The Effect of Acquaintanceship on the Validity of Personality Impressions: A Longitudinal Study

Delroy L. Paulhus and M. Nadine Bruce
University of British Columbia
Vancouver, British Columbia, Canada

Previous studies have used cross-sectional designs to demonstrate the beneficial effect of acquaintanceship on the validity of personality impressions. To counter critiques of those studies, a longitudinal design was used. Participants were randomly assigned to 16 groups of 5–7 members who met once a week for 7 weeks. None of the participants in any group were previously acquainted. Before the first meeting, they completed a battery of self-report measures, including the NEO Five Factor Inventory and the revised Interpersonal Adjective Scales. After Weeks 1, 4, and 7, group members rated each other on single-item measures related to each of the Big Five. All correlations between self-reports and corresponding peer ratings (i.e., validities) were significant by Week 7. The mean Big Five validity increased significantly from .21 to .26 to .30 at Weeks 1, 4, and 7, respectively. Extraversion showed the highest validity and consensus.

It seems intuitively compelling that the validity of person perception should increase with acquaintanceship. Indeed, a string of reputable studies have presented evidence to that effect (Cloyd, 1977; Colvin & Funder, 1991; Funder & Colvin, 1988; Jackson, Neill, & Bevan, 1973; Norman & Goldberg, 1966; Paunonen, 1989; Taft, 1966; Watson, 1989). For example, Jackson, Neill, and Bevan (1973) asked dormitory residents to complete the Personality Research Form (PRF) and then rate one another on corresponding rating scales. The students also rated their degree of acquaintance with each target. Results showed that the personality ratings had higher validities[1] when raters were better acquainted with the target.

More recently, Paunonen (1989) used the PRF to examine the effects of both acquaintanceship and trait observability on accuracy using the PRF. Results again showed a main effect for acquaintanceship. He also found an interaction of acquaintanceship with trait observability; that is, observable traits were more valid, but only for low to moderately acquainted dyads.

Funder and Colvin (1988) also found a clear-cut acquaintanceship effect. A total of 157 undergraduates who completed Q sorts were also Q sorted by both friends and strangers. Results showed significantly higher validities for friends than for strangers (within-target t test). Extending this work, Colvin and Funder (1991) again found that acquaintances predicted self-ratings better than strangers did, although acquaintances and strangers were equally accurate at predicting behavior.

Big Five and Circumplex Dimensions

There is a growing consensus that five relatively orthogonal trait dimensions (the Big Five) form the core of trait psychology. Common labels for the five domains are *Extraversion, Agreeableness, Conscientiousness, Neuroticism,* and *Openness to Experience* (or *Culture*). Evidence for the convergent and discriminant validity of these traits is substantial (e.g., Costa & McCrae, 1989; John, 1990; Wiggins & Trapnell, in press).

Two of the Big Five—Extraversion and Agreeableness—are considered interpersonal in nature. This same plane can also be marked by the two axes of the interpersonal circumplex, commonly labeled *Dominance* and *Nurturance* (Wiggins, 1979). These two axes are considered by many observers to be the fundamental dimensions of personality (e.g., Hogan, 1983; McAdams, 1984). Rather than competing with the Big Five, Dominance and Nurturance may be considered slight rotations of Extraversion and Agreeableness, respectively (McCrae & Costa, 1989b; Trapnell & Wiggins, 1990).

Given the importance of the Big Five and the two circumplex traits, verification of the acquaintanceship effect on these traits

This work was supported in part by a grant to Delroy L. Paulhus from the Social Sciences Research Council of Canada. We thank Peter Borkenau, Rebecca Collins, Randy Colvin, David Funder, David Kenny, Jeff McCrae, Dan Perlman, Paul Trapnell, David Watson, Jerry Wiggins, and two anonymous reviewers for comments on an earlier draft. We also thank Elaine McKay, Steve Moon, and Georgia Stavridis for assistance in conducting ratings.

Correspondence concerning this article should be addressed to Delroy L. Paulhus, Department of Psychology, University of British Columbia, 2136 West Mall, Vancouver, British Columbia, Canada V6T 1Z4.

[1] Note the working definition of *validity* (i.e., *accuracy*) in these studies: A peer rating is valid to the extent that it correlates with an independently obtained self-report. Obviously, there are other possible definitions of *validity* (Funder, 1987).

Journal of Personality and Social Psychology, 1992, Vol. 63, No. 5, 816–824
 0022-3514/92/$3.00

seems critical. We could find no studies of acquaintanceship on the circumplex traits. The only two relevant studies on the Big Five do provide tentative support (Norman & Goldberg, 1966; Watson, 1989).

In their classic article, Norman and Goldberg (1966) examined the correlations between self-report and peer ratings (i.e., the peer validities) for each of the Big Five traits. They reported higher Big Five validities for acquaintances (M = .40) than for strangers (M = .29). Watson (1989) pointed out, however, that Norman and Goldberg did not actually test the acquaintance–stranger differences for significance: In fact, none of the Big Five acquaintance–stranger differences are significant at the .05 level using two-tailed tests.

