Journal of Personality and Social Psychology
1995, Vol. 69, No. 6, 1233-1242

0022-3514/95/$3.00

Enhancing Target Variance in Personality Impressions: Highlighting the Person in Person Perception

Delroy L. Paulhus and Shawn Reynolds
University of British Columbia

D. A. Kenny (1994) estimated the components of personality rating variance to be 15, 20, and 20% for target, rater, and relationship, respectively. To enhance trait variance and minimize rater variance, we designed a series of studies of personality perception in discussion groups (N = 79, 58, and 59). After completing a Big Five questionnaire, participants met 7 times in small groups. After Meetings 1 and 7, group members rated each other. By applying the *Social Relations Model* (D. A. Kenny and L. La Voie, 1984) to each Big Five dimension at each point in time, we were able to evaluate 6 rating effects as well as rating validity. Among the findings were that (a) target variance was the largest component (almost 30%), whereas rater variance was small (less than 11%); (b) rating validity improved significantly with acquaintance, although target variance did not; and (c) no reciprocity was found, but projection was significant for Agreeableness.

The study of person perception plays a pivotal role in linking personality to social psychology. After all, the rating process depends partly on the strength of the signal, that is, the true differences in personality to be detected, and partly on the social cognition of the perceiver. Both groups have contributed to the recent surge in publications on the issue (e.g., Funder & West, 1993; Trope & Higgins, 1993).

Arguably, the most important of these publications is Kenny's (1994) book *Interpersonal Perception,* which represents 15 years of work on partitioning the components of person ratings. In reviewing more than 40 such studies, Kenny came to a number of provocative conclusions. For example, he concluded that target variance, that is, observer agreement on others' personality, was modest, at best. Nor was this target variance improved by increased levels of acquaintance with the target. More important than target variance, according to Kenny, are the personal biases of the raters and the relationship between the rater and the target individual.

To us, these conclusions seemed at odds with the solid evidence for observer agreement in ratings of the Big Five personality dimensions (e.g., McCrae & Costa, 1987; Piedmont, 1994). In examining his review of the literature, we concluded that Kenny's (1994) norms may have underestimated potential target variance because they combined studies of (a) both dyadic and group ratings, (b) varying breadth of acquaintance, and (c) varying rating instructions. In the present article, we designed studies to demonstrate that higher levels of target variance could be achieved. By sharpening both measurement tools and target discriminability, we should also enhance our ability to demonstrate rating validity and the effects of acquaintance.

The Group Rating Paradigm

To this end, we collected three types of data on participants in small groups. As in most of the studies referred to in this report, the paradigm for collecting peer-rating data is the "round-robin" pattern: That is, each person is both a target and a rater of all other members. Because the validity of rating data must be indexed by their correlation with other measures of the same construct, we also included two forms of self-assessment: (a) *self-ratings* on global items (Burisch, 1984) and (b) a standardized personality test, hereafter referred to as the *questionnaire.* The layout of the three assessment modes in our studies is illustrated in Table 1.

Correlational approaches to such data focus on the relation between various target measures: for example, the rating means and questionnaire scores. In contrast, an analysis of variance (ANOVA) approach to such data would focus on the relative contributions of target and rater to the rating variance. Once partitioned, the proportion of variance explained by each source can be used as an index of its importance, and statistical tests can be applied to test hypotheses about these sources. *Target variance* is the variance across the mean ratings *received* by group members: In Table 1, it is manifested in differences across values in the *Target means* column. Higher values are primarily a function of target discriminability (i.e., the targets are perceived as distinctive).

Rater variance is variance across the mean ratings *given* by the group members. Some raters may place every target near the high end of a scale, others may prefer the low end, and still others, the middle range. In Table 1, rater variance is manifested in differences among means on the Rater means row. On evaluative dimensions, higher rater variance is primarily a function of differences in rater leniency.

Delroy L. Paulhus and Shawn Reynolds, Department of Psychology, University of British Columbia, Vancouver, British Columbia, Canada.

This research was supported by a grant from the Social Sciences and Humanities Research Council of Canada. We wish to thank Nadine Bruce and Oliver John for assistance in the data collection and David Kenny, Dan Perlman, Paul Trapnell, Jerry Wiggins, and Michelle Yik for comments on an earlier draft.

