rater agreement and reliability for aggregate scores on forceful and enabling scales*

	Forceful scale			*Enabling scale*		
Rating Source	ICC (single)[1]	ICC (mean)[2]	$Mr_{wg(5)}$	ICC (single)[1]	ICC (mean)[2]	$Mr_{wg(5)}$
Superiors	.55**	.71**	.89	.45**	.62**	.91
Peers	.46***	.72***	.90	.42***	.69***	.88
Subordinates	.51***	.76***	.91	.43***	.69***	.89

Note: Intraclass correlation coefficients (ICCs) are presented as estimates for [1]the reliability of a single rating and [2]for the mean of n raters (superiors n = 2, peers n = 3, subordinates n =3).

*** $p < .001$ ** $p < .01$

Table 6. *Aggregated scale descriptive statistics, alpha reliabilities, and inter-correlations*

		Forceful			*Enabling*			
Rating Source	*N*	*M*	*SD*	α	*M*	*SD*	α	$r_{\text{Forceful \& Enabling}}$
Self	104	1.96	.35	.78	1.88	.26	.63	-.31***
Superiors	98	1.89	.31	.83	1.80	.26	.75	-.40***
Peers	97	1.96	.31	.87	1.74	.25	.83	-.58***
Subordinates	107	1.92	.28	.90	1.75	.23	.86	-.53***
All Coworkers	107	1.91	.26	.93	1.77	.21	.88	-.58***

*** $p < .001$

Despite the unconventional and highly subjective nature of this rating scale, coworkers reached a good deal of consensus and consistency about the extent to which target leaders overdo and underdo forceful and enabling leadership. Thus, aggregating ratings within rating sources is empirically justified.

Aggregated Scale Score Descriptive Statistics

Table 6 provides the descriptive statistics and coefficient alphas for scores on the forceful and enabling scales aggregated within ratings sources for the 107 target executives in our database. The *n*'s vary across rating groups due to no available data for some rating sources (e.g., no peer ratings for CEOs). Additionally provided are the descriptive statistics and alphas for *all coworker* ratings, which reflects the aggregation of data across all coworkers who provided ratings for a given executive.

Also shown in Table 6 is the correlation between forceful and enabling leadership at the aggregate level. Again, evidence is shown for the hypothesized polarity effect—sizable negative correlations. It is noteworthy that the effect is less pronounced in self-rating data than in any of the coworker data.

Convergent Validity Across Rating Sources

Another form of validity evidence can be found in the extent to which ratings from different sources are correlated. But, it is desirable for ratings from the traditional 360° perspectives to correlate higher within sources than between sources. As has been noted (e.g., Borman, 1974; Murphy & Cleveland, 1995), constituents at different organizational levels have different interactions with, expectations of, and opportunities to observe a given manager's performance. Nonetheless, to the extent that a performance trait is characteristic of a manager, there should be some degree of convergence across sources. Table 7 shows the correlations between the various rating perspectives' average ratings on the forceful and enabling scales.

There was indeed a good deal of convergence across rating sources. In fact, these correlations are slightly higher than meta-analytic estimates of correlations for cross-source convergent validity coefficients on managerial performance scales reported in the literature (Conway & Huffcut, 1997). Perhaps forceful and enabling are more robust and stable leadership characteristics than are the myriad other dimensions that have been studied.

It is worth noting that the correlations between self-ratings and each of the other sources are generally lower than the correlations between the other three sources. In part, this may be due to lower reliability of self-ratings compared to the aggregate ratings within other sources. The result is also consistent with the idea that self-ratings are more biased than are observer ratings. Self-ratings are most consistent with superior ratings, suggesting that self-perceptions may be more influenced by superior relationships.

Finally, comparing these correlations between sources to the within source correlations reported in Table 5 [ICC(mean)] suggest that there is more convergence within sources than between sources. This supports the practice of reporting feedback results separately for each rating group. However, given the relatively high correlations between sources, there is credence to also present results for the average ratings across all coworkers.

