Journal of Personality and Social Psychology
1989, Vol. 57, No. 1, 120-128

Strangers' Ratings of the Five Robust Personality Factors: Evidence of a Surprising Convergence With Self-Report

David Watson
Southern Methodist University

Attempted to replicate and extend the results of Passini and Norman (1966), who found surprising evidence of convergent validity (i.e., significant correlations with the targets' self-ratings) in strangers' judgments of 5 broad personality factors. In the current study, 250 previously unacquainted Ss were run in small, same-sex groups of various sizes. Ss rated both themselves and their fellow group members on the same set of 20 bipolar trait scales used by Passini and Norman. Consistent with previous research, significant self–peer agreement correlations were obtained for Extraversion and Conscientiousness. Ratings of Agreeableness also showed significant convergent validity when a sufficient number of peers rated the target. More generally, self–peer agreement correlations tended to rise as the number of peer raters increased. Possible explanations for the validity of strangers' trait ratings are discussed.

Recently, extensive interest has focused on a five-factor structure of personality that provides a reasonably comprehensive scheme for the global classification of personality traits (e.g., Digman & Inouye, 1986; Digman & Takemoto-Chock, 1981; Hogan, 1983; McCrae & Costa, 1985, 1987; Noller, Law, & Comrey, 1987). The five traits in this model—Extraversion or Surgency, Agreeableness, Conscientiousness, Emotional Stability (vs. Neuroticism), and Culture—have repeatedly emerged in factor analyses of personality trait ratings.

Development of the Five-Factor Model

This line of research originated in Allport and Odbert's (1936) effort to compile an exhaustive list of trait-related terms in the English language. Allport and Odbert eventually settled on a list of 4,504 terms that clearly represented trait dispositions. Cattell (1945, 1946) reduced this to a more manageable pool of 171 variables by rationally sorting the terms into synonym groups; these were further reduced to 35 bipolar scales through a cluster analysis of trait ratings. Cattell originally argued that 12–15 factors were necessary to account for the correlations among these terms. However, Fiske (1949) and Tupes and Christal (1961) subsequently showed that five robust factors—which emerged in diverse subject samples and regardless of the degree or type of acquaintance between the rater and target—were sufficient to represent the structure of peer ratings (see Wiggins, 1973, for a review of this early literature).

Norman (1963) created a set of 20 bipolar rating scales by selecting the four best markers of each of the five recurrent factors identified by Tupes and Christal (1961). Using this set of terms, Norman and his colleagues replicated the five-factor structure in several studies involving diverse conditions and populations (Norman, 1963, 1969; Norman & Goldberg, 1966; Passini & Norman, 1966). The robustness of this structure has since been confirmed in several other peer rating studies using different sets of trait terms (Digman & Inouye, 1986; Goldberg, 1981; McCrae & Costa, 1987). Highly convergent structures have also recently been identified in self-ratings (McCrae & Costa, 1985, 1987; Noller et al., 1987), leading McCrae and Costa (1986) to assert that this model has now been established as a basic taxonomic scheme for personality.

Effects of Acquaintance on Peer Ratings

The meaning of this five-factor structure has, however, been a subject of controversy. Passini and Norman (1966), for example, obtained the same five factors when subjects rated the personality traits of complete strangers. More strikingly, both D'Andrade (1965) and Hakel (1969) derived highly convergent five-factor structures from similarity ratings of the trait terms themselves. On the basis of these findings, some writers (e.g., Mischel, 1968) have argued that the five-factor structure merely represents the implicit personality theories of the raters, and that it does not necessarily reflect the organization of traits within the ratees. Thus, it is argued, these trait ratings (and the factors they produce) may have little to do with the actual organization of personality as it exists in individuals (for a discussion of this issue, see Norman & Goldberg, 1966; Wiggins, 1973).

This controversy led Norman and Goldberg (1966) to perform extensive analyses on various sets of peer ratings. Two of these data sets are most pertinent to this discussion. The first is that used in the Passini and Norman (1966) study noted previously. This sample consisted of 84 undergraduates who were divided into same-sex groups of 6–9 individuals. Subjects within each group rated each other on the 20 Norman scales described earlier, despite having had no prior acquaintance with one another—in fact, their contact was limited to being in the

I would like to thank Mary Dieffenwierth, Kathleen Ferrara, Julie Harrison, Eleanor Herndon, Hema Patel, and Lisa Sharp for conducting the rating sessions reported in this article. Thanks also to Lee Anna Clark for her helpful comments on an earlier draft.

Correspondence concerning this article should be addressed to David Watson, Department of Psychology, Southern Methodist University, Dallas, Texas 75275.

same room for less than 15 min with no opportunity for verbal communication. The second sample was composed of 73 Peace Corps trainees tested after 3 months of intensive interpersonal contact. These subjects were also formed into small groups (8–11 individuals per group) and asked to rate each other on the same set of 20 bipolar scales. Thus, Norman and Goldberg were able to compare two strikingly different sets of peer ratings—one generated by well-acquainted peers, the other by complete strangers.

The same five factors were identified in both groups, supporting the notion that this structure is inherent in the semantics of terms themselves. However, Norman and Goldberg (1966) were also able to show important differences between the two sets of peer judgments. Specifically, the ratings in the Peace Corps sample were both more reliable (i.e., they showed better interrater agreement) and more valid (i.e., the averaged ratings correlated more highly with the target's own self-ratings). Thus, Norman and Goldberg demonstrated that the well-acquainted peers were systematically responding to actually observed trait characteristics of the targets.

