

Validity of Observer Ratings of the Big Five Personality Factors

Michael K. Mount, Murray R. Barrick, and J. Perkins Strauss

The authors examined the validity of observer ratings (supervisor, coworker, and customer) and self-ratings of personality measures. Results based on a sample of 105 sales representatives supported the 2 hypotheses tested. First, supervisor, coworker, and customer ratings of the 2 job-relevant personality dimensions—conscientiousness and extraversion—were valid predictors of performance ratings, and the magnitude of the validities were at least as large as for self-ratings. Second, supervisor, coworker, and customer ratings accounted for significant variance in the criterion measure beyond self-ratings alone for the relevant dimensions. Overall, the results suggest that validities of personality measures based on self-assessments alone may underestimate the true validity of personality constructs.

In the past 10 years, the views of many personality psychologists have converged regarding the structure and concepts of personality. Generally, researchers agree that there are five robust factors of personality that can serve as a meaningful taxonomy for classifying personality attributes (Digman, 1990). This taxonomy has consistently emerged in longitudinal studies; across different sources (e.g., ratings by self, spouse, acquaintances, and friends); with numerous personality inventories and theoretical systems; and in different age, sex, race, and language groups. It also has some biological basis, as suggested by evidence of heritability (e.g., Costa & McCrae, 1992; Digman, 1990).

Although the names for these factors differ across researchers, the following labels and prototypical characteristics are representative: (a) extraversion (sociable, talkative, assertive, ambitious, and active), (b) agreeableness (good-natured, cooperative, and trusting), (c) conscientiousness (responsible, dependable, able to plan, organized, persistent, and achievement oriented), (d) emotional stability (calm, secure, and not nervous), and (e) openness to experience (imaginative, artistically sensitive, and intellectual).

The emergence of the five-factor model has enabled researchers to conduct construct-oriented meta-analytic reviews of the predictive validity of personality (Barrick & Mount, 1991; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Tett, Jackson, & Rothstein, 1991). Although these reviews have adopted slightly different personality frameworks, the conclusions can be summarized in terms of the Big Five taxonomy. The Barrick and Mount (1991) and Hough et al. (1990) reviews demonstrated that only one dimension of the Big Five, conscientiousness (achievement and dependability in the Hough et al. frame-

work), is a valid predictor for all occupational groups and all job-related criteria studied. Other dimensions were valid predictors for only some criteria or some occupations. Additional support for this conclusion has been provided by results reported in the U.S. Army Selection and Classification Study (Project A; McHenry, Hough, Toquam, Hanson, & Ashworth, 1990). These authors found that components of conscientiousness (i.e., achievement and dependability) were the best personality predictors of targeted criteria. In contrast, conscientiousness was not the most valid predictor of job performance in a Big Five meta-analysis by Tett et al. (1991). However, as pointed out elsewhere (Ones, Mount, Barrick, & Hunter, in press), the discrepancies may be explained by differences in methodological and statistical approaches used in the study.

The Tett et al. (1991) results notwithstanding, the preponderance of evidence shows that individuals who are dependable, reliable, careful, thorough, able to plan, organized, hardworking, persistent, and achievement oriented tend to have higher job performance in most if not all occupations. Conscientiousness has emerged as perhaps the most important trait motivation variable in personnel psychology (Barrick, Mount, & Strauss, 1993; Schmidt & Hunter, 1992).

The meta-analysis by Barrick and Mount (1991) also revealed that one other personality dimension, extraversion, is a valid predictor of job performance for the sample assessed in this study (sales representatives) as well as for managers. That is, in jobs with a large social component, such as sales and management, Barrick and Mount's results suggest that being sociable, talkative, assertive, ambitious, and active will lead to higher job success.

As this suggests, both conscientiousness and extraversion are relevant personality attributes for the occupation assessed in this study. However, the predictive validity of these personality dimensions is based almost exclusively on a single method of measurement: individual self-assessments. Relatively little is known about the validity of personality constructs when assessed by raters other than the individual, particularly in employment contexts. Therefore, there were two major purposes for this study. The first was to examine the magnitude of validities for these two personality dimensions—conscientiousness and extraversion—when they are assessed by observers (super-

Michael K. Mount and Murray R. Barrick, Department of Management and Organizations, University of Iowa; J. Perkins Strauss, Department of Business Administration, Augustana College.

We thank two anonymous reviewers for their many helpful comments on an earlier version of the article.

Correspondence concerning this article should be addressed to Michael K. Mount, Department of Management and Organizations, University of Iowa, 5380 Pappajohn Building, Iowa City, Iowa 52242-1323.

visors, coworkers, and customers). The second was to examine whether observer ratings explain performance variability over that accounted for by self-ratings.

Self-Ratings Versus Other Ratings

Hogan (1991) pointed out that there is a fundamental difference between self- and others' perspectives of a person's personality characteristics. From the observer's perspective, *personality* refers to a person's public self or social reputation (i.e., the way he or she is perceived by others, such as supervisors, coworkers, customers, friends, and family members). However, from the individual's perspective, *personality* refers to the structures, dynamics, and processes inside a person that explain why he or she behaves in a particular way. As this suggests, ratings obtained from these two perspectives are quite different: One set of ratings is based on the observer's perspective, which incorporates information about one's reputation, whereas the other is based on a self-perspective, which incorporates less observable information about motives, intentions, feelings, and past behaviors.