Table 7. *Between source correlations on forceful and enabling*

	Source			
Source	Self	Superiors	Peers	Subordinates
Self	--	.44	.30	.36
Superiors	.62	--	.64	.52
Peers	.57	.68	--	.57
Subordinates	.46	.63	.69	--

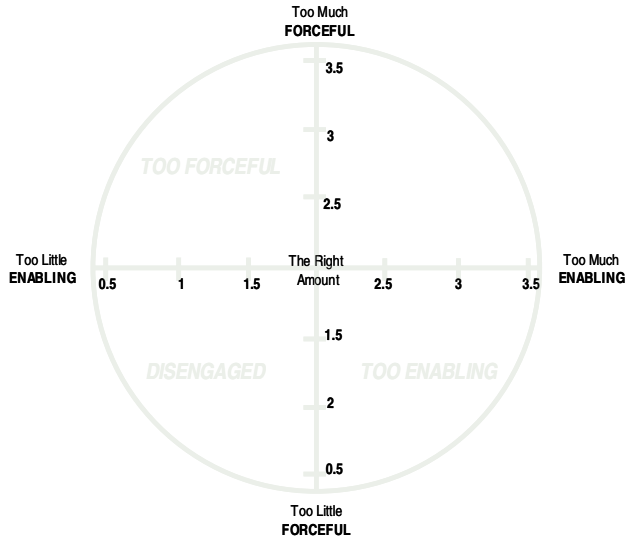
Note: Correlation coefficients above the diagonal are for the enabling scale; coefficients below the diagonal are for the forceful scale. All correlations are significant ($p < .001$).

Four Types of Leaders

More important that examining scores on forceful and enabling leadership separately is considering them in tandem. Patterns of scores across both factors can be used to identify four basic types of leadership style: Too forceful, too enabling, disengaged, and versatile. Three of the four basic patterns are similar to the extreme forms of the three leadership styles identified in Lewin's seminal work (e.g., Lewin, Lippitt, & White, 1939): Too forceful (like Lewin's autocratic style), Too enabling (democratic), and Disengaged (Laissez faire). The fourth leadership style, versatile, is indicated by scores that are not

significantly different from "the right amount" on both forceful and enabling. Figure 4 shows where the three extreme types of leaders are located in a two-dimensional conceptual space defined by forceful and enabling leadership. Note that the upper right-hand quadrant, the area representing overdoing both forceful and enabling, is a conceptual null set.

Figure 4.*Two-dimensional conceptual space defined by forceful and enabling leadership*



The present sample is far too small for establishing normative frequencies for the four types in the executive population. Nonetheless, we did examine the frequencies of the leadership types in our sample represented by the average scores for each rating source. The operational definition of these types were: *Too Forceful* = Forceful $\geq$ 2.0 and Enabling < 2.0; *Too Enabling* = Forceful < 2.0 and Enabling $\geq$ 2.0; *Disengaged* = Forceful < 2.0 and Enabling < 2.0. Leaders were classified as *Versatile* only if their scores on *both* forceful *and* enabling were not significantly different from 2.0.[3] The observed frequency counts are reported in Table 8.

[3] To determine if forceful and enabling scores were not significantly different from 2.0, we used IRT to calculate scores for each rater as well as the standard error for those scores. An advantage to IRT over classical test theory approaches is that it provides unique standard error estimates for each *individual rater* based on patterns of responses across items (Hambleton et al., 1991). We next computed the average score across raters within a rating source as well as the average standard error. Next, we created 95% confidence intervals around the aggregated IRT scores for each target manager using the average standard error. Targets were deemed versatile if this confidence interval included the score corresponding to "the right amount" on *both* forceful *and* enabling scales. Thus, we classified as versatile only those leaders whose scores on both constructs could not be said with 95% confidence to be significantly different from optimal.