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

Table 1
Hypothetical Layout of Rating Data for a Four-Person Group

	Rater					
Target	1	2	3	4	Target means	Questionnaire
1	X(1,1)	X(1,2)	X(1,3)	X(1,4)	*T*(1)	Q1
2	X(2,1)	X(2,2)	X(2,3)	X(2,4)	*T*(2)	Q2
3	X(3,1)	X(3,2)	X(3,3)	X(3,4)	*T*(3)	Q3
4	X(4,1)	X(4,2)	X(4,3)	X(4,4)	*T*(4)	Q4
Rater means	*R*(1)	*R*(2)	*R*(3)	*R*(4)	Overall Mean	

Note. Each X(i,j) is the rating given by Rater j to Target i. For round-robin data the raters and targets are the same persons, hence the diagonal contains the self-ratings.

The remainder is residual variance; that is, variance above and beyond the effects attributable to target and rater variance. In Table 1, such variance results from deviations of observed ratings from those expected from the row and column means for that individual.

In standard ANOVA, residual variance is separated into interaction variance and error. Error variance is, by definition, unpredictable and inconsistent; in contrast, interaction variance shows up consistently across replications. Thus, to distinguish between interaction and error variance, one has to collect one indicator at several points in time or multiple indicators at one time. In round-robin data, such interactions have been labeled *relationship variance,* given that they arise from the unique relationships among participants (Kenny & La Voie, 1984).

Difficulties With Round-Robin Data

Unfortunately, round-robin data raise a host of analysis problems (Warner, Kenny, & Stoto, 1979). Because they do not satisfy the usual statistical assumptions, round-robin data cannot be analyzed with conventional forms of analysis such as ANOVA. The first problem arises from including the self-rating in the peer-rating mean for each target: Including it requires accepting the assumption that the statistical model for self-ratings is comparable to that for peer-ratings. Even in this unlikely event, most researchers would argue that the self-rating should not be included because it is qualitatively different from peer ratings.

If self-ratings are omitted, however, two other problems arise. First, reliance on the remaining peer ratings introduces a systematic bias in the peer-rating mean: Lenient raters do not get the benefit of their own rating and therefore are systematically underrated. Second, standard ANOVA is incapable of analyzing the data. One cell in each row (25% of the total cells in the Table 1 example) is missing. Incomplete block designs are necessary to analyze such data (see Kirk, 1968, pp. 427–440).

Even with such corrections, however, round-robin data are not adequately analyzed with ANOVA because the assumption of independence among levels of the independent variables is not met. Consider the possibility of *projection* operating in the data: That is, there may be an association between the kind of ratings a person gives and the kind of ratings that person receives.[1] A number of studies have found evidence for projection in personality rating data (D. T. Campbell, Miller, Lubetsky, & O'Connell, 1964; Sherwood, 1980; J. D. Campbell, 1986).

In addition, there may be a *reciprocity* in ratings between pairs of persons in a group: That is, the ratings Person A gives to B may correlate with the ratings B gives to A. For example, individuals with similar interests may react positively to one another, whereas liberals and conservatives may actively dislike one another. The operation of reciprocity in discussion groups is predictable from interpersonal theory (e.g., Kiesler, 1983) as well as from the law of similarity and attraction (Byrne, 1971). Both reciprocity and projection contravene the assumptions of independence required for an ANOVA.

Enter the Social Relations Model

The *Social Relations Model* (SRM), developed by Kenny and La Voie (1984), provides a statistical partitioning of rating data into three major sources of variance:[2] target, rater, and relationship. According to the model, a rating, X_{ijk}, comprises five components: (1) the grand mean of all ratings, (2) the mean rating that i receives, (3) the mean rating that j gives, (4) the rating deviation due to i's unique perception of j, and (5) error or instability.

To the extent that various indicators of a construct intercorrelate (i.e., the items are homogeneous), then that construct has stable variance. SRM provides estimates of the stable portion of all three effects (target, rater,[3] and relationship) that are then corrected for attenuation. By estimating only the stable portion of each effect, SRM provides potentially more valid estimates.

Kenny (1990) also developed a computer program (SOREMO) that provides a variety of analyses based on the stable component of

[1] Rather than the term *projection,* Kenny used the term *generalized reciprocity* for the correlation across persons of ratings given and ratings received. He noted that psychoanalysts reserve the former term for cases in which the person denies the rating he or she receives. We feel that the psychoanalytic usage is too restrictive and contradicts widespread current usage (see Paulhus et al., in press).

[2] Note that the SRM partitions variance in two ways—absolute and relative (proportional) variance. Unless explicitly stated, further references to variance should be understood as relative variance for the remainder of this article.

[3] We prefer the traditional terms *target and rater variance* rather than *consensus and assimilation,* which Kenny (1994) prefers. We feel that the latter terms imply within-target and within-rater variance, which in fact, play only a minor role in the formulas.

each effect. If the user includes additional measures on each participant (e.g., questionnaire scores), the program provides all the intercorrelations with these external measures. In short, the program combines ANOVA and correlational approaches.