According to R. Hogan (1991), a person's social reputation may be the most appropriate perspective when the goal is prediction, as in personnel selection. Because past behavior is perhaps the best predictor of future behavior (Wernimont & Campbell, 1968), reputations (which are operationalized in trait terms based on past behavior) should be valid predictors of future behavior (job performance). This is particularly true for those observers who interact almost exclusively with the individual in the work setting. On the basis of this reasoning, observer ratings, which capture one's public self or social reputation at work, would be expected to predict job performance as well (or even better) than ratings based on the individual's perspective, which incorporates self-observations of past behaviors across settings. To our knowledge, this proposition has not yet been tested in the personnel selection literature.

There is some evidence that self-ratings of personality have lower correlations with measures of academic achievement as the criterion than personality ratings obtained from other sources. For example, Hough et al. (1990) conducted a comprehensive literature review of correlations between self-reports on dependability and achievement (components of conscientiousness) and education (i.e., grade point average, or GPA) of high school and college students. Their results indicated an uncorrected correlation of .15 for dependability and .30 for achievement, with a weighted average of .23. In contrast, other studies have shown that the correlations between ratings made by others on conscientiousness and measures of academic achievement are relatively high. Smith (1967) found that college students' scores on the conscientiousness dimension, as rated by peers in the first 9 weeks of classes (assessed before midterm examination), correlated .43 (uncorrected) with first-year grades. Digman (1972) reported correlations in the .50s (uncorrected) between ratings by elementary school teachers on the dimension and high school GPA. Furthermore, Digman found that a composite formed as an unweighted sum of ratings made by elementary school teachers on the conscientiousness dimension correlated .70 with high school GPA. In another study, Takemoto (1979) found a correlation of .65 (uncorrected) be-

tween ratings by eighth-grade teachers on conscientiousness and high school GPA. Overall, these findings suggest that others' ratings of conscientiousness are valid predictors of a variety of academic success criteria.

Other evidence in the personality literature also suggests that observers' ratings of personality predict behavior as well as, if not better than, self-reports. The literature on objective self-awareness demonstrates that observers' judgments of personality have greater predictive validity than do self-ratings of personality about the level of awareness of one's own aggressive behavior or affective reactions (Scheier, Buss, & Buss, 1976). John and Robbins (1991) found that the other participants in a group discussion ranked each actor's contribution to the group more accurately (in comparison with highly reliable criterion rankings by psychologists) than did the actors themselves. Furthermore, Funder, Kolar, and Colvin (1992) reported that close acquaintances predicted interpersonal behaviors as well as if not better than self-reports. Their results showed that personality judgments on the Big Five by close acquaintances were more predictive of four independently evaluated classes of behaviors coded from videotaped interpersonal interactions than were self-descriptions of personality for 140 undergraduate students (each subject had ratings from two friends or roommates).

Other empirical research has shown that self-ratings of personality have rather low correlations with ratings obtained from other sources (e.g., spouses or friends): Uncorrected correlations ranged from the high .20s to .30s (Funder & Colvin, 1988; Funder & Dobroth, 1987; McCrae, 1982; Watson, 1989). Three studies (Funder & Colvin, 1988; Funder & Dobroth, 1987; Watson, 1989) indicated that agreement between observers' ratings was greater than the agreement between self-ratings and observer ratings, with correlations ranging from .30s to .40s (uncorrected). In summary, this research shows that individuals have different views of their own personality than others do and, furthermore, that others' views of personality may be more predictive of behavior than self-reports.

Very little is known about the validity of observer ratings of personality measures in the employment context. However, given the literature cited above, it is likely that observer ratings of job-relevant constructs will be valid predictors of job performance. Two hypotheses were tested in this study. First, we hypothesized that supervisor, coworker, and customer ratings of two job-relevant dimensions—conscientiousness and extraversion—would be valid predictors of sales representatives' performance. (We also examined the validity of agreeableness, emotional stability, and openness to experience when ratings were provided by observers as well as by the sales representative themselves, but no hypotheses were tested.) Second, we examined whether observer ratings account for significant incremental variance in performance ratings over self-ratings. We hypothesized that for the two job-relevant dimensions, conscientiousness and extraversion, observer ratings would account for significant incremental variance in performance over self-ratings. This was based on previous research showing that observer ratings will be valid predictors of performance and that the correlations of observer ratings with self-ratings are relatively low. Although no specific hypotheses were tested, we also examined this for the three other personality dimensions.

Method

Subjects

Subjects were 105 sales representatives from a large appliance-manufacturing organization. They were primarily men (85%), with an average age of 34 years, organizational tenure of 7 years, and job tenure of 4 years. Each sales representative completed a self-rating on a personality questionnaire and also selected other individuals in the work setting to complete the questionnaire (generally, the supervisor, plus five coworkers and five customers). It should be noted that 13 of the subjects were not able to obtain ratings from customers; therefore, the sample size was 92 rather than 105 for all analyses with customers. The average number of years the sales representatives had known their raters was as follows: for coworkers, $M = 2.59$, $SD = 0.88$; for supervisors, $M = 2.54$, $SD = 1.29$; for customers, $M = 2.40$, $SD = 0.97$.

The purpose for obtaining these personality ratings from the various sources was to give developmental feedback to the sales representatives. Performance ratings were obtained from both the supervisor and the coworkers.