These findings have since been replicated by several other investigators. For example, Funder and Colvin (1988) found that ratings made by well-acquainted peers showed significantly better agreement with one another (mean interrater r = .26) and with the target's own self-ratings (mean self–peer r = .27) than did those generated by strangers (mean interrater r = .09; mean self–peer r = .05). Similarly, Jackson, Neill, and Bevan (1973) reported that self–peer correlations were consistently higher when judges were better acquainted with their targets. Weiss (1979) obtained somewhat inconsistent results, but generally found that increasing the information available to raters led to more reliable and differentiated peer judgments. Thus, considerable evidence corroborates Norman and Goldberg's (1966) conclusion that longer target–judge acquaintance produces trait ratings that are both more reliable and more valid.

Reliability and Validity of Strangers' Trait Ratings

Norman and Goldberg's (1966) basic findings are important and have been widely cited. One intriguing aspect of their study has been relatively neglected, however. Although the correlations between the self-ratings and averaged peer ratings tended to be lower in the Passini and Norman (1966) sample (which was composed of unacquainted peers), they were nevertheless significant for three of the five factors: Extraversion (r = .38), Conscientiousness (r = .34), and Culture (r = .32). Only Agreeableness (r = .15) and Emotional Stability (r = .02) failed to produce a significant self–peer correlation in these data. Moreover, given the small sample sizes, none of these correlations were significantly lower than the corresponding coefficients in the Peace Corps data (which ranged from .27 for Agreeableness to .54 for Extraversion).

It is also noteworthy that the same pattern of correlations was obtained in the two samples. That is, Extraversion had the highest self–peer correlation, followed closely by Conscientiousness and Culture, whereas Agreeableness and Emotional Stability had the poorest convergent correlations in both data sets. These results are consistent with others' (using well-acquainted judges) that show that more observable traits produce better interjudge agreement and higher self–peer correlations (e.g., Funder & Colvin, 1988; Funder & Dobroth, 1987; Kenrick & Stringfield, 1980). Furthermore, Funder and Dobroth (1987) and Funder and Colvin (1988) also reported that interrater and self–peer correlations were generally highest for extraverted traits (e.g., ratings of talkative and gregarious) and lowest for those representing low Emotional Stability (e.g., judgments of anxious and thin-skinned). Thus, the strangers in the Passini and Norman (1966) sample produced ratings that were orderly and surprisingly convergent with self-ratings.

Neither Norman and Goldberg (1966) nor subsequent writers have commented much on these aspects of the Passini and Norman (1966) data. Norman and Goldberg simply concluded that "there was no convergence for two of the five factors in the Passini and Norman study and only somewhat tenuous evidence for convergent and discriminant validity for the other three factors" (p. 689). Similarly, Wiggins (1973) merely stated that "the corresponding values for the Passini and Norman subjects, who were not acquainted, were generally low and ranged from near zero to a high of .38" (p. 349).

Although it is true that these convergent validity correlations are not high, it is nevertheless remarkable that the unacquainted peers did as well as they did, given that they had less than 15 min of nonverbal contact with the targets. That significant self–peer correlations can be obtained under such circumstances is an important and puzzling finding.

However, the Passini and Norman (1966) sample was small, and their results obviously require corroboration. Two recent studies have, in fact, replicated certain aspects of these findings. First, Funder and Colvin (1988) compared Q-sort judgments generated by friends and strangers. As noted earlier, the friends produced ratings that showed stronger interjudge agreement and higher self–peer correlations. Nevertheless, the strangers' ratings again proved to be surprisingly systematic: Significant interjudge agreement was reported on 50 of the 100 Q-sort items, and significant self–peer correlations were obtained on 24 items. Moreover, consistent with Norman and Goldberg's (1966) results, self–peer agreement was greatest on items representing extraverted behavior and lowest for those reflecting low Emotional Stability.

Second, Albright, Kenny, and Malloy (1988) studied peer judgments in small groups of previously unacquainted individuals. Ratings were made on a subset (one or two items per factor) of the original Norman (1963) scales. The results in this study generally replicated those reported by Norman and Goldberg (1966). Significant interrater and self–peer agreement was found for the Extraversion scales and, to a lesser extent, for the Conscientiousness items; as expected, ratings on the Agreeableness and Emotional Stability scales showed little reliability or validity. The one significant discrepancy concerned the Culture scales; unlike Norman and Goldberg, Albright et al. found little self–peer or interjudge agreement on these items. This discrepancy could, of course, reflect the fact that Albright et al. used only selected Culture items in their study.

Current Study

To date, no study has replicated Norman and Goldberg's (1966) results using the full set of trait descriptors. This study was designed to replicate and extend these findings, using the same set of 20 scales in a much larger sample. On the basis of Norman and Goldberg's results, I predicted that strangers' ratings of Extraversion, Conscientiousness, and Culture would be significantly related to the targets' own self-ratings on these factors. However, I did not expect ratings on Agreeableness and Emotional Stability to produce significant self–peer correlations.

A second goal of this research was to examine self–peer correlations as a function of the number of raters whose judgements are averaged. Consistent with a more general body of evidence demonstrating the benefits of aggregation (Epstein, 1980; Rushton, Brainerd, & Pressley, 1983; Rushton, Jackson, & Paunonen, 1981), self–peer correlations have been shown to rise as the number of raters increases (McCrae & Costa, 1987; Norman & Goldberg, 1966). This increase likely reflects the increasing reliability of the averaged peer ratings. Aggregation is likely to be especially important for strangers' ratings, given the relatively modest interrater agreement among these judgments (Albright et al., 1988; Funder & Colvin, 1988; Norman & Goldberg, 1966). Thus, I hypothesized that self–peer correlations would generally rise as the number of raters who assessed the target increased.

Method

Subjects

Subjects were 250 undergraduates (101 men, 149 women) enrolled in various psychology courses at Southern Methodist University, a private southwestern university. The subjects participated in return for extra course credit.

Measures

All trait ratings were made on Norman's (1963) 20 bipolar scales. Each of the five factors has four clear and consistent markers within this pool (Norman, 1963; Norman & Goldberg, 1966; Passini & Norman, 1966). The scales—grouped according to the factor they have been shown to define—are listed in Table 1.