Nevertheless, in his own study of acquaintanceship and the Big Five, Watson also found stranger validities (M = .24) that were lower than the acquaintance validities (M = .43) reported in earlier studies. Note that neither Norman and Goldberg (1966) nor Watson (1989) actually collected acquaintance data; instead, they used previously published data to make between-subjects comparisons.

Indeed, all previous studies have used some form of cross-sectional design to examine the effects of acquaintanceship. In most of these studies, judges rated their own degree of acquaintance with targets. In such designs, unfortunately, many extraneous factors are necessarily correlated with acquaintanceship (e.g., self-selected involvement with target, target scrutability, and assimilative projection). Surprisingly, we could find no studies examining changes in validity by testing the same subjects over time. Clearly, a firmer conclusion about the acquaintanceship effect could be drawn from such a longitudinal study.

Therefore, the present study was designed to track the validities of Big Five peer ratings during the course of increasing acquaintanceship. Once a week for 7 weeks, students in an undergraduate personality class met in discussion groups designed to facilitate getting acquainted. Discussion topics were selected to bring out a wide range of personality differences.

After Meetings 1, 4, and 7, participants rated each other on single-item ratings related to the Big Five. In addition, after Week 7, students wrote one-page free descriptions of each member of their group. This procedure permitted us to test the ability of group members to judge personality as acquaintanceship increased.[2] We were also able to compare two criteria for final impressions, namely peer ratings and peer free descriptions.

Hypotheses

Peer Ratings

On the basis of the above literature review, several hypotheses were advanced. First, self-reports should predict peer ratings of the five major personality dimensions; that is, by Week 7, all peer validities should be significant (*Hypothesis 1*). Given that sociability-related traits have consistently been found to be the easiest to judge (Funder & Colvin, 1988; Kenrick & Stringfield, 1980; Norman & Goldberg, 1966; Park & Judd, 1989), validities should be highest for Extraversion and Dominance (*Hypothesis 2*). Given the literature showing substantial validities after minimal acquaintance (Berry, 1990; Norman & Goldberg, 1966; Watson, 1989), the validities for the most judgable traits (Extraversion and Dominance) should be significant even after only one meeting (*Hypothesis 3*). Following Paunonen (1989), however, we predicted that the validities of the less judgable traits would catch up as acquaintanceship increased (*Hypothesis 4*).

Our greatest interest lay in the improvement in validities over time. On the basis of the earlier cross-sectional studies showing that validity increases with acquaintanceship, we predicted that ratings at Week 7 would be more valid than ratings at Week 1 (*Hypothesis 5*).

Initially, raters must rely largely on stereotypic judgments and implicit personality inferences (Albright, Kenny, & Malloy, 1988; Weiss, 1979). As acquaintanceship increases, however, peers should use behavioral cues about the Big Five. These cues should enhance validity and hence yield more orthogonal ratings. Accordingly, we predicted increasing orthogonality of the peer-rated Big Five (*Hypothesis 6*).

Predictions about rater consensus (interrater correlations within groups), however, are more problematic than predictions about accuracy. Increasing the available information raises consensus about targets in some studies (e.g., Funder & Colvin, 1988) but lowers consensus in others (e.g., Weiss, 1979). In principle, Kenny's (1991) model permits prediction of when consensus increases with acquaintanceship; in our study, unfortunately, we could not estimate several of the required parameters. Therefore, we declined to make such predictions.

Free Descriptions

Expert raters rated the Big Five on the basis of behavioral descriptions provided by group members. Apart from their general training as personality psychology students, these experts were given specific training in how to draw inferences about the Big Five traits. Hence, experts should provide more discriminating and more accurate trait inferences. Specifically, we predicted that expert ratings of the Big Five should be more orthogonal than peer ratings (*Hypothesis 7*). Moreover, expert ratings should show higher validities than peer ratings (*Hypothesis 8*).

Method

Subjects

Participants were 89 students enrolled in a third year summer course in personality psychology. They included 55 women and 34 men ranging in age from 19 to 45, with the majority between 20 and 24.

Procedure

During the first week of class, that is, before the first group meeting, participants completed a battery of self-report measures. During the second week of classes, participants were randomly assigned to 1 of 16 heterogeneous groups on the basis of the age, gender, race, and academic interests of the participant. Each group consisted of 5–7 individuals who were previously unacquainted.

---

[2] Seven brief meetings hardly simulates long-term friendship. Nevertheless, this arrangement does permit the study of increasing acquaintanceship without many of the contaminants of long-term friendship (e.g., loyalty, exclusivity, and positivistic distortion).

Before the first group meeting, participants were given the following overview of the weekly procedure: (a) Discussion topics would be assigned for each week's meeting; (b) each meeting would be approximately 20 min long; (c) participants had to complete the assigned rating tasks as soon after each meeting as possible; (d) completed rating forms were to be sealed in the envelopes provided and returned to the instructor before the following meeting; (e) all exchanges that occurred within the meetings were to be kept in strict confidence; and (f) participants were to refrain as much as possible from socializing with other group members outside the group meeting times.[3] Groups then proceeded to a preassigned meeting area.