Among many applications, SRM provides an appropriate analysis for the round-robin data described above. Of particular advantage is its facility for correcting the interdependencies found in round-robin rating data. Hence the well-established phenomena of projection and reciprocity do not distort the calculations of rater and target effects.

In fact, SRM goes further to provide indexes of projection and reciprocity. Instead of viewing these two phenomena as statistical annoyances, SRM estimates their importance and makes them available for correlations with other measures. These correlations are disattenuated according to the degree of unreliability of the peer ratings.

Another important feature of SRM is its provision for the self-rating problem discussed earlier. Recall that a systematic bias is introduced by excluding the self-rating from a target's mean peer rating, but including it requires the assumption that self-ratings and peer ratings are essentially equivalent. SRM omits the self-rating but estimates the missing cell by using the column and row means for that rating (Warner et al., 1979, p. 1747).

Choosing a Domain: The Big Five

In choosing which dimensions to measure, we noted the burgeoning evidence that the so-called "Big Five" factors circumscribe the fundamental dimensions of personality (e.g., Costa & McCrae, 1989; John, 1990; Wiggins & Trapnell, in press). Following common usage, we refer to the five factors as Extraversion, Agreeableness, Conscientiousness, Stability,[4] and Openness to Experience. Because of the robustness of the five-factor solution and their extensive construct validation, the Big Five are now used widely in personality research.

The SRM has already been applied to the Big Five in a number of studies: Thirty-two were listed in a recent review by Kenny, Albright, Malloy, and Kashy (1994). When averaged across the Big Five, target variance was higher in the studies involving long-term (28%) than short-term acquaintance (14%), presumably because long-term acquaintance gave raters more information with which to differentiate the targets. Within the Big Five, the review cites the highest amounts of target variance for Extraversion, with Conscientiousness the next highest (Kenny et al., 1994). Most important for the present study, the review concluded that target variance was the smallest of the four SRM components, estimated at roughly 15% (Kenny, 1994, p. 84) even for group studies such as the present ones (p. 59).

Overview of the Present Studies

We suspected that the figure of 15% was an underestimate of raters' ability to reach consensus on personality differences. Higher figures would require a well-controlled longitudinal study of group perceptions designed to maximize target variance. We could find only three published SRM studies that even approximated this ideal. Malloy and Albright's (1990) study is similar but is not longitudinal. Montgomery's (1984) study is longitudinal but does not include the entire Big Five. Kenny, Horner, Kashy, and Chu, (1992, Study 3) also is similar, but the first wave is zero-acquaintance. None of these studies took steps to control the amount of interaction among group members. In short, none of these studies had all the requisite features to maximize target variance. Therefore, we designed, conducted, and replicated such a study.

We collected three similar data sets of round-robin Big Five ratings of group members. As well as peer ratings, each data set included two other modes of assessment—questionnaire and self-rating. Questionnaires were completed before the participants were randomly assigned to discussion groups of 4–6 members. After Meetings 1 and 7 (hereafter called Waves 1 and 2), participants rated all their group members (including themselves) on Big-Five-related adjectives.

Several design features are worth noting. We chose group over dyadic interactions because target variance is higher in group interactions (Kenny, 1994, p. 59). Seven group meetings of 20 min each approximates the 2–3 h acquaintance of the in-depth longitudinal studies reviewed by Kenny (1994, p. 62). Those studies yielded target variance that matched that of long-term acquaintance studies only for Extraversion. To maximize target variance on the other Big Five dimensions, we installed several other features: (a) we provided a wide variety of group tasks to bring out different personality facets, (b) the tasks were designed so that all group members had to participate, (c) our raters were encouraged to give refined ratings by disallowing ties, and (d) homogeneity of the items for each rating dimension was maximized by selecting correlated items from McCrae and Costa (1987).

To ensure that the consensus levels were not just stereotypes devoid of accuracy, we took further steps to address the validity of the rating data. We administered a well-validated Big Five personality inventory and collected self-ratings at each point in time. These measures are used as criteria for examining differences in rating validity across time and Big Five dimension.

Finally, we built in a number of controls. To ensure that accumulating information was common to all group members, members of each group were initially strangers, and participants were discouraged from interacting with their group members outside of the meetings. Also, to encourage raters to base their judgments on targets' behavior rather than targets' self-descriptions, none of the exercises involved direct self-descriptions of personality.

Apart from our goal of maximizing target variance, we also sought to measure a variety of other rating effects. Use of the SOREMO computer program, based on the SRM, provided estimates of rater and relationship, as well as target variance. In addition, SOREMO provides estimates of interdependency effects (i.e., projection, reciprocity) whose existence would confirm the necessity of using SRM instead of standard analysis techniques. As far as we know, there are no published SRM group studies reporting projection estimates. We examined all of these effects separately for each of the Big Five dimensions. Finally, we conducted two replications of the original study to test the stability of such effects.

