

# **Leadership Versatility Index technical report: Item selection and validation**

## *Forceful and Enabling scales*

Robert B. Kaiser and S. Bartholomew Craig

Kaplan DeVries Inc.

April 2001

Kaiser, R.B. & Craig, S.B. (2001). Leadership Versatility Index technical report: Item selection and validation. Greensboro, NC: Kaplan DeVries Inc.

KAPLAN DEVRIES INC.

Preliminary validation of the forceful and enabling leadership scales used in the Leadership Versatility Index<sup>TM</sup> (LVI) 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 using a measure including an initial set of 11 forceful and 11 enabling items rated on the “curvilinear” response scale shown below. This report includes a summary of those analyses. The net-out of our validation work to date is a refined set of five-item forceful and five-item enabling scales with adequate measurement characteristics. These items form the core of the revised LVI scales, which have been lengthened in light of the results reported below for better content coverage of the forceful and enabling domains.

**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

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

### 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 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, & Fler, 1995). All possible pairwise 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., Fecteau & 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 an equivalent metric 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 items retained after psychometric analyses*

<i>Forceful</i>	<i>Enabling</i>
1f. Strong leader	1e. Enables subordinates
2f. Declares self	2e. Receptive to others' ideas
3f. Makes tough calls	3e. Compassionate
4f. Makes judgments	4e. Shows appreciation
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 distinct and empirically unrelated. The final model tested was the one we hypothesized by the underlying duality 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).

**Table 2.**

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

Model	$\chi^2$	df	Fit Indices					RMSR	RMSEA
			CFI	GFI	AGFI	NNFI			
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 duality 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 best-fitting model.

**Table 3.**

*Factor loadings for forceful and enabling items*

Item (behavioral core of text)	Factor Loading	
	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

## 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. 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). 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*

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

*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 below .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.

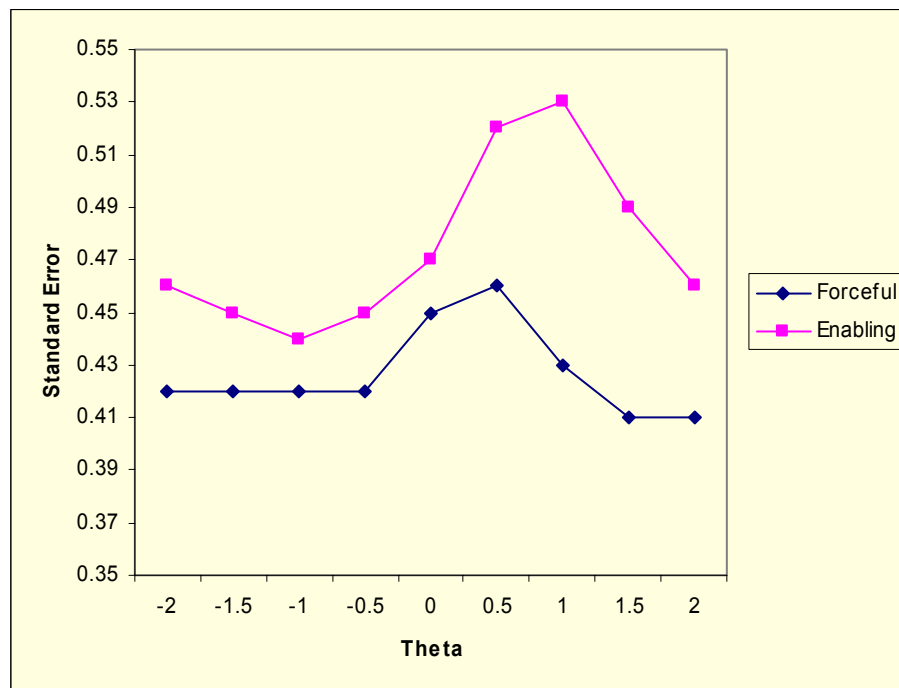
### Test Functioning as Assessed by Item Response Theory

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 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.