Measures

Personality. Each participant completed a shortened version of the personality inventory developed by Goldberg (1992). This personality inventory was developed to provide a set of Big Five factor markers that could replace those developed more than 30 years ago by Norman (1963). On the basis of responses obtained from 867 subjects and 205 peers, Goldberg identified 20 unipolar trait adjective variables for each dimension of the Big Five. In a follow-up study, 175 students completed the Goldberg inventory and two other measures of the Big Five: the NEO Personality Inventory (NEO-PI; Costa & McCrae, 1985) and the Hogan Personality Inventory (R. Hogan, 1986). Correlations among similar personality constructs of the Goldberg inventory and the NEO-PI were .69, .56, .67, .69, and .46, and correlations with the Hogan Personality Inventory were .56, .52, .56, .62, and .39 for extraversion, agreeableness, conscientiousness, emotional stability, and openness to experience, respectively. With both inventories, correlations with dissimilar constructs were considerably lower, ranging from .00 to .32.

Because of time constraints imposed by the organization, we shortened the inventory from 100 to 50 adjectives. Items were selected on the basis of the magnitude of the factor loadings reported by Goldberg (1992); the 10 items with the largest factor loadings were retained for each Big Five personality dimension. Examples of adjectives used were conscientiousness: organized, systematic, thorough, hardworking, careless, inefficient, and sloppy; extraversion: extroverted, talkative, assertive, reserved, introverted, and quiet; agreeableness: sympathetic, cooperative, trustful, cold, rude, and unkind; emotional stability: unenvious, relaxed, calm, moody, touchy, and nervous; and openness to experience: intellectual, creative, artistic, unimaginative, conventional, and simple. For the five factors, coefficient alphas were .75, .73, .79, .73, and .75, respectively, for self-ratings; .83, .84, .86, .73, and .71, respectively, for supervisor ratings; .73, .81, .70, .67, and .71, respectively, for coworker ratings; and .78, .74, .85, .71, and .70, respectively, for customer ratings.

We obtained evidence from 198 undergraduate business students to support the construct validity of the shortened scales used in this study. The students responded to both the Personal Characteristics Inventory (PCI) and the 100-item Goldberg personality inventory. (A thorough description of the PCI is reported in Barrick & Mount, 1993; Barrick et al., 1993.) First, we computed the correlations between the 10 items from the Goldberg inventory used in the present study and the 10 items that were not used. Correlations for the five factors were .78, .79, .76, .75, and .73, for conscientiousness, extraversion, agreeableness, emotional stability, and openness to experience, respectively. The corre-

lations (uncorrected) between the shortened Goldberg questionnaire and similar constructs on the PCI were .61, .66, .60, .64, and .69, respectively. Correlations between the 100-item Goldberg inventory and the PCI were .71, .69, .66, .71, and .61 for conscientiousness, extraversion, agreeableness, emotional stability, and openness to experience, respectively. In both analyses, correlations across dissimilar constructs were much lower. With the shortened version of the Goldberg questionnaire, the correlations across dissimilar constructs ranged from .36 to $-.07$ ($M = .11$). With the 100-item version of the Goldberg questionnaire, the correlations ranged from .39 to $-.03$ ($M = .12$). Overall, these results provide evidence of the construct validity of the shortened Goldberg questionnaire.

In completing the inventory, all sales representatives rated the extent to which the unipolar adjectives were representative of themselves. As mentioned, in addition to the self-ratings, the inventory was also completed by one supervisor and up to five coworkers and five customers, who rated the extent to which the adjectives were descriptive of the sales representative. The response scale ranged from 1 (*strongly agree*) to 5 (*strongly disagree*). High scores represented high levels of conscientiousness, extraversion, agreeableness, emotional stability, and openness to experience.

Job-performance ratings. A nine-dimensional measure of job performance was developed by the researchers on the basis of an analysis of the sales job. The dimensions were job knowledge, quality of work, quantity of work, initiative, customer communications, account management, interpersonal skills, commitment to job, and job attitude. Each dimension was defined by a one-sentence description, followed by three or four interpretative examples illustrating important facets of that dimension. The subjects' supervisors and coworkers rated the sales representatives' performance on a 5-point Likert-type scale ranging from *consistently exceeds job requirements* (1) to *somewhat below job requirements* (5). Raters were informed that ratings were being collected for research purposes. Overall performance was the sum of the ratings across all dimensions. The coefficient alphas were .89 for the supervisors and .94 for the coworkers.

Analysis. Scores on each of the five personality dimensions were obtained by averaging the ratings on the traits for each dimension. Scores on the performance measure were obtained by averaging the supervisor's rating on the nine performance dimensions. Validities were calculated for the sales representatives, supervisors, coworkers, and customers. Our interest was in comparing the magnitude of the validities obtained for self-ratings versus those for the other rating sources. Although data from up to five coworkers and customers were available for each sales representative, the validities were based on personality ratings from only one randomly selected coworker and one randomly selected customer. Averaging all possible coworker or customer ratings would have resulted in higher predictor reliability. This, in turn, could confound the comparison with self-ratings because higher validities could be attributed to either the higher reliability of the personality constructs (on the basis of average ratings) or the effects of different perspectives. (It should be noted that the results based on averages across all coworkers or customers were comparable, although slightly larger than those reported in this study, and are available on request.)

We also report the correlations for each perspective, using coworker performance ratings as the criterion. Such ratings are not traditionally used as the criterion in selection settings; however, their use in this study allows us to assess the generalizability of the relations found across two criteria. Analyses reported using coworker ratings as the criterion are based on the average of all possible coworker ratings (after excluding the coworker who provided the predictor ratings) for each sales representative.