[4] To ensure that all factors were pointed in the positive direction, we reversed the scoring on Neuroticism and relabeled it Stability.

Hypotheses

We anchored our hypotheses in the reviews of person perception studies provided in Kenny et al. (1994) and Kenny (1994). We used the norms from those reviews and the specifics of our methodology to develop five basic hypotheses. For each hypothesis, we also formulated a corollary hypothesis regarding changes in the effects over time.

Hypothesis 1: Rating Components

On the basis of his reviews, Kenny (1994) proposed a set of general "rules," that is, expected values of variance accounted for by rating components. These values were 15, 20, 20, and 45% of rating variance accounted for by target, rater, relationship, and error, respectively (Kenny, 1994, p. 84).

a. Target variance. In addition to the 15% overall figure, Kenny's (1994) review of target variance studies yielded means of 32, 10, 16, 10, and 14% for Big Five Factors I–V, respectively. Because our methodology aimed at maximizing target variance, we expected to exceed those values for all factors but Extraversion. Given its high observability (John & Robins, 1993; Kenrick & Stringfield, 1980), Extraversion should yield the highest target effects at both waves. On the basis of our earlier research (Paulhus & Bruce, 1992), we did not expect to find an increase in target variance over time.

b. Rater variance. Kenny's (1994) review reported rater variances ranging from .06 to .37, with a mean of .20. We expected values at the low end of this range because we did not allow raters to give ties (see Method section), thereby preventing raters from clustering their responses at a preferred level. Also, rater variance should be smaller in Study 1, in which 10-point ratings were collected, than in Studies 2 and 3, in which 15-point ratings were collected. With narrower scales, raters have less opportunity to show level preferences.

c. Relationship variance. Kenny's (1994, p. 84) review reported a mean relationship variance of .20. We had no reason to predict otherwise. Kenny also noted that ratings related to "liking" showed twice the relationship variance of other ratings. Given that, of the Big Five, Agreeableness is most related to liking (Graziano & Eisenberg, in press), we predicted the highest relationship variance for Agreeableness.

Recall that multiple indicators are necessary to make the distinction between error variance and relationship variance. Because Study 1 did not include multiple indicators, relationship variance is indeterminate; hence this hypothesis is relevant only for Studies 2 and 3.

Hypothesis 2: Reciprocity

Any relation between A's rating of B and B's rating of A represents a dyadic reciprocity effect. As Kenny (1990, p. 23) noted, SRM calculates this reciprocity as the correlation of all such rating pairs with the target and rater effects partialed out. Kenny's (1994) review found evidence for dyadic reciprocity on affect-laden judgments (e.g., liking) but not on trait judgments.

Because our participants were initially unacquainted and were randomly assigned to groups, we expected that reciprocity at Wave 1 would certainly be minimal. Even after seven meetings, reciprocity would be unlikely to develop among our participants. After all, in group meetings, no two members can have a private interaction. Admittedly, we could not completely prevent students from interacting outside of the meetings, but we discouraged it. In short, our best hypothesis is that no reciprocity would be observed on any of the Big Five factors.

Hypothesis 3: Projection

Projection is the tendency for high scorers to give high scores to others (see footnote 1). SRM calculates it as the correlation between the target mean and the rater mean across participants. Given that we have three measures of target variance (questionnaire score, self-rating, and mean peer rating), then we have three possible indexes of projection: Each index is calculated by correlating (across persons) the target score with the mean rating given.

Earlier trait rating studies provide mixed support for the prediction of projection. D. T. Campbell et al. (1964) found no evidence for similarity projection but did find evidence for a contrast effect, presumably due to anchoring and adjustment. More recently, J. D. Campbell (1986) found similarity and contrast projection on certain ability traits, but only under limited conditions. Conclusions from recent reviews have also indicated mixed evidence for projection (Paulhus, Fridhandler, & Hayes, in press; Sherwood, 1980).

The only evidence from an SRM group study is Kenny's (1994) reanalysis of Park and Judd's (1989) data. Only Agreeableness showed a consistent positive correlation of rater and target means (Kenny, 1994, p. 109). Given this limited evidence, we hesitated to predict projection on any dimension except Agreeableness.

Hypothesis 4: Convergent Validity

Correlational studies typically index the validity of personality ratings by their correlation with other assessment modes. We have both questionnaire and self-ratings modes. In short-term acquaintance studies, rating validities with such criteria tend to be in the .20–.40 range (e.g., Funder & Colvin, 1988; Paulhus & Bruce, 1992). In fact, significant validities have been found with raters who have minimal information about the targets (e.g., Berry, 1990; Borkenau & Liebler, 1992).