Chi-square analysis of the different frequencies across the cells indicated statistical significance (χ^2 (9) = 19.37, p < .05). As can be seen in the frequency counts in Table 8, overly forceful and disengaged types outnumber overly enabling and versatile types as rated by all coworker sources. In the self-ratings data, there were no differences in frequencies across the four types. Further, versatile and too enabling types occurred in the self-rating data significantly more often than in the three coworker data sources. The disengaged type was significantly less common in the selves' data than in the coworkers' data. Within the three different coworker rating sources, the rarer frequency of versatile types is significantly different from the more common occurrence of too forceful and disengaged types but not too enabling types. The higher incidence of disengaged and lower incidence of too enabling types compared to too forceful and disengaged types in subordinate ratings is statistically significant. Finally, no executives were rated by any source as "overdoing" on both forceful *and* enabling leadership, which constitutes further evidence of construct validity for the measure.

We also computed the frequency counts for the four leadership types based on the average ratings from all coworkers. Those frequencies were: 11 versatile (10.3%), 30 too forceful (28.0%), 15 too enabling (14.0%), and 51 disengaged (47.7%).

Versatility

A key concept in the polarity-based view of forceful and enabling leadership is versatility, managers' tendencies to make appropriate use of both forceful and enabling approaches. A measure of such versatility can be derived from the forceful and enabling scores. This measure–which we call the *versatility index*–considers jointly the extent to which managers use forceful leadership and enabling leadership. It is computationally equivalent to the *squared Euclidean distance metric* and reflects the extent to which the individual is rated at, or close to, the midpoint on both scales marked "does the right amount" (scale value of 2.00).

For example, in the upper-left-hand quadrant of Figure 5 is a plot for an executive who scored in the "does too much" region on forceful and in the "does too little" region on enabling. The geometric distance this leader is from being perfectly versatile—that is, distance from a score of the right amount (2.00) on forceful and a distant from a score of the right amount (2.00) on enabling—can be derived from the Pythagorean theorem. It is calculated as: $c^2 = a^2 + b^2$ where a = (forceful score – 2), b = (enabling score – 2), and c = *distance from optimal versatility.*

To compute the versatility index, we calculated the ratio of each executive's observed *distance from optimal*

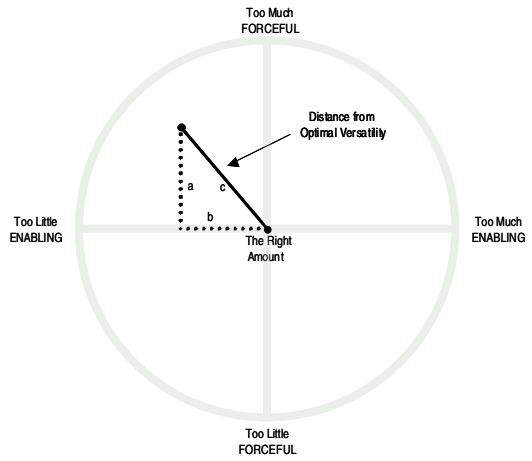
Table 8. *Frequency counts for leadership types*

	Rating Source							
	Self		Superiors		Peers		Subordinates	
Leadership Style	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*
Versatile	23	22.1[a]	16	16.3	15	15.5	12	11.2
Too Forceful	29	27.9	28	28.6	35	36.1	32	29.9
Too Enabling	27	26.0[a]	21	21.4[a,b]	18	18.5[b,c]	13	12.2[c]
Disengaged	25	24.0[a]	33	33.7	29	29.9[a]	50	46.7[b]
N	104		98		97		107	

Note: Percentages with different superscripts between rating groups are significantly different ($p < .05$).

versatility to the maximum possible distance from optimal versatility (i.e., scores on the extreme ends of the response scale, 0.5 and 3.5). This ratio is an inverse measure of versatility. So that higher values would indicate more versatility, we subtracted this value from 1.00. Therefore,

Figure 5. *Computing the "distance from optimal versatility"*



versatility indices can range from 0 to 1.00, where lower values indicate a greater degree of lopsidedness. Table 9 provides descriptive statistics for the versatility index as computed separately for the average ratings within each rating source as well as for the average of all coworker ratings for each target manager.