In the original Passini and Norman (1966) study, subjects rated themselves on a graphic rating scale. Peer scores, however, were obtained using a peer nomination format: Subjects were asked to assign one third of the other members of their group to Pole A (e.g., talkative) and another third to Pole B (e.g., silent) of each scale. An overall peer score for each person on each scale was then derived by summing his or her nominations on Pole A, subtracting the number of nominations on Pole B, and adjusting for the size of the group.

In contrast, this study used a simple 5-point Likert rating format for both self- and peer ratings, in which 1 = *very much like Trait A,* 3 = *about average on this dimension,* and 5 = *very much like Trait B.* I created an overall peer score for each subject on each of the 20 scales by averaging the ratings he or she received for that trait.

I then obtained overall peer scores on the five recurrent factors by summing the average peer ratings on the four marker scales. Similarly, I computed self-scores on the five factors by summing the subject's self-ratings on the four component items. The resulting peer scores were all reasonably reliable (using coefficient α), especially considering they each contain only four items: Extraversion, .88; Agreeableness, .73; Conscientiousness, .78; Emotional Stability, .57; and Culture, .66. These alpha reliabilities reflect moderate to strong average interitem correlations, ranging from .25 (Emotional Stability) to .65 (Extraversion). Thus, consistent with their factor-analytic derivation, the peer factor scores are all internally consistent.

The self-scores were less homogeneous, however: Extraversion, .67; Agreeableness, .48; Conscientiousness, .45; Emotional Stability, .54; and Culture, .49. These reliability estimates reflect low to moderate average interitem correlations, ranging from .17 (Conscientiousness) to .34 (Extraversion). These lower reliabilities raise the possibility that the factor scores may not adequately reflect self–peer agreement at the individual scale level. Because of this, I examined self–peer agreement at both the factor score and individual scale level.

Procedure

The procedure used in the present study differs somewhat from that of Passini and Norman (1966). In their design, all subjects were tested in a single session held on the first day of an introductory experimental psychology course. Subjects had spent less than 15 min together and were allowed no verbal contact.

In this study, subjects were seen in several sessions (15–40 subjects per session) over the course of the spring 1986, fall 1986, and spring 1987 semesters. At the beginning of a session, all subjects sat quietly and completed various questionnaires for approximately 20–40 min; these were collected for other studies and are not discussed here. There was no formal opportunity for interpersonal contact during this period. The subjects were then formed into small, same-sex groups (5–10 subjects per group) by moving their desks into a circular array. A number was then taped to each desk. These numbers corresponded to those found on the peer rating form and enabled the subjects to identify one another.

As noted earlier, Passini and Norman (1966) allowed no verbal contact between subjects. However, previous research has shown that voice quality is an important source of information in peer judgments (e.g., O'Sullivan, Ekman, Friesen, & Scherer, 1985), and it seemed desirable to examine whether verbal contact influenced the level of self–peer agreement. Thus, each subject was allowed one standardized statement: "Hi, my name is . . . ." When the last subject in the group had introduced him- or herself, the subjects began their ratings. The self-ratings were always completed first, followed by the peer ratings.

The subjects were told that the study involved strangers' perceptions of personality, and that it was important that they only rate individuals they had never met before. In most cases, it was possible to separate the subjects into groups in which all members were strangers. However, in a few sessions this proved impossible; in these instances, subjects who were previously acquainted were put into the same group, but did not rate one another.

Altogether, 250 subjects were run in 37 groups (M = 6.8 subjects per group). However, because some subjects were previously acquainted, and others either failed to follow instructions or to complete all of the ratings, many subjects were not rated by all of the other group members. When all unusable ratings were discarded, the final breakdown was the following: 8 subjects were rated by one peer, 16 by two, 11 by three, 49 by four, 73 by five, 33 by six, 53 by seven, 2 by eight, and 5 by nine (M = 5.1 raters per subject).

I should again emphasize that the subjects in this study were previously unacquainted with one another. At the time they began their peer ratings, the subjects had spent anywhere from 30 min to 60 min together in a situation that permitted little formal interaction and that discouraged informal interaction as well. Thus, the subjects were at most only minimally acquainted when they rated one another.

However, I should also note that over the course of the rating session the raters had the opportunity to observe other subjects if they so chose. Moreover, some minimal social interaction was possible while the subjects were in the small groups. Indeed, even though extraneous talking and social interaction were actively discouraged, one experimenter reported that it proved almost impossible to keep certain subjects (presumably extraverts) quiet. Thus, it seems likely that subjects had the opportunity to discern at least some salient personality characteristics during the session, a point that is discussed later.

Results

Self–Peer Agreement in the Overall Sample

Table 1 presents correlations between the targets' self-ratings and the averaged peer judgments. These were computed on the five factor scores and on each of the 20 individual scales. Simple self–peer agreement correlations are displayed in the first column of the table.

Before examining these results, it is important to investigate the possibility that significant correlations partly or largely reflect substantial sex differences on a scale or trait. For example, a significant correlation between self- and peer-ratings of Conscientiousness would result if women tend to rate themselves higher on this dimension and strangers correctly perceive this fact. Therefore, to rule out sex differences as a possible explanation for such findings, I also computed partial correlations, controlling for subject sex. These partial correlations are shown in the second column of Table 1, and it is noteworthy that in every case they are virtually identical to the simple zero-order correlations. Thus, subject sex did not have an important influence on self–peer agreement in this sample and is not considered further.