In such studies, however, all variance other than the systematic relation between target measures is unaccounted for and is thus seen as error. On the peer-rating mode, for example, rater and relationship variance simply add noise to the target variance, thereby reducing correlations with other measures of target variance. As noted earlier, the SRM stabilizes the target variance by isolating it from other variance sources. Correlations are also disattenuated according to the degree of unreliability of the peer-rating measures.

For these reasons, we hypothesized that, by Wave 2, our rating validities would exceed those found in non-SRM studies: Ours should be in the .40–.60 range for all Big Five dimensions. Finally, on the basis of the only longitudinal Big Five validity study (Paulhus & Bruce, 1992), we hypothesized an increase in validity with acquaintance.

Method

Participants

Three similar data sets (N = 79, 58, and 59 participants in 16, 11, and 12 groups, respectively) were collected.[5] In each study, the participants were third-year undergraduates in a personality course at a large Canadian university. Overall, 38% were male, and 36% had East Asian ancestry.

As a class exercise, they participated in discussion groups oriented around course topics. The participants later used their ratings as the basis of a term paper concerning how their impressions of their discussion group members changed over time. After the course, participants were asked if their ratings could be used as part of a personality study. None refused.

Materials

Questionnaire. In all three studies, Costa and McCrae's (1989) Five Factor Inventory (FFI) was used as the questionnaire measure of the Big Five factors. It comprises 60 items (12 items for each of the Big Five) and requires less than 10 min to complete.

Peer ratings. These are the adjective scales on which participants rated their group members on the basis of behavior observed in the discussions. The same set of scales were completed at home after Meetings 1 and 7.

Participants were asked to write the initials of each group member over a number on the scale itself. It was explained clearly that tie ratings were not allowed: That is, they could write only one initial over any number on the scale. This requirement was designed to counteract the usual tendency for participants to rate other participants as highly positive and therefore highly similar; raters would be forced to put more effort into making distinctions across targets.

In Study 1, there was a total of five scales—one indicator for each Big Five factor. Each was a unipolar scale ranging from *not at all* (1) to *very much* (10). To help clarify the construct, all (but one) adjective labels were followed by two related adjectives. The exact labels were: *assertive* (vocal, dominant), *prosocial* (cooperative, likable), *work oriented* (deliberate, organized), *insecure,* and *intellectual* (original, clever).

In Studies 2 and 3, participants rated 15 bipolar adjective scales, that is, 3 indicators per Big Five factor (e.g., *outgoing, peppy,* and *sociable* for Extraversion). The indicators were selected from the set validated by McCrae and Costa (1987). Again, ties were not allowed.

Self-ratings. Participants were asked to include themselves when rating group members on the above scales. This requirement yielded self-rating scores on the same scales as for peer ratings at Waves 1 and 2 of the study.

Procedure

After the first class, participants were asked to complete the FFI questionnaire at home and return it at the next class meeting. The discussion group assignments were then completed randomly. As each group assembled, the instructor verified that all were unacquainted; if two students were acquainted, one was replaced with an unacquainted student from the remaining pool of students. This assignment yielded groups of 4–6 previously unacquainted students. These groups met once a week during class for 7 weeks. In each meeting, they spent 20 min discussing a course topic assigned to them.

After Meetings 1 and 7, participants were provided with a rating form to complete at home and return to the instructor at the next class meeting. They were told to seal the completed form in the envelope provided to ensure confidentiality. They did not know in advance what traits they would be rating. As noted above, the rating form involved rating themselves as well as other group members on a list of traits.

The ratings were returned confidentially to the raters near the end of the course to be used as the basis for their term paper. They were told not to share their ratings with other group members.

Analyses

We performed the bulk of the analyses using Kenny's (1990) SOREMO program. Each data set submitted to SOREMO consisted of several groups of round-robin ratings on one construct. SOREMO begins by partitioning variance in a 2 × 3 analysis, breaking it into stable and unstable components across target, rater, and relationship. Therefore, for each study, this analysis had to be performed a total of 10 times—once for each of the Big Five factors at each of the two waves.

Unfortunately, SOREMO computes significance values only for individual indicators (e.g., *outgoing*), not for the constructs (e.g., Extraversion). Therefore, to test the constructs, the summed indicators for each construct had to be manually tested for significance.[6] Most of the tests below are based on group-level analyses. In Study 1, for example, the df = 15 for most tests because there are 16 groups of raters (see SOREMO manual, Kenny, 1990, p. 12). Minimal sex differences were found; hence the data for men and women were pooled.