Concurrent Validity

An important form of construct validity evidence for a measure of leadership is the degree to which it is related to important outcomes. For a sub-sample of the present sample, we had overall effectiveness ratings available. Raters were asked in a semi-structured interview conducted at a different time from the forceful and enabling rating task to "Please give a rating of X's overall effectiveness as an executive on a ten-point scale, where 10 is outstanding and 5 is adequate." Effectiveness ratings were available for a total sample of 78 target executives. Effectiveness ratings were collected from a total of 76 self-raters, 142 superiors (who rated 74 target leaders), 282 peers (70 targets), and 403 subordinates (78 targets).

Validity of effectiveness measure. Although single-item measures are not inherently flawed (Judge & Ferris, 1993), they are often suspected to lack adequate measurement characteristics. We looked at the psychometric properties of the effectiveness ratings in terms of inter-rater agreement and inter-rater reliability within rating sources (cf. Fleenor et al., 1996) and convergent validity between rating sources. Inter-rater agreement for each rating source was assessed with James' r_{wg} statistic (James et al., 1984; 1993), calculated separately for each target where two or more raters provided data. This index is appropriate when a group of raters rate a single target on a single variable or construct and the researcher wants to know the extent to which the overall *level* of ratings is similar across the individual raters. Similar to the interpretation of indices of reliability, r_{wg} values closer to 1.00 indicate better measurement properties.

Table 9. *Versatility index descriptive statistics*

Rating Source	*N*	*M*	*SD*	*skew*	*minimum*	*maximum*
Self	104	.81	.10	-.45	.53	.95
Superiors	97	.82	.13	-.93	.39	1.00
Peers	97	.78	.12	-.89	.37	1.00
Subordinates	107	.79	.11	-.84	.42	.95
All Coworkers	107	.79	.09	-.79	.53	.93

Table 10. *Descriptive statistics and validity evidence for effectiveness ratings*

Rating Source	*N*	*M*	*SD*	ICC (single)[1]	ICC (mean)[2]	Self	Superiors	Peers	Subs
Self	76	7.35	1.23	--	--	--			
Superiors	74	7.82	1.37	.71***	.83***	.35**	(.83)		
Peers	70	7.50	1.14	.40***	.72***	.27*	.70***	(.78)	
Subordinates	78	7.82	.93	.42***	.75***	.36**	.50***	.31**	(.83)

Note: Coefficients along the diagonal are Mr_{wg}, computed as the average r_{wg} across target executives within coworker rating sources. Intraclass correlation coefficients (ICCs) are presented as estimates for [1]the reliability of a single rating and [2]for the mean of *n* raters (superiors *n* = 2, peers *n* = 4, subordinates *n* =4).
*** $p < .001$ ** $p < .01$ * $p < .05$

It is worth noting that multi-item scales of overall managerial effectiveness often include an item that is very similar in wording to the present measure. For example, the scale used in the program of research inspired by Quinn's (1988) competing values framework comprises five items, including one worded, "Overall effectiveness as a manager." Within each 360° rating source, this item has been shown to be the highest loading of all five items on the underlying overall effectiveness factor, usually exceeding .90 (e.g., Hooijberg & Choi, 2000). Thus it appears that variance in such a general item tapping perceptions of overall effectiveness contains a good deal of the common variance among multiple indicators of this construct.

With the inter-rater agreement and reliability results, convergent validity correlations, and conceptual mapping onto similar multi-item scales, we interpreted this single-item effectiveness rating as a reasonably valid and reliable measure of perceived overall effectiveness.

Forceful and enabling leadership and effectiveness. Recall that scores on the forceful and enabling scales range from "too little" (underdo) to "too much" (overdo). Therefore, to assess their relationships with the effectiveness ratings, we applied quadratic, or "curvilinear," regression analyses. Each predictor in such a model is represented by parameter estimates for the following terms: constant, the predictor, and the predictor squared. Quadratic regression models were constructed by regressing the effectiveness measure onto forceful and enabling scores separately for each rating source and for scores derived from the average of all coworkers' ratings. The results appear in Table 11.