In interpreting the results shown in Table 1, recall that Norman and Goldberg (1966) found significant self–peer correlations for Extraversion, Conscientiousness, and Culture, and that Albright et al. (1988) reported significant self–peer agreement for marker scales of Extraversion and Conscientiousness. Table 1 indicates that these results are largely replicated in the current data. Most important, there is again evidence that strangers' trait ratings are surprisingly convergent with the target's own self-ratings. In the current data, Extraversion again shows the highest convergent validity, and the factor score correlation ($r = .41$) is very similar to the value reported by Norman and Goldberg ($r = .38$). It is also important to note that all four marker items of Extraversion display significant self–peer agreement, ranging from .17 (frank, open vs. secretive) to .41 (sociable vs. reclusive). Thus, this self–peer convergence was general across the factor and not confined to one particular aspect of Extraversion.

Conscientiousness also has a significant self–peer correlation ($r = .16$), but the value in the current sample is substantially lower than that reported by Norman and Goldberg ($r = .34$). Interestingly, among the marker scales only fussy, tidy vs. careless ($r = .21$) showed a significant level of self–peer agreement. In other words, although subjects were able to rate the neatness aspect of Conscientiousness with some accuracy, they were not able to judge its responsibility–dependability component. These results suggest that self–peer convergence on this factor may largely reflect the fact that strangers can actually observe individual differences in neatness among the targets; responsibility, however, must be inferred from behavioral manifestations that are not readily available to unacquainted peers. This conclusion is further supported by the data of Albright et al. (1988), who found that interrater consensus on their Conscientiousness items was largely a function of how formally and neatly the target was judged to be dressed.

The results for Agreeableness and Emotional Stability also replicate findings from earlier studies: There was no significant self–peer agreement on either factor. Moreover, none of the marker scales for Agreeableness showed any significant self–peer convergence. Among the Emotional Stability items, only nervous, tense vs. poised displayed a significant, but low ($r = .14$), level of agreement.

Table 1
Self–Peer Agreement on the Factor Scales and Their Component Items: Zero-Order and Partial Correlations (Controlling for Sex)

	Self–peer agreement correlations	
Factor scales and component items	Zero order	Partial
Extraversion	.41*	.40*
Talkative vs. silent	.25*	.24*
Frank, open vs. secretive	.17*	.16*
Adventurous vs. cautious	.27*	.27*
Sociable vs. reclusive	.41*	.40*
Agreeableness	.08	.06
Good-natured vs. irritable	.06	.03
Not jealous vs. jealous	−.02	−.02
Mild, gentle vs. headstrong	.06	.06
Cooperative vs. negativistic	.09	.04
Conscientiousness	.16*	.15*
Fussy, tidy vs. careless	.21*	.20*
Responsible vs. undependable	.07	.06
Scrupulous vs. unscrupulous	.00	−.01
Persevering vs. quitting, fickle	.05	.04
Emotional Stability	−.01	.00
Nervous, tense vs. poised	.14*	.16*
Anxious vs. calm	−.01	.00
Excitable vs. composed	−.02	−.01
Hypochondriacal vs. not so	.05	.04
Culture	.10	.09
Artistically sensitive vs. insensitive	.09	.07
Intellectual vs. unreflective, narrow	.08	.09
Polished, refined vs. crude, boorish	.16*	.15*
Imaginative vs. simple direct	.09	.09

Note. $N = 250$.
* $p < .05$ (two-tailed test).

Finally, the current data produced little self–peer agreement on the Culture factor. The factor score correlation ($r = .10$) was nonsignificant, and only one component item (polished, refined vs. crude, boorish) showed significant agreement ($r = .16$). These results differ from those of Norman and Goldberg (1966), who reported a moderate correlation ($r = .32$); however, they are consistent with the findings of Albright et al. (1988), who found little self–peer agreement on their Culture items. The reasons for these inconsistent findings on the Culture factor are as yet unclear.

Summarizing across the three relevant studies, there is consistent evidence of convergent validity in strangers' ratings of two traits (Extraversion, Conscientiousness), mixed results with regard to another (Culture), and no evidence of convergence for two others (Agreeableness, Emotional Stability). Item-level analyses indicate that self–peer agreement is general across the various markers of Extraversion, but is largely confined to the neatness component of Conscientiousness.

Effects of Aggregation

Analyses of one to five raters. As noted earlier, a second goal of this study was to examine the reliability and convergent va-

Table 2
Self–Peer Agreement Correlations and Interrater Reliabilities for the Factor Scales as a Function of Increasing Aggregation

Factor scales and number of raters	Self–peer agreement correlations			Spearman–Brown reliability
	M	Range[a]	% significant[a]	
Extraversion				
One rater	.33*	.20–.42	100	.36
Two raters	.39*	.30–.49	100	.52
Three raters	.43*	.35–.50	100	.62
Four raters	.45*	.41–.49	100	.69
Five raters	.46*	—	—	.73
Agreeableness				
One rater	.10	.02–.22	20	.21
Two raters	.13	.04–.23	40	.33
Three raters	.14	.06–.20	50	.44
Four raters	.16*	.10–.18	60	.51
Five raters	.16*	—	—	.57
Conscientiousness				
One rater	.14	.07–.19	40	.29
Two raters	.18*	.13–.21	70	.45
Three raters	.19*	.16–.23	100	.55
Four raters	.21*	.19–.23	100	.62
Five raters	.21*	—	—	.67
Emotional Stability				
One rater	.03	−.07–.09	0	.09
Two raters	.03	−.03–.09	0	.15
Three raters	.04	−.01–.09	0	.21
Four raters	.04	.01–.09	0	.27
Five raters	.05	—	—	.31
Culture				
One rater	.03	−.07–.18	20	.14
Two raters	.03	−.09–.18	10	.23
Three raters	.04	−.08–.12	0	.32
Four raters	.04	−.02–.08	0	.39
Five raters	.04	—	—	.44

Note. These correlations are calculated on subjects who were rated by five or more peers ($n = 166$).
[a] For each trait, there were 5 one-rater, 10 two-rater, 10 three-rater, and 5 four-rater composites. See text for more details.
* $p < .05$ (two-tailed test).

lidity of strangers' trait ratings as a function of the number of judges who rated the target. I examined this issue in the subsample of 166 subjects who were rated by at least five peers. If a subject had been rated by more than five peers, I used only the 1st five peer ratings (the order of the peer judges was random and simply reflected the arbitrarily assigned subject numbers within each rating group). These analyses were restricted to the five factor scores.