Results

The means and standard deviations for the personality dimensions for the four rating perspectives and the performance

Table 1
Means, Standard Deviations, and Correlations Between the Big Five Factors and Performance Ratings by Rater Source

Rating source	n	M	SD	Performance rating			
				Supervisor		Coworker ^a	
			r_{xy}	ρ	r_{xy}	ρ	
Conscientiousness							
Self	105	4.18 _b	0.48	.18*	.26	.17*	.23
Supervisor	105	3.97 _a	0.54	.45**	.64	.25**	.34
Coworker	105	4.10	0.66	.26**	.37	.24**	.32
Customer	92	4.15	0.41	.30**	.42	.24**	.32
Extraversion							
Self	105	3.87 _b	0.46	.06	.09	.12	.16
Supervisor	105	3.65 _a	0.53	.26**	.37	.24**	.32
Coworker	105	3.83	0.58	.24**	.34	.24**	.32
Customer	92	3.86 _b	0.50	.27**	.38	.21*	.28
Agreeableness							
Self	105	4.43 _b	0.40	.05	.07	.04	.05
Supervisor	105	4.13 _a	0.47	.34**	.48	.15	.20
Coworker	105	4.22 _a	0.54	.12	.17	.15	.20
Customer	92	4.40 _b	0.40	.30**	.42	.34**	.46
Emotional stability							
Self	105	3.44	0.55	.05	.07	.06	.08
Supervisor	105	3.27	0.50	.16	.23	.12	.16
Coworker	105	3.31	0.55	.08	.11	.04	.05
Customer	92	3.42	0.46	.09	.13	.02	.03
Openness to experience							
Self	105	3.69 _a	0.51	.09	.13	.08	.11
Supervisor	105	3.48 _b	0.46	.20*	.28	.09	.12
Coworker	105	3.59	0.49	.23*	.33	.10	.13
Customer	92	3.62	0.42	.15	.21	.15	.20
Performance ratings							
Supervisor	105	3.89 _a	0.60				
Coworker	105	4.13 _b	0.41				

Note. Validities based on personality and performance ratings provided by raters from the same source are in boldface. Means with different subscripts are statistically different. r_{xy} = observed validity coefficient; ρ = validity coefficient corrected for attenuation in the criterion.

^a This performance criterion was based on an average of 1.6 coworker responses per sales representative.

* $p < .05$. ** $p < .01$.

ratings are shown in Table 1. As previously noted, the sample size for the customers is smaller than for the other perspectives. Although all sales representatives rated themselves and were rated by the supervisor and at least one coworker, 13 respondents were not able to obtain customer ratings. For each of the personality dimensions, we examined whether there were significant differences among rating sources by conducting a one-way analysis of variance, followed by Tukey's honestly significant difference test. The mean ratings were significantly different for each of the five personality dimensions across rater

sources. Those means that were significantly different are denoted with subscripts in Table 1. For conscientiousness, $F(3, 403) = 3.13, p < .05$, with self-ratings greater than supervisors' ratings ($\omega^2 = .02$). For extraversion, $F(3, 403) = 4.08, p < .01$, with self-ratings and customer ratings greater than supervisors' ratings ($\omega^2 = .02$). For agreeableness, $F(3, 403) = 10.11, p < .01$, with the following comparisons significantly different: self- and customer ratings greater than supervisors', and self- and customer ratings greater than coworkers' ($\omega^2 = .07$). For emotional stability, $F(3, 403) = 2.61, p < .05$, but none of the contrasts were significantly different ($\omega^2 = .02$). For openness to experience, $F(3, 403) = 3.55, p < .01$, with self-ratings greater than supervisors' ratings ($\omega^2 = .03$). The most consistent finding from this analysis is that self-ratings were significantly higher than supervisor ratings. Overall, however, the omega-square values show that the proportion of variance in the ratings of personality attributable to rating sources is quite small for all dimensions. With respect to the two sets of criterion ratings, coworker performance ratings were significantly higher than supervisor ratings, $F(1, 208) = 11.38, p < .01$ ($\omega^2 = .05$).

The corrected (ρ) and uncorrected (r_{xy}) correlations between the Big Five personality scales for the four rating sources and the supervisor and coworker performance ratings are also shown in Table 1. There was only one supervisor rating of performance for each sales representative. Therefore, we corrected the validities (ρ) for unreliability in the criterion by using the average single-rater reliability of .50 obtained by Rothstein (1990), which was based on 9,975 first-line supervisors. The true validities when using coworker performance ratings as the criterion were corrected based on the correlation between two randomly selected coworkers' performance ratings for each sales representative. On the basis of 105 pairs of performance ratings, the reliability of a single coworker's ratings was .53. To avoid problems associated with common method variance for coworker ratings, we randomly selected one coworker's ratings as the predictor and used the average of the remaining coworkers' ratings as the criterion measure. Because there was an average of 1.6 coworkers for each sales representative, the Spearman-Brown prophecy formula was used to adjust the reliability upward. Consequently, we used .55 as the reliability of the composite of the coworkers' performance ratings.

The validities for the two job-relevant predictors, conscientiousness and extraversion, are presented first. As shown in Table 1, all validity coefficients for conscientiousness for self- and observer perspectives were statistically significant for both criterion measures ($p < .05$, one-tailed tests). True validities (ρ) for the supervisor criterion ranged from .26 for self-ratings to .64 for supervisor ratings. For the coworker criterion, true validities ranged from .23 for self-ratings to .34 for supervisor ratings. For extraversion, the correlations based on self-ratings were not statistically significant ($\rho_s = .09$ and .16 for the two criteria), whereas all those based on observer ratings were significant. For the supervisor criterion, ρ true validities ranged from .34 for coworker ratings to .38 for customer ratings. For the coworker criterion, ρ true validities ranged from .28 for customer ratings to .32 for coworker and supervisor ratings.