As expected, the significant quadratic functions relating forceful and enabling scores to effectiveness ratings curve upward as lower scores in the underdo range on forceful (and enabling) approach 2.00 (the optimal point) and then curve downward as they approach the overdo extreme. This suggests that the response scale format works as intended: as greater departures from "does the right amount" are associated with decreased effectiveness.

Also, it is clear that in these data forceful leadership is more strongly related to effectiveness than is enabling leadership. This may be due to the somewhat more restricted range on enabling in this sample: very few of the executives were rated in the overdo range on enabling and the few that were tended to be rated as only slightly overdoing enabling.

It is also noteworthy that self-ratings on both sides of leadership were unrelated to self-ratings of effectiveness. This points to the various biases that influence self-ratings, including the biased view that one's own style of leadership is more effective than alternative styles.

Recognizing that using ratings on leadership and effectiveness from the same source likely inflates their relationships because of common method effects, we also looked at the relationships between subordinate-rated leadership and superior-rated effectiveness. Those results are presented in Table 12.

Table 11. *Regressions of effectiveness onto forceful and enabling scores separately*

	Forceful				Enabling			
Rating Source	β0	β1	β2	R^2	β0	β1	β2	R^2
Self	4.97	1.96	-0.37	.02	6.63	1.60	-0.63	.03
Superiors	-10.00	17.75***	-4.30**	.39***	-9.29	20.16***	-5.80**	.19***
Peers	-6.67	14.07**	-3.40**	.30***	-2.49	11.40**	-3.19**	.09*
Subordinates	-5.68	13.49**	-3.28**	.33***	4.47	2.98*	-0.60	.05*
All coworkers	-10.54	5.66**	-5.35**	.36***	0.54	1.87	-1.74	.04

Note: β0 = intercept, β1 = beta weight for predictor, β2 = beta weight for predictor-squared.
*** $p < .001$ ** $p < .01$ * $p < .05$

Table 12. *Regression of superior-rated effectiveness onto subordinate-rated forceful and enabling scores*

	Forceful				Enabling			
	β0	β1	β2	R^2	β0	β1	β2	R^2
	-4.66	2.63**	-2.47**	.14**	8.90	-.38	0.54	.03

Note: β0 = constant, β1 = beta weight for predictor, β2 = beta weight for predictor-squared.
** $p < .01$

These results were similar to those within rating sources in that the significant function relating subordinate-rated forceful to superior-rated effectiveness peaked around the point corresponding to "the right amount" (2.00). Important was the non-significant relationship between subordinate ratings of enabling leadership and superior-rated effectiveness. Again, this may be attributable to the very few ratings of overdo on enabling. When only the underdo ratings on enabling (i.e., enabling < 2.00) were correlated with superior ratings of effectiveness, the effect size increased, although it was still a smaller effect than the corresponding one for forceful scores.

Simultaneous consideration of forceful and enabling leadership and effectiveness. The dynamic tension between forceful and enabling leadership inherent in a polarity-based conception highlights the importance of examining them in concert. The underlying hypothesis is that the most effective managers will be versatile, as indicated by scores that approach 2.00 ("the right amount") on both scales. Said alternatively, managerial ineffectiveness is thought to be associated with stronger tendencies to overdo and underdo across the two sides of leadership.

To first test the hypothesis that versatility is related to effectiveness, we examined the correlation between the versatility index and effectives ratings within each rating source. Also examined were these correlations between rating sources. Recall that the versatility index represents the degree to which scores on both forceful and enabling approach 2.0, "the right amount." Lower scores on the versatility index indicate greater departure from this optimal pattern. These results are shown in Table 13.

It is evident that there is a fairly strong link between versatility on the forceful and enabling polarity and effectiveness within and between coworker rating sources. As expected, the correlations within rating sources (e.g., versatility and effectiveness both based on subordinates data) are greater than the correlations between sources (e.g., versatility based on subordinates, effectiveness based on superiors). This is due to the fact that ratings of leadership and effectiveness within groups are both based on the same expectations, observations, and prototypes as well as sources of rating error within groups, but not between groups. The correlations between sources are nonetheless practically significant (except for subordinate-rated versatility and peer-rated effectiveness) and rule out attributing the relationship between versatility and effectiveness to an artifact of common method bias.