Discussion topics were chosen to meet three specific objectives: (a) to parallel course topics, (b) to encourage group discussion and debate, and (c) to create situations eliciting behaviors relevant to each of the Big Five. The topics were as follows: (a) early memories, (b) conceptions of adjustment, (c) types of intelligence, (d) controversial social issues, (e) sources of anxiety, (f) positive and negative qualities, and (g) fantasy and creativity.[4]

To permit assessment of changes in personality impressions as well as final impressions, we had peers rate group members on the Big Five at three times: once after the first week's discussion, once again at the midpoint after Week 4, and once more after the final meeting, at Week 7. Participants who missed a meeting were asked to complete the relevant rating forms before the next meeting (on the basis of their impressions to that point).[5]

A week after the final ratings, participants wrote free descriptions of each group member as a take-home project. They were limited to one page per member and had to justify conclusions with specific behaviors.

Self-reports. The most widely used measures of the Big Five were developed by Costa and McCrae (1989). Both the 60-item NEO Five Factor Inventory (NEO-FFI), and the 181-item NEO Personality Inventory (NEO-PI), have been extensively validated (e.g., McCrae & Costa, 1983, 1987). Because of its shorter length, the NEO-FFI was chosen as the self-report inventory to measure the Big Five in this study.

To assess the circumplex traits, the revised Interpersonal Adjective Scales were used (IAS-R; Wiggins, Trapnell, & Phillips, 1988). Sixty-four adjectives are rated on 8-point scales ranging from *extremely inaccurate* to *extremely accurate.* This measure permits the scoring of eight octants of interpersonal traits: These include assured–dominant (PA), which we term *Dominance,* and warm–agreeable (LM), which we term *Nurturance.*

Peer ratings. After Weeks 1, 4, and 7, participants rated one another on five unipolar scales. All five were scored on 10-point scales with endpoints labeled *very low* and *very high.* Each was designed to tap one of the Big Five dimensions. To help clarify the construct, all (but one) adjective labels were followed by two related adjectives. In the usual Big Five order, the labels were *assertiveness (vocal, dominant), prosocial orientation (cooperative, likable), work orientation (deliberate, organized), insecurity,* and *intellect (original, clever).* Note that these are not the usual labels for the Big Five; the three-adjective combinations were chosen to include various common terms related to these constructs. Note also that the first two ratings are also general enough to provide validity criteria for the two circumplex traits (Dominance and Nurturance).

Free descriptions. Following the last discussion group, all participants were asked to prepare a project in which they described their final impressions of each group member and the behavioral evidence they used as the basis for these conclusions.

Two teams of two raters were trained to code Big Five descriptions. Half of the projects were coded by one team and half by the other. Each description of each member was rated separately on the Big Five. Ratings were made on 7-point Likert scales with endpoints labeled *low* and *high.*

As a reliability check before beginning to code the free descriptions, a random set of five descriptions was chosen for each dimension and coded by all four raters as well as by a personality psychologist (Delroy L. Paulhus). The interrater reliabilities on this small set ranged from .80 to .90—acceptable enough to proceed with coding (see final reliabilities below).

Results

The descriptive statistics for the self-report scales are listed in Table 1. Means, standard deviations, and alpha reliabilities of the NEO-FFI scales are in keeping with a large student sample collected by Trapnell and Wiggins (1991). The statistics on the circumplex scales are also similar to the student norms reported by Wiggins et al. (1988).

The intercorrelations among the self-reports are provided in Table 2. Although relatively small, all intercorrelations among the Big Five are positive except those with Neuroticism. The mean (absolute) correlation is .18. The highest correlation among the Big Five (−.37) is that between Conscientiousness and Neuroticism.

Peer Ratings

Minimal sex differences were observed in the analyses below. Hence the data for men and women were pooled.

Consensus. Table 3 presents several forms of interrater reliability for each Big Five dimension. The first three data columns contain a measure of consistency within groups: the mean interrater Pearson correlation. Each tabled value is the mean across 16 groups of the mean interrater correlation within each group. Consensus did not increase over time: Indeed, consensus was highest—.41 averaged across all five dimensions—in Week 1. Averaged over the three occasions, consensus was greatest on ratings of Extraversion (.61) and least for ratings of Openness (.25).

Reliability of means. The last two columns in Table 3 provide two forms of reliabilities for the final (Week 7) mean. Both are reliabilities of the Week 7 mean rating stepped up for 4–6 raters. The fourth data column is based on the final Pearson correlations. The fifth data column is the intraclass correlation (Shrout & Fleiss, 1979, Type I) calculated across all 89 participants. This value is the appropriate index for evaluating the reliability of the mean peer rating, that is, the criterion for calculating the final validities.[6]

---

[3] This restriction was required to control the amount of exposure among participants. Without such a restriction, factors such as the participant's Extraversion would influence the amount of contact with other participants.

[4] Note that the weekly topics do not map one-to-one with the Big Five. The precise instructions for each discussion meeting may be obtained from Delroy L. Paulhus by request.

[5] Unfortunately, this procedure leads to a conservative estimate of validities for such participants. Fortunately, it minimizes problems with missing data.