Relevant results are displayed in Table 2. The 1st three columns of the table show self–peer agreement correlations for composites representing increasing numbers of raters. For example, at the one-rater level, the targets' self-rated Extraversion score was individually correlated with the Extraversion ratings of each of the five judges (Peer 1, Peer 2, Peer 3, etc.). At the two-rater level, self-rated Extraversion was correlated with the average Extraversion rating of all possible pairs of judges (the mean of Peers 1 and 2, 1 and 3, 2 and 3, etc.). Similarly, the targets' self-scores on each trait were also correlated with all possible three-rater and four-rater composites. Finally, at the five-rater level self-rated Extraversion was correlated with the average of all five judges combined.

The first column of Table 2 shows the mean self–peer agreement correlations—averaged across all possible composites at each level—for the five factors. (Fisher's r to z transformation was used for the computation of all average correlations shown in Tables 2 and 3). The second column of the table indicates the range of these correlations, and the third column reports the percentage of the individual convergence correlations that were significant ($p < .05$, two-tailed test).

Before analyzing self–peer convergence, however, it is informative to examine the extent to which the peer raters agreed among themselves. The fourth column of Table 2 displays Spearman–Brown reliability estimates based on the average interrater correlation. Because the order of the raters was random, I computed these reliability estimates using intraclass correlations. Consistent with previous research (e.g., Albright et al., 1988; Funder & Colvin, 1988; Funder & Dobroth, 1987), interjudge agreement was strongest for the Extraversion ratings and lowest for the Emotional Stability ratings. The interjudge correlations for the Emotional Stability ratings were low enough to suggest that these minimally acquainted peers were essentially making random guesses as to the targets' standing on this factor.

Turning now to the convergence correlations, Table 2 indicates that increasing the number of peer judges enhanced self–

peer agreement on some factors but not on others. Specifically, ratings on Emotional Stability and Culture were completely unaffected by aggregation. It appears that raters were simply unable to discern personality characteristics in their fellow group members that were relevant to these factors. As noted before, the results for Emotional Stability are consistent with previous research in the area (e.g., Albright et al., 1988; Norman & Goldberg, 1966), but the findings on the Culture factor have been inconsistent across studies. It appears that strangers can attain significant self–peer agreement on this factor under certain circumstances, but the specific parameters that influence the level of convergence have not yet been identified.

In contrast, strangers' ratings on Extraversion, Conscientiousness, and Agreeableness benefited from increasing the number of judges: On these factors, the convergence correlations generally increased in magnitude and significance with the inclusion of additional raters. It is especially noteworthy that ratings of Agreeableness—which did not yield a significant self–peer correlation in the overall sample (see Table 1)—showed a low but significant level of convergence in the four- and five-judge composites. Thus, it appears that, given a sufficient number of raters, strangers may be able to achieve a significant level of convergent validity on this dimension.

Ratings on Extraversion showed the strongest effect due to aggregation: The self–peer convergence correlation increased with the addition of each new rater, and the coefficient for the five-factor composite ($r = .46$) was significantly higher than the average correlation observed at the disaggregated (one-rater) level ($r = .33$; $Z = 1.97$, $p < .05$). Nevertheless, it is noteworthy that the strangers' ratings were moderately correlated with the self-reports even at the disaggregated level; in fact, as shown in the third column of Table 2, *all* of the individual judges produced ratings that were significantly related to the targets' self-rated Extraversion. It is striking that significant self–peer convergence can be demonstrated on Extraversion even when a single, minimally acquainted peer rates the target.

Aggregation beyond five raters. The analyses in Table 2 are informative but incomplete, in that they terminate with five-rater composites. There is no reason to believe, of course, that the effects of aggregation will necessarily asymptote at this level, which raises the question of what happens to the convergent validity of strangers' trait ratings beyond five judges. A comprehensive treatment of this issue is beyond the scope of this article, but it can be examined in a preliminary fashion by calculating self–peer agreement correlations in the subsample of subjects ($n = 93$) who were assessed by six or more peers.

These correlations are reported in Table 3 (in the row labeled *Current sample*). Similar to the findings presented in Table 2, these data indicate that further aggregation enhanced convergent validity on some of the factors but not on others. Specifically, self–peer agreement was essentially unchanged on Extraversion and Emotional Stability; that is, the convergence correlations in Table 3 for Extraversion ($r = .43$) and Emotional Stability ($r = -.04$) are virtually identical to the corresponding values for the five-rater composites reported in Table 2. In contrast, ratings on Conscientiousness (from .21 to .28), Culture (from .04 to .20), and Agreeableness (from .16 to .31) showed a modest increase in self–peer agreement. It is also noteworthy that the convergence coefficient for Culture approached significance ($p < .06$, two-tailed) in this subsample. These results are only exploratory, but they suggest that adding in additional raters may further enhance the convergent validity of strangers' trait ratings, at least on some of the factors and under certain circumstances. This would seem to be a promising area for further research.[1]

Effects of Acquaintance on the Validity of Peer Ratings

Reviews of the literature on self–peer agreement generally conclude that convergence increases as the judge becomes better acquainted with the target (e.g., Funder, 1987; Funder & Colvin, 1988). This is certainly true in the sense that agreement correlations are generally higher when well-acquainted peers are used, but often the comparisons have involved studies using different measures and methodologies. The data reported here offer an interesting opportunity to compare the convergent validity of ratings made by strangers and well-acquainted peers on the same set of trait measures.