Again, it is important to note the lack of an association between target managers' ratings on forceful and enabling leadership and effectiveness as rated by themselves or the three coworker rating sources. This is consistent with the idea that most leaders tend to view their preferred style as most effective. It also highlights the importance of developmental feedback: these executives did not see the link between their lopsidedness and ineffectiveness that was so clearly apparent to their coworkers.

A more complete model of forceful and enabling leadership and effectiveness. The preceding analyses offer support for the construct validity of the forceful and enabling theory and measure by demonstrating a sizable relationship between the versatility index and effectiveness ratings. However, those correlational analyses provide an incomplete picture of the link between forceful and enabling leadership and effectiveness. Specifically, correlating the versatility index with effectiveness across all four leadership style types (i.e., versatile, too forceful, too enabling, and disengaged) assumes that the relationship is the same for all four types. However, there is theoretical reason to suspect that this may not be the

Table 13.
Correlation between versatility index and effectiveness ratings within and between sources

Source of	*Source of Effectiveness Rating*				
Versatility Index	Self	Superiors	Peers	Subordinates	All coworkers
Self	.08	.22	.21	.17	.24*
Superiors	.16	.68***	.52***	.42***	.62***
Peers	.09	.44***	.55***	.31**	.47***
Subordinates	.11	.31**	.16	.59***	.44***
All coworkers	.06	.49***	.36**	.50***	.53***

Note: *** $p < .001$ ** $p < .01$ * $p < .05$

case. For one, the too forceful and too enabling styles are both active leadership patterns, which are quite likely less detrimental to effectiveness than the relatively inactive disengaged pattern (Bass, 1990). Thus, the correlation between versatility and effectiveness may be weaker for disengaged types.

Another reason is based on social psychological research in the interpersonal circumplex tradition. Similar to our conception of forceful and enabling leadership, interpersonal theorists view social behavior as involving two basic dimensions, some form of agency and communion (Wiggins, 1991). Greater departure from a balanced use of both dimensions, referred to as "amplitude" in circumplex research, is differentially related to various social consequences according to different patterns of agency and communion (e.g., high on both, high on one and low on the other; Wiggins & Pincus, 1992).

Thus, we conducted analyses that modeled variance on effectiveness ratings as a function of leadership type, versatility, and the interaction between type and versatility. This was done by constructing a one-way Analysis of Covariance (ANCOVA) model with the four-level leadership type categorical variable and the versatility index as a covariate to predict effectiveness ratings. The procedure was repeated separately for the self-rating and three separate coworker ratings sources. The results are presented in Table 14.

As with prior analyses, we also ran an ANCOVA model using the all coworkers' data. Like the within-source models, this too was highly significant ($F(7,71) = 9.66$, $p < .001$, $eta^2 = .49$).

The results of the ANCOVA models warrant several points. First, the type by versatility index interactions were significant in all three coworker rating sources. Thus, the correlation between the versatility index and effectiveness was not uniform across the four types of leadership styles. As expected, the correlation was weakest in magnitude for the disengaged type. Second, the significant main effect for leadership type revealed that in all three coworker rating sources the mean effectiveness rating for the versatile type was significantly higher than that for the disengaged type. Additionally, in the superior data, the mean effectiveness rating for the versatile type was significantly higher than that for the too forceful and too enabling types. The pattern of mean effectiveness ratings across all three coworker rating sources showed that the versatile managers tended to be rated highest, followed by too forceful and too enabling managers whose mean effectiveness ratings were about the same, and the disengaged managers were rated the lowest. Third, as seen in the prior analyses, self-ratings of leadership were not related to self-ratings of effectiveness, once again emphasizing the tendency for managers to be blind to the sources of their own ineffectiveness.