In contrast, few of the validities for the other personality dimensions were significant. None of the self-ratings for these non-job-related personality dimensions were significant pre-

dictors (ρ s ranged from .05 to .13) for either criterion type. Two significant validities (for agreeableness and openness to experience) occurred when the supervisor provided both predictor and criterion information. (Validities based on predictor and criterion data provided by the same person are identified in boldface in the table by the underlined coefficients.) Of the remaining observer-based validities for the three non-job-related predictors, only three were statistically significant. For agreeableness, customer ratings were significant predictors for both supervisor and coworker ratings of performance (ρ s = .42 and .46, respectively). For openness to experience, coworker ratings were significantly correlated with supervisor ratings of performance (ρ = .33).

As suggested by a reviewer, an argument could be made to correct the validities between personality dimensions and performance for unreliability in the predictors. This correction is not traditionally made, but it may be appropriate here because, at a theoretical level, our intention was to examine the relations between the Big Five constructs and performance and because we knowingly used an imperfect measure of the Big Five (e.g., the Goldberg questionnaire was only half of its original length). Therefore, to determine the true validities, we corrected the correlations for each predictor dimension for each rating source, using the reliabilities reported earlier. The resulting true validities are corrected for unreliability in both the predictor and criterion. For example, the true validity shown in Table 1 for self-ratings of conscientiousness is .26, but it would be .30 if corrected for predictor unreliability. The true validity for customer ratings of conscientiousness and agreeableness using supervisor ratings is .42, but it would be .48 if corrected for predictor unreliability. Overall, the true validities were approximately 16% higher than those reported in Table 1 (which are corrected for unreliability in the criterion only).

Our second purpose was to examine the incremental validity of observer ratings over self-ratings. We conducted hierarchical regression analyses, using supervisor performance ratings as the criterion. (Results from using coworker performance ratings as the criterion are available on request.) Self-ratings were entered in the first step; in the second step, each observer source was entered separately to assess the incremental validity of each source over self-ratings alone; in the third step, all rating sources (self, supervisor, coworker, and customer) were entered together as a block to determine the percentage of variance accounted for by all sources.

The regression results are presented in Table 2. Results for conscientiousness are considered first. As shown, each rating source accounted for significant variance in performance ratings beyond that accounted for by self-ratings alone ($p < .05$ in each case). Considering all sources together (Step 3), observer ratings of conscientiousness account for an additional 21% of the variance in performance beyond that accounted for by self-ratings alone ($p < .05$).

Results for extraversion were similar to those for conscientiousness. Each rater perspective accounted for a significant amount of performance variability beyond self-ratings alone ($p < .05$ in each case). The analysis in which we used all observer perspectives indicated that the three rating perspectives accounted for an additional 11% of variance beyond self-ratings ($p < .05$).

Table 2
Results of Hierarchical Regression Analyses
for the Big Five Factors

Rating source	Supervisor performance ratings			
	Total ΔR^2	Adjusted R^2	ΔR^2	p
Conscientiousness				
Self	.035		.035	.06
Observers				
Supervisor	.205		.170	.01
Coworker	.083		.048	.02
Customer	.115		.080	.01
All sources	.243	.205		.01
Extraversion				
Self	.004		.004	.55
Observers				
Supervisor	.062		.058	.01
Coworker	.058		.054	.02
Customer	.078		.074	.01
All sources	.114	.072		.04
Agreeableness				
Self	.003		.003	.59
Observers				
Supervisor	.131		.128	.01
Coworker	.015		.012	.28
Customer	.103		.100	.01
All sources	.181	.141		.01
Emotional stability				
Self	.002		.002	.63
Observers				
Supervisor	.030		.028	.11
Coworker	.007		.005	.48
Customer	.008		.006	.43
All sources	.024	.000		.73
Openness to experience				
Self	.007		.007	.40
Observers				
Supervisor	.062		.055	.02
Coworker	.075		.068	.01
Customer	.033		.026	.14
All sources	.097	.053		.07

Note. For observers, each rating source was entered individually after controlling for self-ratings (i.e., three separate regressions were run for each personality dimension).

Results for agreeableness showed that supervisor and customer ratings accounted for significant variance beyond self-ratings ($p < .05$ in each case). When all observer perspectives were used, they accounted for an additional 18% of variance beyond self-ratings ($p < .05$).

None of the perspectives accounted for significant variability in performance for emotional stability. For openness to experience, two perspectives—supervisor and coworker ratings—accounted for significant performance variability beyond self-ratings ($p < .05$). When all observer ratings of openness to experience were used together, they accounted for an additional 9% of

variance beyond self-ratings, although this was only marginally significant ($p = .07$).

We also examined the correlation matrix of the five personality dimensions by four rating sources. (This 20×20 correlation matrix is available on request.) We cast this in terms of a multitrait, multimethod framework, and three types of information were yielded: the correlation between different rating sources when rating the same personality dimension (convergent validity); the correlation between ratings obtained from the same source when rating different dimensions (common method variance); and the correlation between ratings from different sources rating different dimensions (discriminant validity).

The convergent validities based on the average correlation of the six possible pairs of raters rating the same personality dimension were .40 for extraversion, .31 for conscientiousness, .29 for agreeableness, .28 for emotional stability, and .24 for openness to experience. The average across the five dimensions was .30. The common method correlations, based on the average correlation within each of the four sources across traits, was .14 for supervisor ratings, .21 for self-ratings, .37 for coworkers, and .41 for customers. The average across the four sources was .28. The average discriminant validity based on all possible pairs of raters rating different personality dimensions was .03, and the validities ranged from $-.34$ to .35.