[6] The Type I intraclass correlation (Shrout & Fleiss, 1979) is actually an underestimate of the reliability of the means in these data. This index is appropriate for cases in which each target is rated by a different judge. Here, participants in same groups are rated by the same judges, although those in different groups are rated by different judges. Nonetheless, we opted for the conservative estimate of this parameter.

Table 1
Descriptive Statistics for Self-Report Scales

Trait	*M*	*SD*	α
Big Five			
Extraversion	30.4	6.7	.80
Agreeableness	32.0	6.7	.79
Conscientiousness	30.7	6.9	.83
Neuroticism	24.2	9.3	.88
Openness	30.7	6.2	.70
Circumplex			
Dominance	42.0	7.2	.81
Nurturance	49.8	6.1	.88

Note. These are unweighted means controlled for sex. Big Five traits were scored from the NEO Five Factor Inventory. Possible scores range from 0 to 48. Circumplex traits were scored from the revised Interpersonal Adjective Scales. Possible scores range from 8 to 64.

Prediction of Peer Ratings: Peer Validities

Correlations of the self-report scales with the appropriate peer-rated criteria[7] (i.e., the validities) are presented in Table 4.

Confirming Hypothesis 1, all seven individual validities were significant at Week 7. Even the lowest validity—Neuroticism ($r = .18$)—was significant. The two most judgable traits, Extraversion ($r = .40$) and Dominance ($r = .51$), had the highest validities, confirming Hypothesis 2. Both of these judgable traits were significant at Week 1, confirming Hypothesis 3. In fact, only Agreeableness and Nurturance were not significant at Week 1.

Hypothesis 4 was not confirmed because, at Week 7, the mean validity of the two judgable traits (.46) remained significantly higher than the mean (.26) for the less judgable traits ($Z = 3.12$, $p < .01$). It is true that, between Weeks 1 and 7, the judgable trait validities did increase less than those of the less judgable traits (.05 vs. .11).[8] A closer look, however, reveals that only Agreeableness and Nurturance showed any substantial increase.

Of special note, the mean validity at Week 7 ($M = .30$) was significantly higher than that at Week 1 ($M = .21$), $t(86) = 2.13$, $p < .02$.[9] Hence, Hypothesis 5, the central hypothesis of the study, was also supported.

Orthogonality. The intercorrelations among the final peer ratings are presented in Table 5. The pattern is similar to that in previous rating studies; that is, all intercorrelations except those with Neuroticism are positive. Unfortunately, the absolute values of the intercorrelations are disconcertingly high, ranging from .35 to .67 ($M = .50$). It is clear from Table 5 that the major source of overlap is Extraversion, which correlates highly with all other ratings.

Nonetheless, Hypothesis 6 was supported in that the mean intercorrelation showed a gradual decline from Week 1 (.66) to Week 4 (.57) to Week 7 (.50). The value at Week 1 was significantly higher than that at Week 7, $Z = 2.71$, $p < .01$ (Steiger, 1980).

Discriminant validity. The high intercorrelation of the peer ratings noted above raises questions about their discriminant validity. Accordingly, correlations of all self-report scales with all final peer ratings are displayed in Table 6. Note that the final validities from Table 4 appear on the diagonal.

With the exception of Neuroticism, the highest value in each row is with the corresponding rating criterion. Thus, some discrimination is evident in the validities. On the other hand, less discrimination is evident down the columns, indicating that the criterion ratings were conceptually too broad.

In general, this overlap in peer ratings does not compromise the Big Five validities; that is, if orthogonal predictors are all significant, then they must be predicting independent parcels of variance. Unfortunately, there is some overlap among the predictors, primarily originating from Extraversion and Neuroticism.

To provide an estimate of the independent contributions of the self-report predictors, we included all five predictors in regression analyses of each Big Five criterion. Table 7 shows the resulting pattern of first-order and partial beta coefficients. The partial betas represent the validity of each predictor with the other four predictors partialed out. The mean beta decreased as expected from .30 to .24, but four of five remained significant. Only the weakest of the partial betas, Neuroticism, virtually disappeared (.01). When the discriminant predictors were similarly controlled in the Week 1 and Week 4 validities, the drop was more dramatic: The partial beta coefficients averaged only .10 and .15, respectively. This finding is consistent with the above finding that the final ratings were more orthogonal than were the earlier ratings. In short, the final validities were not only higher in value but also more discriminating.

Expert Ratings of Free Descriptions

Recall that participants also wrote free descriptions of all their group members. Each description was later coded on the Big Five dimensions by two expert raters. As indexed by Pearson correlations, the interrater reliabilities[10] were all quite adequate: Extraversion (.89), Agreeableness (.86), Conscientiousness (.86), Neuroticism (.79), and Openness (.80). Hence, to provide a more reliable estimate, we took the mean of the two ratings for each dimension. The intercorrelations among the Big Five are given in Table 8. It is worthwhile to compare these relations with those among the peer ratings (i.e., Table 5). Confirming Hypothesis 7, the mean intercorrelation (absolute value) was significantly lower among the expert ratings ($M =$

[7] Note that there are no specific peer ratings for Dominance and Nurturance. Instead, their validities are calculated using the Extraversion and Agreeableness criteria, which were broad enough to be appropriate targets.