Table 3 reports self–peer agreement correlations in four samples, two using well-acquainted peers, two involving strangers, and all using the same bipolar rating scales (note, however, that the Culture factor was not assessed in the data of Norman, 1969). The data involving well-acquainted raters were originally reported by Norman and Goldberg (1966, Table 3, p. 689) and by Norman (1969, Table 8, p. 434). The correlations involving strangers come from the Passini and Norman (1966) sample (reported in Norman & Goldberg, 1966, Table 3), and from the current study, using the subsample of subjects who were rated by six or more peers. I used this subsample so that the number of peer raters would be comparable across studies. Norman (1969) does not report the number of raters used in his study, but the mean number of peers is similar in the other three data sets (8.2 in Norman & Goldberg, 6.2 in Passini & Norman, 6.8 in the current subsample).

Table 3 also presents weighted mean self–peer agreement correlations for each trait, calculated separately for the two types of raters. These results generally indicate that well-acquainted peers do, in fact, produce ratings that converge more strongly with the target's own self-view. The agreement correlations are always higher for the well-acquainted peers than for the strangers, and on two factors (Conscientiousness and Emotional Stability) the difference between the groups is significant. In the case of Conscientiousness, strangers' ratings show a moderate level of convergent validity ($r = .31$), but the well-acquainted peers do significantly better, achieving a mean correlation of .56 with the self-ratings. For Emotional Stability, well-acquainted peers are only moderately successful, but, as we have seen, strangers display no convergent validity whatsoever on this factor.

Nevertheless, it is remarkable that ratings made by minimally acquainted peers converge significantly with the targets' self-ratings on four of the five factors. It is true that strangers' ratings do not achieve the reliability and validity shown by well-ac-

[1] It should be noted that this specific group of 93 subjects may not be entirely representative of the overall sample. Thus, these higher self–peer correlations may not simply be due to the effects of greater aggregation per se, but may reflect—to some extent at least—idiosyncratic characteristics of this particular subsample of targets and judges. Because of this, these results can only be considered suggestive.

Table 3
Comparison of Self–Peer Agreement Correlations in Ratings Made by Strangers and Well-Acquainted Peers

		Trait				
Type of rating and sample	*n*	EXT	AGR	CON	STAB	CULT
Well-acquainted peers						
Norman & Goldberg (1966)	73	.54*	.27*	.47*	.32*	.45*
Norman (1969)	169	.51*	.32*	.60*	.31*	—
Strangers						
Passini & Norman (1966)	84	.38*	.15	.34*	.02	.32*
Current sample (six or more peers)	93	.43*	.31*	.28*	−.04	.20
Weighted mean correlations						
Well-acquainted peers	242	.52*	.31*	.56*[a]	.31*[a]	.45*
Strangers	177	.41*	.24*	.31*	−.01	.26*

Note. EXT = Extraversion, AGR = Agreeableness, CON = Conscientiousness, STAB = Emotional Stability, CULT = Culture.
[a] Indicates that this correlation is significantly higher than the strangers' mean for the same trait. All other pairwise comparisons are nonsignificant.
* $p < .05$ (two-tailed test)

quainted peers. Still, given a sufficient number of raters, strangers can achieve a significant level of convergent validity on most of these factors.

Discussion

Summary of Results

Replicating and extending Norman and Goldberg's (1966) results, I have demonstrated in this study that strangers' trait ratings are surprisingly congruent with the targets' own self-reports. In fact, with the single exception of Emotional Stability, ratings on all of the factors have now shown some evidence of significant self–peer agreement (see Table 3).

Consistent with previous research (Albright et al., 1988; Funder & Colvin, 1988; Norman & Goldberg, 1966), the strongest and most general convergence was seen on Extraversion. Averaging across the current data and those of Passini and Norman (1966), strangers produced a self–peer agreement correlation of .41 for Extraversion, which is comparable in magnitude to the coefficient obtained with well-acquainted peers (see Table 3). Furthermore, Table 2 shows that all of the individual judges produced ratings that were significantly correlated with the targets' self-rated Extraversion score. The implications of these findings are remarkable: In judging targets' standing on extraversion, a single, minimally acquainted peer can achieve a significant level of convergent validity.

As expected, this study's data also indicated that the convergence correlations generally rose as the number of peer raters increased. Only ratings of Emotional Stability were completely unaffected by aggregation; overall, it appeared that subjects were simply unable to discern relevant trait characteristics in the targets and were therefore forced to make random judgments on this factor. In contrast, convergent validity on the other factors tended to increase with the addition of more peer judges. These data again demonstrate the importance of having a sufficient number of judges in peer-rating studies, a point first emphasized by Norman and Goldberg (1966).

Sources of Self–Stranger Convergence

It is now well established that minimally acquainted judges can generate significant self–peer correlations, at least for certain traits and given a sufficient number of raters. Obviously, the most important remaining question is this: How do they achieve this convergence with self-report? That is, what information do they use that enables them to detect significant aspects of the targets' personalities?

Further research is needed to identify the sources of this surprising validity in strangers' trait ratings. It is important to emphasize that the critical issue here is not what stimuli strangers use in their judgments (much is already known about this), but what stimuli they use to achieve significant agreement with self-ratings. This will undoubtedly be a complex issue to investigate, as it is likely that different cues will be involved in the estimation of different traits; that is, the cues that are important for Extraversion will likely differ from those that are useful for Conscientiousness.

What is needed are controlled experimental studies that directly manipulate the various sources of information normally available to unacquainted peers. It would be interesting, for example, to evaluate the role of voice characteristics by directly comparing ratings made with and without any verbal contact. In this regard, it is noteworthy that Passini and Norman (1966) permitted no formal verbal contact among their subjects, whereas participants in this study were allowed to introduce themselves to one another. The fact that the two designs yielded similar results suggests that formal verbal contact of this sort does not significantly influence the validity of the strangers' ratings. Nevertheless, this procedural change may be responsible for some of the findings that were inconsistent across studies. This is an interesting topic for subsequent research.