**Figure 2.**

*Standard error functions for the forceful and enabling scales*

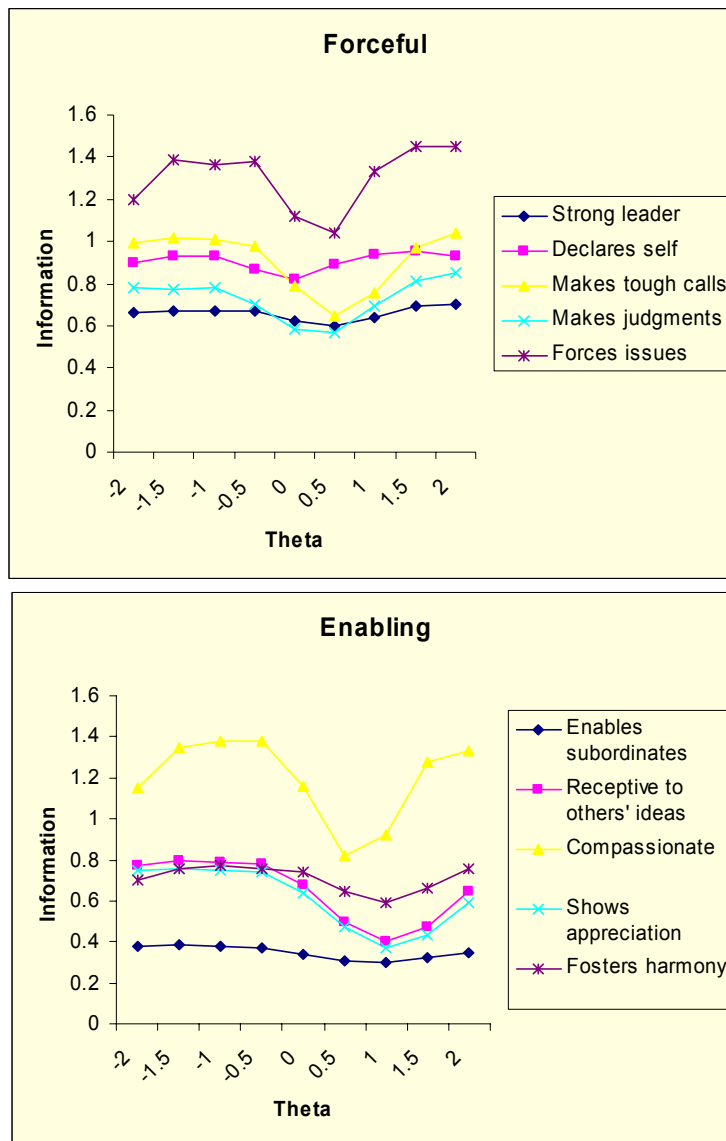


IRT is also a powerful analytic tool for understanding how much information each item in a scale provides relative to the other items. This is particularly useful when it comes time to generate additional items: items can be written to target ranges on theta at which the current test items are less precise. 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).

### Aggregating Ratings: 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*

Rating Source	Forceful scale			Enabling scale		
	ICC (single) <sup>1</sup>	ICC (mean) <sup>2</sup>	$Mr_{wg(5)}$	ICC (single) <sup>1</sup>	ICC (mean) <sup>2</sup>	$Mr_{wg(5)}$
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 <sup>1</sup>the reliability of a single rating and <sup>2</sup>for the mean of  $n$  raters (superiors  $n = 2$ , peers  $n = 3$ , subordinates  $n = 3$ ).  
 \*\*\*  $p < .001$  \*\*  $p < .01$

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.

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

<i>Rating Source</i>	<i>N</i>	<i>Forceful</i>			<i>Enabling</i>			<i>r</i> <sub>Forceful &amp; Enabling</sub>
		<i>M</i>	<i>SD</i>	<i>α</i>	<i>M</i>	<i>SD</i>	<i>α</i>	
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***

\*\*\*  $p < .001$

### 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.

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

<i>Source</i>	<i>Source</i>			
	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 ( $p < .001$ ).

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.

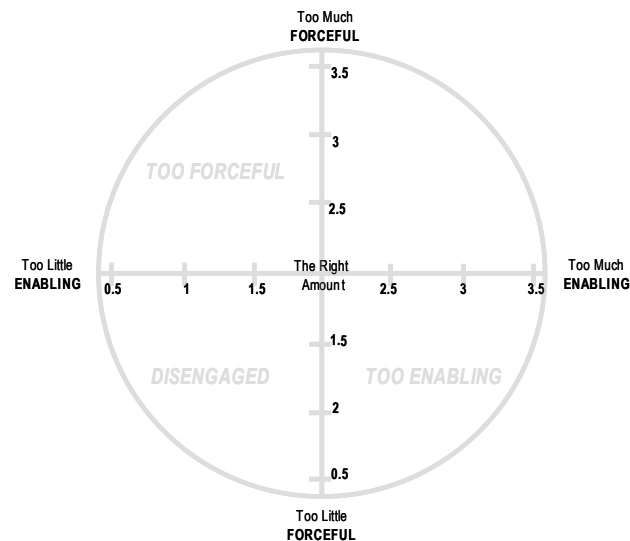
Finally, comparing these correlations between sources to the within source correlations reported in Table 5 [ICC(mean)] suggest that there is more convergence within sources than between sources. This supports the practice of reporting feedback results separately for each rating group. However, given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

#### Four Types of Leaders

More important than 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.<sup>1</sup> The observed frequency counts are reported in Table 8.

**Table 8.**  
*Frequency counts for leadership types*

<i>Leadership Style</i>	<i>Rating Source</i>							
	<i>Self</i>		<i>Superiors</i>		<i>Peers</i>		<i>Subordinates</i>	
	<i>n</i>	<i>%</i>	<i>n</i>	<i>%</i>	<i>n</i>	<i>%</i>	<i>n</i>	<i>%</i>
Versatile	23	22.1 <sup>a</sup>	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 <sup>a</sup>	21	21.4 <sup>a,b</sup>	18	18.5 <sup>b,c</sup>	13	12.2 <sup>c</sup>
Disengaged	25	24.0 <sup>a</sup>	33	33.7	29	29.9 <sup>a</sup>	50	46.7 <sup>b</sup>
<i>N</i>	104		98		97		107	

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

Chi-square analysis of the different frequencies across the cells indicated statistical significance ( $\chi^2(9) = 19.37, p < .05$ ). 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%).