[8] We know of no appropriate statistical test for the difference between two changes in correlations.

[9] The mean validity is calculated only on the Big Five to permit comparisons with other studies. Also, if the two circumplex scales were included, then two of the Big Five would be doubly represented.

[10] Because means and standard deviations of the two rating teams were very similar, we calculated Pearson reliabilities across all 89 targets and stepped up these values to represent the reliability of the mean. Pearson correlations are appropriate here because no linear transformation of a rater's scores can affect the mean rating such that calculation of validity is affected.

Table 2
Intercorrelations of Self-Reports

Trait	1	2	3	4	5	6	7
Big Five							
1. Extraversion	—						
2. Agreeableness	.29**	—					
3. Conscientiousness	.11	.15	—				
4. Neuroticism	−.28**	−.14	−.37***	—			
5. Openness	.18*	.02	.03	−.19*	—		
Circumplex							
6. Dominance	.33***	−.30**	.39***	−.43***	.14	—	
7. Nurturance	.18*	.52***	.06	.10	.17	−.30**	—

Note. Scales 1–5 are from the NEO Five Factor Inventory; Scales 6–7 are from the revised Interpersonal Adjective Scales.
* $p \leq .05$, two-tailed. ** $p \leq .01$. *** $p \leq .001$.

.32) than among the peer ratings ($M = .50$), $Z = 4.52$, $p < .001$ (see Steiger, 1980). Hence, the expert ratings exhibited improved discrimination among the Big Five. The singular exception is a strong correlation remaining between Extraversion and Neuroticism ($r = -.67$).

The obtained validities were as follows: Extraversion, .38; Agreeableness, .32; Conscientiousness, .20; Neuroticism, .19; and Openness, .37; Overall, the validities for expert ratings ($M = .29$) and peer ratings ($M = .30$) were very similar; hence, Hypothesis 8 was not confirmed.

Discussion

This study provides clear-cut evidence that the validity of personality impressions increases with acquaintanceship. Our data confirm longitudinally the result found previously with between-subjects indicators of acquaintanceship.

As cautioned by Jackson et al. (1973), between-subjects studies yield ambiguous evidence for the acquaintanceship effect. For example, instead of indicating little time spent with the target, low acquaintanceship ratings may simply indicate that a target is difficult to judge; thus, the low acquaintanceship rating and the low validity both follow from the fact that this person is difficult to get to know (Colvin & Funder, 1991). Our longitudinal design rules out this artifact and other confounds by comparing ratings of the same targets over time.

The fact that the acquaintanceship effect was confirmed on the Big Five and circumplex traits is a further contribution. It is reassuring that the acquaintanceship effect holds for traits from the two most important structural models in contemporary personality assessment. A study using any other traits could have been criticized for not testing the fundamental personality traits.

The mean validity was significant even after the first meeting. This finding should not be surprising given that the so-called "zero acquaintance" studies have found substantial validities with far less than the 20 min of interaction our study provided at Week 1 (e.g., Albright et al., 1988, Watson, 1989). Nonetheless, the fact that some of our validities were already near ceiling at Week 1 worked against our attempt to demonstrate an increase over the 7 weeks. Presumably, a comparison with "absolute zero acquaintance" (no information) would have been more impressive.

Validity Versus Consensus

Although no statistical verification was possible, it appeared that group consensus was actually highest at Week 1. Note,

Table 3
Rater Reliabilities

	Group consensus			Reliability of *M*	
Trait	Week 1	Week 4	Week 7	Pearson *r*	Intraclass correlation
Extraversion	.68***	.58***	.56***	.86***	.75***
Agreeableness	.29**	.30**	.30**	.68***	.46***
Conscientiousness	.33***	.34***	.19*	.57***	.55***
Neuroticism	.43***	.28**	.31**	.69***	.34***
Openness	.30**	.21*	.23*	.60***	.61***
M	.41***	.34***	.32**	.68***	.54***

Note. Entries in the first three data columns are means across 16 groups of the mean intercorrelation among two to six ratings of the same individual. The final two columns are reliabilities based on the mean of four to seven raters at Week 7.
* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

Table 4
Peer Validities for Seven Traits Over Time

Trait	Week 1	Week 4	Week 7
Big Five			
Extraversion	.35***	.28**	.40***
Agreeableness	.01	.25*	.31**
Conscientiousness	.29**	.25*	.33***
Neuroticism	.25*	.10	.18*
Openness	.27**	.35***	.31**
Big Five *M*	.21*	.26**	.30**
Circumplex			
Dominance	.45***	.50***	.51***
Nurturance	.07	.05	.18*

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

however, that the mean Week 1 validity was a mere .10 after other predictors were partialed out. The explanation seems to be that all the Week 1 ratings were highly confounded with extraversion. In contrast, the final ratings showed considerable discriminant validity.

The increasing validity over time may seem to conflict with decreasing consensus, although Kenny's (1991) model shows how this can occur (p. 159). In our data, we take this pattern to mean that the high initial validity and consensus derived partly from stereotypic inferences (Weiss, 1979). That is, group members initially used extraversion and common implicit personality theories to arrive at similar judgments. With increasing exposure to the targets, these judgments became more idiosyncratic; at the same time, because they were behavior based, the judgments became more valid.