In future studies it will also be desirable to control the information available to raters in a more rigorous manner. The design used in this study (and in most of the other relevant studies as well, e.g., Albright et al., 1988; Passini & Norman, 1966) did not completely eliminate informal observation and extraneous social interaction. Subjects had various opportunities to observe the targets whom they were rating over the course of the rating session. Furthermore, although talking and social interaction were actively discouraged by the experimenters, it proved impossible to eliminate such interaction entirely. It may be these data that enable strangers to achieve significant convergent validity in their ratings. A greater degree of experimental control could be achieved by videotaping the targets, so that judges could complete their ratings without directly observing or interacting with them. This would allow one to control or manipulate the presentation of such potentially important cues as voice characteristics, facial expression, physical appearance, dress, and so on (see Funder & Colvin, 1988, for an example of this type of design).

However, even if specific cues are found to affect the level of self–peer convergence, this will not completely answer the question of how and why this agreement occurs. A more fundamental issue concerns the nature and interpretation of these significant self–peer correlations. There are several (not mutually exclusive) possibilities, two of which are briefly noted here. First, strangers may observe significant cues that are a direct expression of the target's personality. For example, extraverts may simply act in a more socially dominant or exhibitionistic manner, even in highly structured situations that allow for little or no social interaction. Thus, during the rating sessions extraverts may have talked more, sat closer to other group members, engaged in more attention-seeking behaviors, and so on, whereas introverts sat quietly and unobtrusively at their desks. Similarly, high scorers on the Conscientiousness factor may have dressed more neatly than low scorers—recall that among the marker scales of Conscientiousness, only fussy, tidy vs. careless showed significant self–peer convergence (see also Albright et al., 1988). In other words, strangers may achieve a significant level of convergent validity because they perceive important aspects of personality that are readily manifested, even in the course of superficial social interaction.

This explanation is consistent with the general pattern of findings regarding the convergent validity of peer ratings. As was discussed earlier, studies with well-acquainted peers have generally found that behaviorally based, externally observable dispositions (such as Extraversion) produce higher self–peer correlations than do more internal, subjective traits (such as Emotional Stability; see Funder & Dobroth, 1987; Kenrick & Stringfield, 1980). This study's results, as well as those of Passini and Norman (1966) and Albright et al. (1988), indicate that strangers' ratings follow a very similar pattern. Thus, strangers—like well-acquainted peers—may be able to observe direct manifestations of certain traits, even with minimal interpersonal contact.

Another possibility is that strangers are calling on common stereotypes that, for whatever reason, have a grain of truth to them. For example, several stereotypes involve various aspects of physical appearance—"fat people are jolly," "redheads are hot-tempered," and so on. Interestingly, recent research has shown that at least one of these appearance-based stereotypes may help to produce significant self–peer correlations. Berry and McArthur (1985) found that babyfaced adults are perceived as being weak, submissive, naive, and approachable. Berry and Brownlow (in press) replicated these results and also reported evidence indicating that babyfaced adults have congruent self-impressions—that is, they view themselves as weaker, more submissive, and more affiliative than other individuals. Thus, consensually defined stereotypes involving facial configuration or other aspects of physical appearance are potentially a source of self–peer agreement.

However, there are at least two possible theoretical explanations for any appearance-based convergence. One possibility is that it reflects a self-fulfilling prophecy. Research has shown that subjects' expectations of others affect their behavior toward them in subsequent interactions. These others, in turn, must modify their own behavior, and often do so in a way that confirms the original expectation (e.g., Anderson & Bem, 1981; Snyder, Tanke, & Berscheid, 1977). For example, the self-impressions of babyfaced adults may reflect the fact that because others expect them to be weak and submissive, they are treated as if they were weak and submissive. This, in turn, may ultimately force these individuals to behave in a weak and submissive manner. Eventually, they may internalize these behavioral patterns and come to view themselves as weak and submissive. In other words, consistent peer-perceptions may eventually lead to the development of a congruent self-view.

Note, however, that this explanation leads, in turn, to further and currently unresolved questions. First, how and why do these stereotypic conceptions originally develop? Second, why do only certain traits show this effect—that is, why does Extraversion show a significant self–peer convergence correlation, whereas Emotional Stability does not?

An alternative (and not mutually exclusive) explanation is that there is an intrinsic biological link between physical appearance and personality. If so, then strangers may correctly use aspects of physical appearance that are natural, biologically based correlates of personality. Sheldon and his colleagues (Sheldon & Stevens, 1942; Sheldon, Stevens, & Tucker, 1940), for example, proposed a constitutional theory of personality in which various body types were differentially associated with certain personality traits and temperamental variables. Specifically, they argued that endomorphs (who are plump and have an underdeveloped musculature) tend to be relaxed and easygoing, whereas ectomorphs (thin and delicate) are apprehensive and inhibited, and mesomorphs (athletic build) are bold, energetic, and assertive. Sheldon and others have published evidence supporting this link between somatotype and personality (see Lindsey, 1967; Rees, 1968); however, the data are all correlational, and thus do not address the ultimate question of *why* physique and personality are related.

Again, it must be emphasized that the available data offer no compelling basis for choosing among these differing interpretations. Furthermore, as noted before, these views are not necessarily mutually exclusive, and different explanations may underlie the accurate prediction of different traits. The important point is that identifying the sources of validity in strangers' trait ratings—and isolating their ultimate explanation(s)—is an important goal for future investigation. It is hoped that the data presented here will stimulate further research along these lines.