Ideally, the convergent validities should have been higher than those for common method correlations and discriminant validities. The present results show that the discriminant validities were near zero, as desired; however, the convergent validities were only slightly larger than the common method correlations. This indicates that the degree of agreement between two raters from different perspectives when rating the same personality construct was relatively low because the correlation was greater than .30 only for conscientiousness and extraversion. The common method correlations showed that the degree of common method variance was relatively high, particularly for coworker and customer ratings.

Discussion

Relatively little is known about the validity of personality measures when they are assessed from perspectives other than individuals' self-ratings. We hypothesized that job-relevant personality constructs would be valid predictors when assessed from the perspective of observers and that observer ratings on these personality dimensions would account for significant variance beyond self-ratings alone. The results confirmed both of our hypotheses.

The magnitudes of the zero-order correlations for self-ratings of conscientiousness were .26 and .23 (corrected) for the two criteria, which is very similar to the value reported for conscientiousness for sales representatives (.23) by Barrick and Mount (1991). None of the validities based on self-assessments for the other four personality constructs were statistically significant, which is also similar to the results reported by Barrick and Mount (where the 90% credibility value included zero for the four constructs).

The results illustrate that observer ratings of the job-relevant constructs, conscientiousness and extraversion, are valid pre-

dictors of performance. For these two constructs, the magnitude of the validities was at least as high as for self-assessments. For conscientiousness, corrected correlations ranged from .32 to .42 for the two criteria (excluding those based on common methods); for extraversion, they ranged from .28 to .38. The findings for extraversion are particularly striking because results based on self-reports indicated that extraversion was not a valid predictor of performance for this job, whereas observer ratings indicate that it was for both supervisor and coworker performance ratings.

The results for the agreeableness dimension provided evidence that one observer source, customer ratings, was a significant predictor of performance. The corrected correlations were .42 and .46 for supervisor and coworker performance ratings, respectively. These are very similar or slightly higher than those for conscientiousness and extraversion, both of which were believed to be more job relevant.

Overall, the zero-order correlations show that observer ratings of relevant personality constructs (conscientiousness and extraversion) are valid predictors of performance, that their validities are at least as high as for self-assessments, and that their validities differ according to rater perspective. Furthermore, customer ratings of agreeableness were among the most valid.

Our second major purpose was to examine whether observer ratings accounted for significant variance in performance ratings beyond self-ratings. As predicted, each observer perspective accounted for significant variance beyond self-ratings for both conscientiousness and extraversion. In addition, for agreeableness, supervisor and customer ratings accounted for significant variance, and for openness to experience, supervisor and coworker ratings accounted for significant variance. Overall, these results show that observer ratings are valid predictors of performance ratings and that they account for significant variability beyond self-ratings, particularly for job-relevant dimensions.

Taken together, these results underscore the distinction between the validity of personality constructs per se and the validity of the constructs based on a particular method of measurement. Nearly all empirical findings obtained in prior research reflect the validity of personality constructs based on self-reports. Conclusions about the validity of personality measures may be a function of the perspective of the rater-self versus observer. Our results demonstrate that job-relevant personality constructs are valid predictors when they are assessed from the observers' perspective but that they are less so when assessed from the individual's perspective.

As discussed, Hogan (1991) distinguished between personality from the perspective of self and personality from the observer's perspective. From the observer's perspective, *personality* refers to a person's public self or social reputation; it reflects the way that he or she is perceived by others. From the perspective of the self, *personality* refers to an individual's perceptions of himself or herself in numerous situations that may incorporate less observable information about motives, intentions, feelings, and past behaviors (Hogan, 1991).

Thus, one explanation for our results is that the frame of reference of observers is different from that of individuals when rating themselves because observers' assessments are based on job-related observations. That is, observers' assessments of an

individual's personality may be more valid predictors than self-assessments because they are based on observations of the individual almost exclusively in the work environment. On the other hand, individuals see themselves in numerous situations, such as at home, at church, at play, and at work. Consequently, observer ratings may be better predictors than self-ratings because there is greater point-to-point correspondence between the predictor and the criterion.

One way to test this frame of reference hypothesis would be to instruct individuals to rate their personality only as they see themselves at work. If the hypothesis is true, then the resulting validities would be higher than for the traditional self-assessment approach, in which no frame of reference is specified.

A second explanation for the relatively high validities for observer ratings is possible distortion in the ratings. Both impression management, whereby individuals intentionally distort information to create a favorable impression, and self-deception, whereby individuals unconsciously present themselves in a favorable light, may account for the differing validities for the rating sources (Paulhus, 1986). With respect to self-ratings, there is evidence that individuals can fake their scores on personality inventories, but this does not appear to affect the validities (Hough et al., 1990). Although less is known about distortion in observer ratings, the possibility exists here as well. For example, in the present study, customer ratings of agreeableness were valid predictors, whereas coworker ratings were not. It is reasonable to assume that the sales representatives are highly motivated to present themselves in a favorable light to their customers but are less motivated to do so with coworkers. Because impression management is likely to be characteristic of effective sales representatives, these individuals will also tend to be better performers. However, sales representatives are less likely to put on an act for coworkers; therefore, high performers are no more likely to be perceived as agreeable than are low performers, hence the lower validity for coworker ratings.

Use of Observer Ratings

Our results raise several issues regarding the use of alternative rating sources to assess personality attributes. Perhaps the most important of these is whether it is feasible to use such ratings for external selection purposes. One common external selection practice in most firms is reference checking; estimates of use range as high as 95% of firms, according to some surveys (Lilenthal, 1980). Traditionally, reference checking entails collecting information about prospective job applicants from people who have contact with the job applicants, typically to verify information about the applicant or as the basis for making predictions about job success. In this sense, observer ratings are already an integral part of the external selection process.