Underestimates?

Although our final validities were reasonable, the mean figure ($M = .30$) is still somewhat less than the mean peer validities reported by Costa and McCrae (1989) in the NEO-FFI manual ($M = .40$). There are several reasons why our figures are likely to be conservative estimates of the "true" validities in such discussion groups.

The most obvious factor is the unreliability ensuing from our peer-rating scales. We asked peers to make a single global rating on a (amplified) trait descriptor related to each Big Five dimension. By contrast, Costa and McCrae (1989) asked their raters to complete the full 181-item NEO-PI comprising an average of 36 items per factor. Indeed, most peer-rating studies of accuracy have used multiple ratings (exceptions are Albright et al., 1988; Jackson et al., 1973). Consider first the obvious benefits of aggregation for reliability: Tripling the number of items per factor would boost our mean validity from a mean of .31 to an expected .40 (using the Spearman–Brown formula). In addition, the use of multiple indicators would permit the extraction of factor scores: The latter provide the ideal in reliable criteria (Costa & McCrae, 1989).

Second, it should be noted that the Costa and McCrae (1989) validity figures resulted from the use of validimax factor analysis, a rotational procedure designed to maximize validities (McCrae & Costa, 1989a). In the present study, by contrast, neither the self-reports nor the peer ratings were factored; nor were they rotated to fit. Consequently, our validities depended critically on our wording of the single rating scale. As noted below, this wording did affect our results.

Third, it is likely that our observed validities would have continued to rise with increased acquaintance. After all, seven meetings of 20 min each represent a total contact time with the group of only 2 hr and 20 min. Given that each participant held the floor for only a fraction of that time, the information gathered per member could only be termed modest.

Overestimates?

On the other hand, by tailoring conversation topics to highlight behaviors relevant to each of the Big Five, we may have overestimated participants' ability to distinguish these facets of personality in typical group settings. Ordinarily, surgency-related traits (Extraversion–Dominance) might overwhelm less salient personality variables in leaderless and otherwise unstructured groups. We argue that the discussion topics are not unusual for group discussions, at least in academia. Admittedly, such a full range of topics is seldom found on the agenda of the same group.

Trait Specificity

Apart from these general effects, several dimensions warrant individual comment. Of the Big Five, Extraversion consistently showed the highest interrater consensus and the highest validities. This finding has often been replicated (e.g., Funder & Colvin, 1988; Kenrick & Stringfield, 1980; Norman & Goldberg, 1966; Park & Judd, 1989).

The highest peer validity of all was the .51 prediction of the extraversion criterion from the IAS-R Dominance scale. A look at the criterion descriptor—*assertiveness (vocal, dominant)*—reveals an ideal correspondence between predictor and criterion. Dominance was also more narrow banded than Extraversion in predicting across the five peer ratings; that is, the mean cross validities were .19 and .29 for Dominance and Extraversion, respectively. Indeed, these two measures were specifically designed to tap narrow and broad traits, respectively (Wiggins & Trapnell, in press).

Note that the superiority of the IAS-R Dominance scale over the NEO-FFI Extraversion scale cannot be attributed to the inherent reliability of a longer scale. On the contrary, the Dominance scale comprises only eight adjectives, whereas the NEO-FFI Extraversion scale contains 12 questionnaire statements.

Table 5
Intercorrelations of Final Peer Ratings

Trait	1	2	3	4	5
1. Extraversion	—				
2. Agreeableness	.54	—			
3. Conscientiousness	.57	.38	—		
4. Neuroticism	−.67	−.44	−.42	—	
5. Openness	.66	.35	.45	−.55	—

Note. All correlations are significant at $p = .001$, two-tailed.

Table 6
Correlations of All Self-Reports With Final Peer Ratings

Self-reports	Peer ratings 1	2	3	4	5
Big Five traits					
1. Extraversion	**.40*** **	.31**	.22*	−.34***	.20*
2. Agreeableness	−.02	**.31****	.27**	.01	.10
3. Conscientiousness	.27**	.06	**.33*****	−.17	.20*
4. Neuroticism	−.33***	−.11	−.20*	**.18***	−.14
5. Openness	.31**	.22*	.28**	−.31**	**.31****
Circumplex traits					
Dominance	**.51*****	.16	.10	−.37***	.15
Nurturance	−.06	**.18***	.15	−.02	.16

Note. The Big Five scales are scored from the NEO Five Factor Inventory; the circumplex scales are scored from the revised Interpersonal Adjective Scale. Final validities are in boldface.
* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

Apparently, after a single meeting, group members were able to grasp with precision each others' level of extraversion. The ease of judging extraversion may actually place a ceiling on possible improvement with acquaintanceship (Kenny, 1991; Paunonen, 1989). Recent studies suggest that this high judgability of Extraversion derives from its visual and auditory cues (Berry, 1990; Borkenau & Liebler, 1992; Warner & Sugarman, 1986).