Results

Hypothesis 1: Rating Components

Table 2 contains three panels corresponding to our three studies. Each panel shows the percentage of variance[7] attributable to various sources: that is, target, rater, relationship, and error. The panels are further broken down by Big Five factor and wave (separated by slashes).

Each entry was tested for significance by a one-tailed t test comparing the amounts of variance to 0.[8] Although this table is broken down by Big Five and wave, the hypotheses below focus on the mean results, which may be found in the rightmost column.

Hypothesis 1a: Target variance. In all three of our studies, target variance was substantial. At Wave 2, Studies 1, 2, and 3 yielded 29, 28, and 28%, respectively. Given that these figures are almost twice as high as the Kenny "rule" of 15% target vari-

[5] Part of the data from Study 1 were previously analyzed with purely correlational methods and reported in Paulhus and Bruce (1992). The present analyses with the SRM provide a much richer and more comprehensive picture of the data.

[6] One would think that these tests could be performed in SOREMO by first summing across the indicators and then determining significance values. Unfortunately, relationship variance can only be separated from error variance with multiple indicators, which would be lost if one were to sum across the indicators.

[7] Variance attributable to each source is calculated by determining the *absolute* variance within each group and pooling across all groups. SOREMO then converts these values into relative variances.

[8] Notice the apparent inconsistency in some significance values: For instance, on Conscientiousness in Study 1, the .24 for rater at Wave 2 is not significant, whereas the .07 for target at Wave 1 is. The reason for the inconsistency is that significance is determined across groups; the group is the unit of measure. Hence, a large value may not be significant if there is a great deal of variation across groups. A test of variance against 0 seems counterintuitive; however, it is appropriate within SOREMO because the values being tested are not the actual variances but a transformation that has an expected value of 0 under the null hypothesis.

Table 2
Percentage of Variance Accounted for by Rating Source Across Factor and Wave

Source	E	A	C	S	O	*Mean*
			Study 1			
Target	68*/62*	35*/26*	35*/24	7/0	44*/34*	38/29
Rater	7*/6*	9*/4	7*/0	17*/0	10*/1	10/2
Rel'p/error	25/32	56/70	58/76	76/100	46/65	52/68
			Study 2			
Target	47*/49*	11*/24*	10/20*	25*/23*	23*/23*	23/28
Rater	1/4	17*/10*	18*/18	15*/4	12/13*	13/10
Rel'p	23*/23*	19*/28*	22*/28*	26*/20*	21*/21*	22/24
Error	29/23	54/39	50/34	35/54	44/43	42/39
			Study 3			
Target	29*/46*	19*/30*	13*/24*	8*/20*	11*/18*	17/28
Rater	5/8	19*/15*	23*/20*	12*/13*	24*/17*	17/15
Rel'p	28*/17*	33*/36*	21*/15*	24*/23*	16*/16*	26/21
Error	38/29	29/19	44/41	46/44	54/49	42/36

Note. Values for Waves 1 and 2 are separated by a slash. Relationship effects could not be estimated in Study 1 because only one indicator was used to measure each factor. E = Extraversion; A = Agreeableness; C = Conscientiousness; S = Stability; O = Openness; Rel'p = relationship.
* $p < .05$.

ance, our hypothesis was strongly supported. In fact, even after the first meeting, our target variance averaged 26%. Across the 39 groups the target variance did not differ significantly from Wave 1 to Wave 2, $t(38) = 1.44$, *ns*. Although Extraversion was highest, as predicted (*Mdn* = 49%), it is noteworthy that we found median values above 15% for all of the Big Five factors.

Hypothesis 1b: Rater variance. We hypothesized that rater variance would be smaller than the Kenny norm of 20%. At Wave 2, we found means of 2, 10, and 15% in Studies 1, 2, and 3, respectively, at Wave 2. Hence, this hypothesis was strongly supported. Note that, at Wave 1, rater effects were slightly higher at 10, 13, and 17%.

Hypothesis 1c: Relationship variance. Relationship variance could be separated from error in Studies 2 and 3. Here, relationship variance was consistently significant: Indeed, all 12 effects in Table 2 were significant.

Averaged across traits, the values were slightly stronger than the Kenny norm of 20%. Specifically, in Studies 2 and 3, 23% and 24% of the variance was accounted for by relationship effects. As predicted, Agreeableness showed the highest value, with a mean of 32% at Wave 2, compared with a mean of 20% for the other four traits.

Hypothesis 2: Reciprocity

Reciprocity is the tendency for the rating A gives to B to correlate with the rating B gives to A. This tendency was measured in the present studies by correlating stable relationship effects across all pairs of individuals who rated each other.