Finally, and perhaps most important, the amount of variance in effectiveness ratings accounted for by the full model of forceful and enabling leadership indicates a large and practically important effect size. Across the rating sources, scores on forceful and enabling leadership accounted for 46 to 62 percent of the variance in effectiveness. When one considers the multi-dimensional nature of perceptions of overall effectiveness, it is remarkable that scores on two dimensions of leadership have such strong explanatory power.

We also used ANCOVA to model the simultaneous effects of forceful and enabling leadership as rated by subordinates in predicting effectiveness as rated by superiors. Again the issue was to rule out common method bias as an explanation for significant relationships show between effectiveness and leadership type, versatility, and the type by versatility interaction. This model was also significant, $F(7,66) = 3.65$, $p < .01$, $eta^2 = .28$, and mirrored those based on the within-source data. However,

Table 14.
ANCOVA results for predicting effectiveness from leadership type and versatility index

	Rating Source															
	Self				Superiors				Peers				Subordinates			
Source	*df*	*MS*	*F*	eta^2	*df*	*MS*	*F*	eta^2	*df*	*MS*	*F*	eta^2	*df*	*MS*	*F*	eta^2
Full Model	7	2.02	1.38	.13	7	12.12	15.33***	.62	7	5.94	7.68***	.46	7	4.72	9.88***	.50
Leadership Type	3	1.76	1.20		3	4.88	6.17***		3	3.28	4.24***		1	1.47	3.08**	
Versatility Index	1	.03	.02		1	40.19	50.84***		1	15.26	19.73***		3	14.09	29.51***	
Type x Versatility	3	1.77	1.20		3	4.36	5.52***		3	3.07	3.97**		1	1.09	2.28*	
Residual	66	1.47			65	.79			62	.77			70	.48		
Total	74	55.59			73	62.80			70	57.45			78	62.00		

Note: *** $p < .01$ ** $p < .05$ * $p < .10$

the effect size was smaller: the full model of leadership type, versatility, and the type by versatility interaction based on subordinate ratings accounted for 28 percent of the variance in superiors' ratings of effectiveness. This effect size is about half that for the within-source data, but nonetheless indicates a very meaningful practical effect. Clearly, versatility on the forceful and enabling polarity is profoundly important to executive effectiveness.

Summary of Construct Validity Evidence for the Measure

The analyses reported above offer compelling support for the construct validity of the five-item forceful and enabling scales as an operationalization of the forceful and enabling polarity theory. The structural analyses of the instrument provided solid support for the internal characteristics of the measure. A two correlated-factors measurement model represents a good structural fit for the observed ratings on the ten items. The scales demonstrate acceptable levels of internal consistency reliability. And as a 360^{o} rating instrument, raters demonstrate an acceptable level of agreement and reliability within sources as well as good convergence between rating sources in their evaluations of target executives using the two scales. Moreover, scores from the traditional 360^{o} sources on the two scales are on a common metric and thus are directly comparable.

The two scales can be improved, however. In particular, they seem to suffer some degree of construct under-representation. Including more valid items to each scale will likely reduce the amount of error variance in each as estimated with structural equation models. This also will likely lead to enhanced inter-rater reliabilities—especially for the enabling scale—which wavered around the critical value of .70 in the present sample. Also, targeting items to better measure the optimal points on the response scale should reduce the degree of measurement imprecision in that range observed in the IRT analyses.

The instrument works in a way that is consistent with the larger theory in demonstrating the polarity effect, suggesting the advantages of the new response scale format created for this application. Also consistent with theory was the demonstration of a strong link between versatility and effectiveness ratings. In sum, the current measure surpasses basic standards for psychometric adequacy.

Appendix References

Bass, B.M. (1990). Bass and Stogdill's handbook of leadership: Theory, research, and managerial applications (3rd Ed.). New York: Free Press.

Borman, W.C. (1974). The rating of individuals in organizations: An alternate approach. Organizational Behavior and Human Performance, 12, 105-124.

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.

Facteau, J.D., & Craig, S.B. (2001). Are performance appraisal ratings obtained from different rating sources comparable? Journal of Applied Psychology, 86, (in press).