The NEO-FFI Agreeableness scale outperformed the IAS-R Nurturance scale (validities were .31 vs. .18) in predicting our Agreeableness criterion rating, *prosocial orientation (cooperative, likable)*. Across time, the Agreeableness validities were .01, .25, and .31 at Weeks 1, 4, and 7, respectively, showing marked improvement. The Nurturance validities also showed substantial improvement over the three ratings (.07, .05, and .18). This increase over time from a null relation to significance suggests that judgments of this domain were based on the actual content of the group interactions, not from easily observed cues.

Because it is not easily observed, Neuroticism may be the most difficult Big Five trait to rate (e.g., John, 1990; Watson, 1989). Nevertheless, significant validities were evidenced for both peer and expert ratings. This validity, however, appeared to be an artifact of confusion with Extraversion: Regression analysis showed that the independent contribution of Neuroticism was minimal. Peers and experts alike seemed to have difficulty distinguishing genuine trait anxiety from quietness due to Introversion.

What exactly was it about the group situation that rendered peers unable to rate Neuroticism? There are at least two possible explanations. First, it is already known that anxiety is difficult for peers to discern, regardless of the extent of the acquaintance (John, 1990). Alternatively, because of the negative social desirability of anxiety and insecurity, participants may have actively attempted to hide their Neuroticism in the group setting.

Note that the substantial validities for Neuroticism reported in Costa and McCrae (1989) were based on spouse ratings; indeed, of the Big Five, Neuroticism had the highest validity. Apparently, one's Neuroticism can be hidden from group-discussion members but not from one's spouse.

Peer Ratings Versus Expert-Rated Free Descriptions

Although expert ratings of free descriptions were still based on peers' impressions, we hoped that experts might be less subject to certain rating biases. Expert raters could draw their own, presumably more objective, conclusions from the behaviors described by peers in their free descriptions. For example, they were instructed to disregard inferences of traits from irrelevant behavior.

Overall, we found very similar validities for the peer and expert ratings. Nonetheless, expert ratings did show some im-

Table 7
Discriminant Validities: Independent Effects of Predictors

Trait	Beta coefficients First order	Partial
Extraversion	.40	.35**
Agreeableness	.31	.25*
Conscientiousness	.33	.29**
Neuroticism	.18	.01
Openness	.31	.29**

Note. Self-reports are scored from the NEO Five Factor Inventory.
* $p \leq .05$. ** $p \leq .01$.

Table 8
Intercorrelations of Expert Ratings of Free Descriptions

Trait	1	2	3	4	5
1. Extraversion	—				
2. Agreeableness	.13	—			
3. Conscientiousness	.32*	.13	—		
4. Neuroticism	−.67**	−.16	−.32*	—	
5. Openness	.31*	.44**	.39**	−.35**	—

* $p \leq .01$, two-tailed. ** $p \leq .001$, two-tailed.

provement over peer ratings. For one thing, the orthogonality of the Big Five was significantly improved. Although they still confused introversion and neuroticism, expert raters were clearly able to distinguish between extraversion and standing on the other three dimensions.

What Is Special About Discussion Groups?

The opportunity to study personality in leaderless group discussions allowed us to track acquaintanceship and accuracy. At the same time, we have to consider the special implications of evaluating validities in this setting. How unique is this context? For example, should we ever expect the level of validity after 7 weeks to approach the validities obtained with close-friend or spouse judgments of personality? There are at least two reasons to be wary of the group context.

First, the discussion group may fundamentally alter the manifestation of traits; that is, novel (or at least select) facets of personality may appear in this situation. For example, social facilitation induced by the group presence could polarize the manifestation of certain traits. If so, we should have observed very high validities; we did not. More likely, the high demand for impression management in groups might mask the manifestation of certain traits (Paulhus, 1986). More generally, the topology of personality dynamics in groups might highlight a unique set of individual differences. If such group effects are operative, the modest validities we obtained might simply evidence the inappropriateness of Big Five self-reports for predictions of behavior emerging in this setting.

We hesitate to draw that conclusion. For one thing, the cumulative validity evidence for our self-report predictors (NEO-FFI and IAS-R) is substantial and broadly based. For example, the NEO-FFI was validated on a variety of quality criteria (spouse ratings, long-time acquaintance ratings, and behavior). Moreover, spouses and friends have observed the target in a wide variety of situations (including group situations) over an extended period of time. Nonetheless, we cannot rule out the potential value of developing self-report measures tailored to group-relevant personality.

A second reason for being wary of the evaluation context is that the discussion group may alter the perception of traits. We have already detailed some evidence to this effect. Indeed, we did find evidence for differential judgability of the Big Five traits. Because participation must mediate most manifestations of personality in discussion groups, the situation magnifies further the usual salience of Extraversion. Clearly, Extraversion outshone the others with consistently higher validity and interrater consensus. Nonetheless, the fact that our observed rank-order of the validities corresponded closely to that in previous studies reassures us that our discussion format did not reorder the importance of the Big Five dimensions.

Limitations and Recommendations

We recommend that future studies of the acquaintanceship effect continue to exploit the longitudinal design. Nonetheless, various problems raised throughout the Discussion section suggest some recommendations for the design of follow-up studies.