We computed significance with a two-tailed *t* test of the covariance of each construct across all groups. The covariances are used because, unlike correlations, they can be combined across groups. Only 2 of these 30 values (across the Big Five factors at the two waves in the three studies) were significant; by chance alone, we would expect 1.5 of 30 to achieve significance. Thus, the hypothesis of no reciprocity was supported.

Hypothesis 3: Projection

Projection is the tendency for high scorers to give high scores to others. Given that we had three kinds of target measures available, there are three ways of measuring projection. For each factor, study, and wave, we correlated the mean rating given by each person with (a) the mean rating received, (b) the questionnaire score, and (c) the self-rating. Two sets of correlations are provided in Table 3: The third set, correlations of ratings

Table 3
Projection Effects Separated by Big Five Factor and Wave

Study	E	A	C	S	O
		Peer ratings given and peer ratings received			
1	3/−3	47/−35	51/−38	−100/13	4/0
2	44/8	5/21*	4/17	14/47	42/−8
3	13/38	44/32	−28/−13	53/14	−11/−53
		Peer ratings given and questionnaire score			
1	26/−30	82*/0	35/23	−91*/06	23/0
2	0/27	48*/74*	45*/−1	19/27	29/29
3	34/26	48*/44*	40/14	28/02	6/−2

Note. Values for Waves 1 and 2 are separated by a slash. Each entry is a disattenuated correlation. Decimal points have been removed to save space. E = Extraversion; A = Agreeableness; C = Conscientiousness; S = Stability; O = Openness.
* $p < .05$.

given to others with ratings given to self, is inherently problematic. Rating leniency tendencies will tend to create spurious correlations. Hence this set is not reported here.

In the top panel of Table 3, the Peer-given with Peer-received correlations seem erratic, and only one was significant; hence, we concluded that no evidence was found for projection of this type. By contrast, the Peer-given with questionnaire panel shows consistent evidence for projection on Agreeableness (5 of 6 significant) but not on the other Big Five factors (only 2 of 24). Stronger results with the questionnaire may result from its higher reliability.

Hypothesis 4: Convergent Validity

Validity of the peer ratings is indicated by the correlations with the questionnaire score and self-ratings. Invariably we found the self-rating results were similar though weaker than those with the questionnaire. Hence, to save space we present only the questionnaire results in Table 4. Each value is a disattenuated correlation.

We hypothesized that validities would be in the .40–.60 range at Wave 2. In fact, pooled across the three studies, we found significant validities at both waves for all Big Five factors. Overall, we found mean correlations of .36 at Wave 1 and .43 at Wave 2. To examine the effects of acquaintance, we pooled the 39 groups and compared mean validities (excluding Extraversion) at the two waves.[9] As predicted, the improvement in validity was significant, $t(38) = 2.82$, $p < .01$.

Discussion

We began with the norms for various rating effects summarized in Kenny's (1994) review of person perception research. For example, he cited typical figures of 15, 20, and 20% variance for target, rater, and relationship, respectively. In response, our three studies were designed to increase target variance and reduce rater variance. To assess the consequences for validity, we included questionnaire and self-ratings to examine convergence with peer ratings. We also sought replicated estimates of reciprocity and projection. Given that our success in these goals varied across hypotheses, the discussion below is organized by hypothesis.

Hypothesis 1: Rating Components

One clear finding was that our studies yielded higher estimates of target variance than did Kenny's (1994) review of short-term acquaintance studies. At both waves, our target variance matched that achieved by long-term acquaintances (e.g., Malloy & Albright, 1990). This success may derive in part from our prohibition of ties. Consider, for example, a group in which some raters are extremely lenient—they give all group members perfect ratings on Agreeableness—and other raters are hostile—they give all targets identical low ratings. For this group, consensus is difficult because no distinctions are made among targets. Rater variance, however, would be large, because raters' preferred rating level is the sole determinant of the ratings given. By contrast, our prohibition of ties required raters to spread out their ratings rather than cluster them at some preferred level. Perhaps more important, the prohibition of ties required our raters to think in more depth about the personality of their group colleagues.

Another explanation for our higher levels of target variance lies in our improved choice of indicators for each of the Big Five factors.[10] Note first that stable target variance is a direct function of the intercorrelations of the indicators (Kenny, 1990). If the indicators do not correlate highly, target variance may emerge, but it will be largely unstable; stable target variance occurs only when the ratings of the indicators overlap. We chose indicators known to be highly intercorrelated, whereas in earlier studies the multiple indicators may share less variance. For example, in Kenny et al. (1992, Study 2), the two indicators of Factor V were Intelligence and Imagination, two descriptors that do not correlate highly (Trapnell, 1994). Of course, the choice of construct indicators is always a trade-off of bandwidth and fidelity: Choosing indicators that are virtually synonymous (i.e., *peppy* and *full of energy*) may reduce validity because the indicators do not span the possible meanings of the construct (Ozer, 1989).