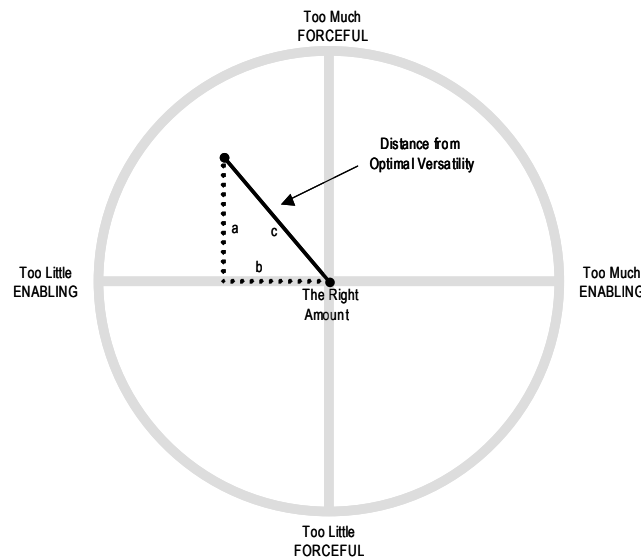
<sup>1</sup> To determine if forceful and enabling scores were not significantly different from 2.0, we used IRT to calculate scores for each rater as well as the standard error for those scores. An advantage to IRT over classical test theory approaches is that it provides unique standard error estimates 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.

## 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  $a = (\text{forceful score} - 2)$ ,  $b = (\text{enabling score} - 2)$ , and  $c = \text{distance from optimal versatility}$ .

**Figure 5.**  
*Computing the “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 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, 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.

**Table 9.**  
*Versatility index descriptive statistics*

Rating Source	<i>N</i>	<i>M</i>	<i>SD</i>	<i>skew</i>	<i>minimum</i>	<i>maximum</i>
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

### 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. Fleener 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.

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 and random samples of four peers and four subordinates—for superiors ICC[2,2] ( $n = 34$ ), for peers ICC[4,4] ( $n = 41$ ) and for subordinates ICC[4,4] ( $n = 59$ ) (see Shrout & Fleiss, 1979). Estimates for the reliability of a single rater (e.g., ICC[1,4] in the case of peer data) were also calculated. As reported in Table 2, this aggregated single item rating exceeded minimum standards of agreement and reliability (i.e., .70; Nunally, 1978; James et al. 1993).

Convergent validity evidence is demonstrated in the between-rating source correlations (all are positive and significant). Further, these measurement properties are in line with meta-analytic

estimates of ratings of middle managers on multiple-item scales (Conway & Huffcut, 1997). Descriptive statistics and statistical validity evidence for this measure is summarized in Table 10.

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

<i>Rating Source</i>	<i>N</i>	<i>M</i>	<i>SD</i>	ICC (single) <sup>1</sup>	ICC (mean) <sup>2</sup>	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 <sup>1</sup>the reliability of a single rating and <sup>2</sup>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.

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

Rating Source	Forceful				Enabling			
	$\beta_0$	$\beta_1$	$\beta_2$	$R^2$	$\beta_0$	$\beta_1$	$\beta_2$	$R^2$
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:  $\beta_0$  = intercept,  $\beta_1$  = beta weight for predictor,  $\beta_2$  = beta weight for predictor-squared.

\*\*\*  $p < .001$  \*\*  $p < .01$  \*  $p < .05$

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

	Forceful				Enabling			
	$\beta_0$	$\beta_1$	$\beta_2$	$R^2$	$\beta_0$	$\beta_1$	$\beta_2$	$R^2$
	-4.66	2.63**	-2.47**	.14**	8.90	-.38	0.54	.03

Note:  $\beta_0$  = constant,  $\beta_1$  = beta weight for predictor,  $\beta_2$  = 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 effectiveness 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.

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

<i>Source of Versatility Index</i>	<i>Source of Effectiveness Rating</i>				
	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:* \*\*\*  $p < .001$  \*\*  $p < .01$  \*  $p < .05$

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 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.

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

<i>Source</i>	<i>Rating Source</i>															
	Self				Superiors				Peers				Subordinates			
	<i>df</i>	<i>MS</i>	<i>F</i>	<i>eta</i> <sup>2</sup>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>eta</i> <sup>2</sup>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>eta</i> <sup>2</sup>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>eta</i> <sup>2</sup>
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:* \*\*\*  $p < .01$  \*\*  $p < .05$  \*  $p < .10$

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 shown 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, 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 per cent 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 fit for the observed ratings on the ten items. The scales demonstrate acceptable levels of internal consistency reliability. And as a 360° 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° 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.

## References

- Bass, B.M. (1990). Bass and Stogdill's handbook of leadership: Theory, research, and managerial applications (3rd Ed.). New York: Free Press.
- 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, 215-227.
- 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. & 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.