A study of more extended acquaintanceship is required. Although our participants met over seven occasions, the total meeting time (2 hr and 20 min) is unlikely to be sufficient to yield maximum validities. The validities rose as much in the last 3-week interval as in the first 3-week interval. More frequent or longer meetings should bring out even richer personality information and thus yield stronger validities.

Other improvements may stretch the limits of practicability for many researchers. The longer form of the NEO-PI with its more reliable scales could be used instead of the NEO-FFI. Use of more than one peer-rating scale would improve reliability (Paunonen, 1989). More participants would allow statistical confirmation of moderator effects (Chaplin, 1991). Finally, direct operationalization of various parameters theoretically linked to acquaintanceship by Kenny (1991) could provide empirical tests of his hypotheses.

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.

Berry, D. S. (1990). Taking people at face value: Evidence for the kernel of truth hypothesis. *Social Cognition, 8,* 343–361.

Borkenau, P., & Liebler, A. (1992). Trait inferences: Sources of validity at zero acquaintance. *Journal of Personality and Social Psychology, 62,* 645–657.

Chaplin, W. F. (1991). The next generation of moderator research in personality psychology. *Journal of Personality, 59,* 143–178.

Cloyd, L. (1977). Effect of acquaintanceship on accuracy of person perception. *Perceptual and Motor Skills, 44,* 819–826.

Colvin, C. R., & Funder, D. C. (1991). Predicting personality and behavior: A boundary on the acquaintanceship effect. *Journal of Personality and Social Psychology, 60,* 884–894.

Costa, P. T., Jr., & McCrae, R. R. (1989). *The NEO-PI/FFI manual supplement.* Odessa, FL: Psychological Assessment Resources.

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.

Hogan, R. (1983). A socioanalytic theory of personality. In M. M. Page (Ed.), *Personality: Current theory and research* (pp. 55–89). Lincoln, NE: 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.

John, O. P. (1990). The "Big Five" factor taxonomy: Dimensions of personality in the natural language and in questionnaires. In L. A. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 66–100). New York: Guilford Press.

Kenny, D. A. (1991). A general model of consensus and accuracy in interpersonal perception. *Psychological Review, 98,* 155–163.

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

McAdams, D. P. (1984). Love, power, and images of the self. In C. Malatesta & C. E. Izard (Eds.), *Emotion in adult development* (pp. 159–174). Beverly Hills, CA: Sage.

McCrae, R. R., & Costa, P. T., Jr. (1983). Joint factors in self-reports and

ratings: Neuroticism, extraversion, and openness to experience. *Personality and Individual Differences, 4,* 245–255.

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

McCrae, R. R., & Costa, P. T. (1989a). Rotation to maximize the construct validity of factors in the NEO Personality Inventory. *Multivariate Behavioral Research, 24,* 107–124.

McCrae, R. R., & Costa, P. T. (1989b). The structure of interpersonal traits: Wiggins' circumplex and the five-factor model. *Journal of Personality and Social Psychology, 56,* 586–595.

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

Park, B., & Judd, C. M. (1989). Agreement on initial impressions: Differences due to perceivers, trait dimensions, and target behaviors. *Journal of Personality and Social Psychology, 56,* 493–505.

Paulhus, D. L. (1986). Self-deception and impression management in test responses. In A. Angleitner & J. S. Wiggins (Eds.), *Self-report via questionnaire* (pp. 142–165). New York: Springer-Verlag.

Paunonen, S. V. (1989). Consensus in personality judgments: Moderating effects of target–rater acquaintanceship and behavior observability. *Journal of Personality and Social Psychology, 56,* 823–833.

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

Steiger, J. H. (1980). Tests for comparing elements of a correlation. *Psychological Bulletin, 87,* 245–251.

Taft, R. (1966). Accuracy of empathic judgments of acquaintances and strangers. *Journal of Personality and Social Psychology, 3,* 600–604.

Trapnell, P. D., & Wiggins, J. S. (1990). Extension of the Interpersonal Adjective Scales to include the Big Five dimensions of personality. *Journal of Personality and Social Psychology, 59,* 781–790.

Trapnell, P. D., & Wiggins, J. S. (1991). [Item factor analysis of the Five Factor Inventory]. Unpublished raw data, University of British Columbia, Vancouver, British Columbia, Canada.

Warner, R. M., & Sugarman, D. B. (1986). Attributions of personality based on physical appearance, speech, and handwriting. *Journal of Personality and Social Psychology, 50,* 792–799.

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.

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. (1979). A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of Personality and Social Psychology, 37,* 395–412.

Wiggins, J. S., & Trapnell, P. D. (in press). Personality structure: The return of the Big Five. In S. R. Briggs, R. Hogan, & W. H. Jones (Eds.), *Handbook of personality psychology.* San Diego, CA: Academic Press.

Wiggins, J. S., Trapnell, P. D., & Phillips, N. (1988). Psychometric and geometric characteristics of the revised Interpersonal Adjective Scales (IAS-R). *Multivariate Behavioral Research, 23,* 517–530.

Received October 1, 1991
Revision received April 27, 1992
Accepted June 11, 1992 ■