The issue then becomes how to best use the observer ratings in the reference check. One approach would be to develop the reference check information on the basis of relevant Big Five personality dimensions as determined through a job analysis. The job candidate would then nominate others (supervisors, customers, coworkers, and direct reports) from previous work situations to complete the forms. Such an approach would have the advantage of being more job relevant, systematic, and theoretically based than most approaches to reference checking. An

issue that would need to be addressed is what the validities of the observer ratings are when the purpose is administrative rather than developmental, as was the case in the present study. Similarly, from the perspective of internal staffing, there are a number of advantages to using alternative rating sources. It would take relatively little time and cost very little, and the resulting validity is likely to be relatively high if coworkers and supervisors in the organization, as well as current customers, were to evaluate the sales representatives' personality traits. Our results indicate that ratings based on these sources have validities as high as, if not higher than, self-ratings; a composite of these ratings accounts for 11% and 20% more variance in performance (on average) than self-ratings alone for the two job-relevant predictors.

The results for the customer perspectives were particularly intriguing in view of the relatively high validities (e.g., $\rho = .42$ for conscientiousness, $\rho = .38$ for extraversion, and $\rho = .42$ for agreeableness) we found when using supervisor performance ratings as the criterion. These were as high as validities for cognitive ability tests, work samples, and biodata cited by Schmitt, Gooding, Noe, and Kirsch (1984) (corrected only for sampling error), although not as high as validities for ability tests reported by Hunter and Hunter (1984; corrected for sampling error, range restriction, and unreliability in the criterion). From the customer's perspective, sales representatives who are highly conscientious, extroverted, and agreeable are better performers than those who are low on these dimensions. Ratings made by customers should be of great importance to sales representatives because customers are ultimately the key to their success as well as to the organization's success. This suggests that organizations should provide feedback to sales representatives about customer perceptions on these three personality dimensions.

It is becoming more common to use multiple raters in performance appraisal systems; such 360° feedback programs typically include peers, subordinates, supervisors, and self-ratings as sources of performance feedback. Results of the present study suggest that multiple raters of personality (particularly customers) can provide useful developmental feedback for sales representatives. The value of the 360° approach is underscored by the fact that the convergent validities in the present study were relatively low, ranging from .24 to .40 for the five dimensions. These values correspond to the convergent validities reported in previous studies (e.g., Funder & Colvin, 1988; Funder & Dobroth, 1987; McCrae, 1982; Shrauger & Osberg, 1981; Watson, 1989) in which uncorrected correlations generally ranged from the high .20s to .30s. Given that the reliability of a single rater's rating of personality attributes is low, it is not surprising that validities of personality measures are also relatively low. Aggregating raters will result in more reliable measures of personality constructs and probably more valid measures as well.

Although we found evidence of high discriminant validity (mean $r = .03$), we also found that the common method variance (mean $r = .28$) was nearly as high as the convergent validities (mean $r = .30$). As previously reported, coworkers ($r = .36$) and customers ($r = .41$) exhibited the most common method bias. For coworkers, this was probably due to lack of familiarity because coworkers did not have extensive interaction with each other in part because sales representatives cover geographically

separated territories and tend to work out of their homes. Relatively little is known about customer ratings of personality. Therefore, it is not possible to explain why the relatively high common method bias was obtained. Given the relatively high validities for these ratings of conscientiousness, extraversion, and agreeableness, this seems to be a fruitful area for further research.

Other Issues

These results raise several questions pertaining to training and development activities. First, little is known about the extent to which individuals' relative standing on personality dimensions can be changed through developmental activities. Given that personality traits are thought to be relatively stable over time and situations, can developmental activities bring about relatively permanent changes in these areas? A related question is whether the behaviors summarized by the traits on some personality dimensions are more trainable than those summarized by the traits on other personality dimensions. For example, is it easier (or more beneficial) to train sales representatives to be agreeable, good-natured, and cooperative (agreeableness) than to train them to be sociable, talkative, and assertive (extraversion) or responsible, dependable, and achievement oriented (conscientiousness)? These issues may serve as a useful framework for future research pertaining to personality and performance linkages.

Two final notes are in order. First, the results of the present study do not address whether observer ratings of personality are more accurate than self-ratings. As pointed out by others (Gordon, 1970; Sulsky & Balzer, 1988), both validity and reliability are necessary but not sufficient conditions for accuracy. That is, accuracy of measurement is a function of both the strength and kind of relation between a set of observed scores and a corresponding set of true scores (Sulsky & Balzer, 1988). Therefore, our results do not necessarily mean that observers can more accurately assess a person's personality characteristics. However, the validity coefficients reported in our study demonstrate that observer ratings of job-relevant personality characteristics predict a supervisor's ratings of job performance at least as well as, if not better than, self-ratings.

Second, we are not arguing that the use of self-report personality inventories should be abandoned. As has been pointed out elsewhere, most personality inventories were not developed for personnel selection purposes (Schneider & Schmitt, 1986). It is likely that the validity of self-report personality inventories will be higher if they are designed to assess job-relevant constructs, such as conscientiousness (Barrick & Mount, 1991, 1993; Barrick et al., 1993; Hough et al., 1990; McHenry et al., 1990), or other job-relevant constructs based on a job analysis (Tett et al., 1991). As prior research has shown, conscientiousness is assessed to varying degrees or not at all in some cases, such as the in the Minnesota Multiphasic Personality Inventory (Barrick & Mount, 1991; Hogan, 1991; Johnson, Null, Butcher, & Johnson, 1984). In view of these findings, additional research is needed regarding the validity of self-report measures that assess conscientiousness (and other job-relevant personality variables) in construct-valid ways.