Fleenor, J.W., Fleenor, J.B., & Grossnickle, W. F. (1996). Interrater reliability and agreement of performance ratings: A methodological comparison. Journal of Business and Psychology, 10, 367-380.

Fornell, C. & Larcker, D.F. (1981). Evaluating structural equation models with observable variables and measurement error. Journal of Marketing Research, 18, 39-50.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). Fundamentals of Item Response Theory. Newbury Park, CA: Sage.

Hooijberg, R. & Choi, J. (2000). Which leadership roles matter to whom? An examination of rater effects on perceptions of effectiveness. Leadership Quarterly, 11, 341-364.

Hu, L., & Bentler, P.M. (1995). Evaluating model fit. In Hoyle, R.H. (Ed.) Structural equation modeling: Concepts, issues, and applications (pp. 76-99). Thousand Oaks, CA: Sage.

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure modeling: Conventional criteria versus new alternatives. Structural Equation Modeling, 6, 1-55.

James, L.J., Demaree, R.G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. Journal of Applied Psychology, 69, 85-98.

James, L.J., Demaree, R.G., & Wolf, G. (1993). r_{wg}: An Assessment of within-group interrater agreement. Journal of Applied Psychology, 78, 306-309.

Judge, T. A. & Ferris, G. R. (1993). Social context of performance evaluation decisions. Academy of Management Journal, 36, 80-105.

Kaiser, R.B. & Craig, S.B. (2001). [Item selection and validation for forceful and enabling leadership scales: Exploratory and confirmatory analyses of data from 1,140 raters of 107 executives.] Unpublished data analyses. Kaplan DeVries Inc.

Kaiser, R.B. & Kaplan, R.E. (2000, April). Getting at leadership versatility: Theory and measurement of the forceful and enabling polarity. Paper presented at the 15th annual meeting of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Lewin, K., Lippitt, R., & White, R.K. (1939). Patterns of aggressive behavior in experimentally created social climates. Journal of Social Psychology, 10, 271-301.

Lord, R.G. & Maher, K.J. (1993). Leadership and information processing: Linking perceptions and performance. Boston, MA: Rutledge.

Messick, S. (1995). Validation of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. American Psychologist, 50, 741-749.

Mount, M.K., Judge, T.A., Scullen, S.E., Sytsma, M.R., Hezlett, S.A. (1998). Trait, rater and level effects in 360-degree performance ratings. Personnel Psychology, 51, 557-576.

Murphy, K. R. & Cleveland, J. N. (1995). Understanding performance appraisal: Social, organizational, and goal-based perspectives. Thousand Oaks, CA: Sage.

Nunnally, J.C. (1978). Psychometric theory. New York: McGraw-Hill.

Quinn, R.E. (1988). Beyond rational management. San Francisco: Jossey-Bass.

Raju, N.S., van der Linden, W., & Fleer, P. (1995). An IRT-based internal measure of test bias with applications for differential item functioning. Applied Psychological Measurement, 19, 353-368.

SAS (1996). SAS/STAT user's guide. Vol. 3. Carey, NC: SAS Institute.

Scullen, S.E., Mount, M.K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. Journal of Applied Psychology, 85, 956-970.

Shrout, P. & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86, 420-428.

Sivasubramaniam, N., Kroeck, K.G., & Lowe, K.B. (1997). "In the eye of the beholder": A new approach to studying folk theories of leadership. Journal of Leadership Studies, 4, 27-42.

Thissen, D. (1995). MULTILOG 6.3: A computer program for multiple, categorical item analysis and test scoring using item response theory. Chicago: Scientific Software, Inc.

Wiggins, J.S. (1991). Agency and communion as conceptual coordinates for the understanding and measurement of interpersonal behavior. In W. Grove & D. Cicchetti (Eds.), Thinking clearly about psychology: Essays in honor of Paul E. Meehl Vol. 2, (pp.89-113). Minneapolis: University of Minnesota Press.

Wiggins, J.S. & Pincus, A.L. (1992). Personality: Structure and assessment. Annual Review of Psychology, 43, 473-504.