References

Albright, L., Kenny, D. A., & Malloy, T. E. (1988). Consensus in personality judgments at zero acquaintance. *Journal of Personality and Social Psychology, 55,* 387–395.

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycholexical study. *Psychological Monographs, 47* (1, Whole No. 211).

Anderson, S. M., & Bem, S. L. (1981). Sex typing and androgyny in dyadic interaction: Individual differences in responsiveness to physical attractiveness. *Journal of Personality and Social Psychology, 41,* 74–86.

Berry, D. S., & Brownlow, S. (in press). Were the physiognomists right? Personality correlates of facial babyishness. *Personality and Social Psychology Bulletin.*

Berry, D. S., & McArthur, L. Z. (1985). Some components and consequences of a babyface. *Journal of Personality and Social Psychology, 48,* 312–323.

Cattell, R. B. (1945). The principal trait clusters for describing personality. *Psychological Bulletin, 42,* 129–161.

Cattell, R. B. (1946). *The description and measurement of personality.* New York: World Book.

D'Andrade, R. G. (1965). Trait psychology and componential analysis. *American Anthropologist, 67,* 215–228.

Digman, J. M., & Inouye, J. (1986). Further specification of five robust factors of personality. *Journal of Personality and Social Psychology, 50,* 116–123.

Digman, J. M., & Takemoto-Chock, N. K. (1981). Factors in the natural language of personality: Re-analysis and comparison of six major studies. *Multivariate Behavioral Research, 16,* 149–170.

Epstein, S. (1980). The stability of behavior: II. Implications for psychological research. *American Psychologist, 35,* 790–806.

Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *Journal of Abnormal and Social Psychology, 44,* 329–344.

Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin, 101,* 75–90.

Funder, D. C., & Colvin, C. R. (1988). Friends and strangers: 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.

Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 2, pp. 141–165). Beverly Hills, CA: Sage.

Hakel, M. D. (1969). Significance of implicit personality theories for personality research and theory [Summary]. *Proceedings of the 77th Convention of the American Psychological Association, 4,* 403–404.

Hogan, R. T. (1983). A socioanalytic theory of personality. In M. Page (Ed.), *1982 Nebraska Symposium on Motivation* (pp. 55–89). Lincoln: University of Nebraska Press.

Jackson, D. N., Neill, J. A., & Bevan, A. R. (1973). An evaluation of forced-choice and true-false item formats in personality assessment. *Journal of Research in Personality, 7,* 21–30.

Kenrick, D. T., & Stringfield, D. O. (1980). Personality traits and the eye of the beholder: Crossing some traditional philosophical boundaries in the search for consistency in all of the people. *Psychological Review, 87,* 88–104.

Lindsey, G. (1967). Behavior and morphological variation. In J. N. Spuhler (Ed.), *Genetic diversity and human behavior* (pp. 227–240). Chicago: Aldine.

McCrae, R. R., & Costa, P. T., Jr. (1985). Updating Norman's "adequate taxonomy": Intelligence and personality dimensions in natural language and in questionnaires. *Journal of Personality and Social Psychology, 49,* 710–721.

McCrae, R. R., & Costa, P. T., Jr. (1986). Clinical assessment can benefit from recent advances in personality psychology. *American Psychologist, 41,* 1001–1002.

McCrae, R. R., & Costa, P. T., Jr. (1987). Validation of a five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52,* 81–90.

Mischel, W. (1968). *Personality and assessment.* New York: Wiley.

Noller, P., Law, H., & Comrey, A. L. (1987). Cattell, Comrey, and Eysenck personality factors compared: More evidence for the five robust factors? *Journal of Personality and Social Psychology, 53,* 775–782.

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.

Norman, W. T. (1969). "To see oursels as ithers see us!": Relations among self-perceptions, peer-perceptions, and expected peer-perceptions of personality attributes. *Multivariate Behavioral Research, 4,* 417–433.

Norman, W. T., & Goldberg, L. R. (1966). Raters, ratees, and randomness in personality structure. *Journal of Personality and Social Psychology, 4,* 681–691.

O'Sullivan, M., Ekman, P., Friesen, W., & Scherer, K. (1985). What you say and how you say it: The contribution of speech content and voice quality to judgments of others. *Journal of Personality and Social Psychology, 48,* 54–62.

Passini, F. T., & Norman, W. T. (1966). A universal conception of personality structure? *Journal of Personality and Social Psychology, 4,* 44–49.

Rees, L. (1968). Constitutional psychology. In D. L. Sills (Ed.), *International encyclopedia of the social sciences* (pp. 66–76). New York: Macmillan.

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin, 94,* 18–38.

Rushton, J. P., Jackson, D. N., & Paunonen, S. V. (1981). Personality: Nomothetic or idiographic? A response to Kenrick and Stringfield. *Psychological Review, 88,* 582–589.

Sheldon, W. H., & Stevens, S. S. (1942). *The varieties of temperament: A psychology of constitutional differences.* New York: Harper.

Sheldon, W. H., Stevens, S. S., & Tucker, W. B. (1940). *The varieties of human physique: An introduction to constitutional psychology.* New York: Harper.

Snyder, M., Tanke, E., & Berscheid, E. (1977). Social perception and interpersonal behavior: On the self-fulfilling nature of social stereotypes. *Journal of Personality and Social Psychology, 35,* 656–666.

Tupes, E. C., & Christal, R. E. (1961). Recurrent personality factors based on trait ratings (Tech. Rep. Nos. 61–67). Lackland, TX: U.S. Air Force Aeronautical Systems Division.

Weiss, D. S. (1979). The effects of systematic variations in information on judges' descriptions of personality. *Journal of Personality and Social Psychology, 37,* 2121–2136.

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment.* Reading, MA: Addison-Wesley.

Received July 8, 1988
Revision received December 9, 1988
Accepted January 11, 1989 ■