Summary

In summary, there is a distinction between the validity of personality constructs based on self-reports and the validity of the constructs based on observer ratings. Our results show that supervisor, coworker, and especially customer ratings of conscientiousness and extroversion are valid predictors of sales performance. They also show that observer ratings account for significant variance beyond that of self-ratings for the job-relevant personality constructs. The substantive message here is that the validity of personality constructs may be understated through reliance on the self-report method alone.

References

- Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance. *Personnel Psychology, 44*, 1-26.
- Barrick, M. R., & Mount, M. K. (1993). Autonomy as a moderator of the relationships between the Big Five personality dimensions and job performance. *Journal of Applied Psychology, 78*, 111-118.
- Barrick, M. R., Mount, M. K., & Strauss, J. P. (1993). Conscientiousness and performance of sales representatives: Test of the mediating effects of goal setting. *Journal of Applied Psychology, 78*, 715-722.
- Costa, P. T., Jr., & McCrae, R. R. (1985). *The NEO-PI personality Inventory*. Odessa, FL: Psychological Assessment Resources.
- Costa, P. T., Jr., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences, 13*, 653-665.
- Digman, J. M. (1972). The structure of child personality as seen in behavior ratings. In R. Dreger (Ed.), *Multivariate personality research* (pp. 587-611). Baton Rouge, LA: Claitor's Publishing.
- Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41*, 417-440.
- Funder, D. C., & Colvin, C. R. (1988). Friends and acquaintanceship, agreement, and the accuracy of personality judgment. *Journal of Personality and Social Psychology, 55*, 149-158.
- Funder, D. C., & Dobroth, K. M. (1987). Differences between traits: Properties associated with interjudge agreement. *Journal of Personality and Social Psychology, 52*, 409-418.
- Funder, D. C., Kolar, D. W., & Colvin, R. (1992). *When do others know us better than we know ourselves? Perspectives on personality by the self and by others*. Unpublished manuscript.
- Goldberg, L. R. (1992). The development of markers for the Big Five factor structure. *Psychological Assessment, 4*, 26-42.
- Gordon, M. E. (1970). The effect of the correctness of the behavior observed on the accuracy of the ratings. *Organizational Behavior and Human Performance, 5*, 366-377.
- Hogan, R. (1986). *Manual for the Hogan Personality Inventory*. Minneapolis, MN: National Computer Systems.
- Hogan, R. (1991). Personality and personality measurement. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (pp. 873-919). Palo Alto, CA: Consulting Psychologists Press.
- Hough, M. H., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effects of response distortion on those validities. *Journal of Applied Psychology, 75*, 581-595.
- Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72-95.
- John, O. P., & Robbins, R. W. (1991). *Accuracy in self-perception, self-enhancement bias, and narcissism: The power of individual differences*. Unpublished manuscript, Institute of Personality Assessment and Research, University of California, Berkeley.
- Johnson, J. H., Null, C., Butcher, J. N., & Johnson, K. N. (1984). Rep-

- licated item level factor analysis of the full MMPI. *Journal of Personality and Social Psychology*, 47, 105-114.
- Lilenthal, R. A. (1980). *The use of reference checks for selection*. Washington, DC: U.S. Office of Personnel Management.
- McCrae, R. R. (1982). Consensual validation of personality traits: Evidence from self-report ratings. *Journal of Personality and Social Psychology*, 43, 293-303.
- McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1990). Project A validity result: The relationship between predictor and criterion domains. *Personnel Psychology*, 43, 335-354.
- Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology*, 66, 574-583.
- Ones, D. S., Mount, M. K., Barrick, M. R., & Hunter, J. E. (in press). Personality and job performance: A critique of the Tett, Jackson & Rothstein (1991) meta-analysis. *Personnel Psychology*.
- Paulhus, D. L. (1986). Self-deception and impression management in test responses. In A. Angleitner & J. S. Wiggins (Eds.), *Personality assessment via questionnaire* (pp. 143-165). New York: Springer-Verlag.
- Rothstein, H. R. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology*, 75, 322-327.
- Scheier, M. F., Buss, A. H., & Buss, D. M. (1976). Self-consciousness, self-report of aggressiveness, and aggression. *Journal of Research in Personality*, 44, 637-644.
- Schmidt, F. L., & Hunter, J. E. (1992). Causal modeling of processes determining job performance. *Current Directions in Psychological Science*, 1, 89-92.
- Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology*, 37, 407-422.
- Schneider, B., & Schmitt, N. (1986). *Staffing organizations*. Glenview, IL: Scott, Foresman.
- Shrauger, J. S., & Osberg, T. M. (1981). The relative accuracy of self-predictions and judgments by others in psychological assessment. *Psychological Bulletin*, 90, 322-351.
- Smith, G. M. (1967). Usefulness of peer ratings of personality in educational research. *Educational and Psychological Measurement*, 27, 967-984.
- Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology*, 73, 497-506.
- Takemoto, N. K. (1979). *The prediction of occupational choice from childhood and adolescent antecedents*. Unpublished master's thesis, University of Hawaii, Honolulu.
- Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology*, 44, 703-742.
- Watson, D. (1989). Strangers' ratings of the five robust personality factors: Evidence of a surprising convergence with self-report. *Journal of Personality and Social Psychology*, 57, 120-128.
- Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology*, 53, 372-376.

Received February 5, 1993

Revision received August 9, 1993

Accepted August 10, 1993 ■