It also appears that the number of points on the rating scale has an impact on rating components. Use of the 10-point scale in Study 1, rather than the 15-point scales used in Studies 2 and 3, yielded the highest target variance and lowest rating variance of the three studies.

Across the Big Five, Extraversion exhibited the most target variance by far and Stability, the least. This finding accords with many previous studies showing that Extraversion is the most easily and validly rated construct.

By contrast, rater effects were less evident for Extraversion (5%) than for the other constructs (an average of 12%). (These means were found by averaging the results within each Big Five factor, across both waves and the three studies.) Apparently rater style plays less of a role when raters are able to make clear distinctions across targets. Factor differences in personalism were not dramatic: Agreeableness was the highest at 29%, but following closely were Stability (26%), Extraversion (23%), Conscientiousness (21%), and Openness (18%).

Hypothesis 2: Reciprocity

On the basis of previous studies of group interactions, we predicted no reciprocity. This hypothesis was supported. Reciprocity seems to be minimal in group interactions because the same information is available to all raters.

Hypothesis 3: Projection

Except for Agreeableness, our data show little evidence of projection, that is, a relation between an individual's standing on a trait and the ratings that the individual gives to others. Re-

[9] Note that these significance tests are rather conservative because they have to be performed at the group level. Recall that we have 16, 11, and 12 groups in Studies 1, 2, and 3, respectively.

[10] Alternatively, the increase in relative target variance could have resulted indirectly from our efforts to reduce rater variance. If so, relationship variance should also have been double that found in previous studies. This was not the case.

Table 4
Validities Separated by Big Five Factor and Wave: Questionnaire Score Versus Peer Rating

Study	E	A	C	S	O	*Mean*
1	38/34	24/54	23/38	12/11	31/58	26/39
2	52/55	29/26	31/50	22/33	51/39	37/41
3	50/37	58/37	37/34	56/74	21/70	44/50
Mean	47*/43*	37*/39*	30*/41*	30*/39*	34*/55*	36*/43*

Note. All values are disattenuated correlations. Decimals have been removed to save space. E = Extraversion; A = Agreeableness; C = Conscientiousness; S = Stability; O = Openness.
* $p < .05$.

call that we had two such indexes. The evidence was weaker for the given/received index of projection—that is, when ratings given were correlated with ratings received by the same rater. Evidence for projection was strongest when ratings given were correlated with the questionnaire measure of rater personality. A possible explanation is that the questionnaire (FFI) is a reliable, well-validated instrument, whereas the peer ratings are psychometrically weaker. Possibly, the FFI was simply more valid, that is, better able to pick up true Agreeableness differences than were the peer ratings.

The emergence of projection only on Agreeableness is intriguing. Although Kenny (1994) reported a similar finding for dyadic studies, it was not obvious that the same result would obtain for group perceptions. It is well known that individuals differ in the so-called "Pollyanna effect" (Matlin & Stang, 1978); that is, likable individuals tend to consider others to be likable as well, on the basis of their prior (likely positive) experiences. At the other end of the spectrum, cynical, hostile individuals may see others as generally disagreeable, on the basis of their prior (likely negative) experiences. This effect may conversely be called a "sourpuss effect." This line of reasoning does not predict projection for the other constructs (e.g., a conscientious person will not necessarily see others as conscientious).

Note that any confirmation of projection in our data supports the utility of the SRM analysis approach. That is, projection contravenes the assumption of the independence of ratings required for analysis by standard ANOVA. In short, standard ANOVA is incapable of analyzing round robin rating data—at least for Agreeableness.

Hypothesis 4: Convergent Validity

In general, the present studies provide strong evidence for the validity of the Big Five factors by demonstrating a convergence of three distinct modes of measurement. In general, the questionnaire measure outperformed the self-ratings in predicting peer ratings, but all correlations among the modes were substantial. Differences in validities across waves and factors generally substantiated previous findings.

Note that these correlations are disattenuated for rating unreliability. This adjustment allows us to more fairly compare validities across studies. Otherwise differences in reliabilities across studies would be confounded with validity differences.

Differences across wave. Some overall improvement in validity was observed with increased acquaintance: The mean correlation at Wave 1 was .36, and the average correlation at Wave 2 was .43. As in previous studies, Extraversion was the exception, showing little change. Accordingly, when Extraversion was excluded, improvement with acquaintance was more apparent. That is, the mean correlation was significantly higher for Wave 2 ($M = .44$) than Wave 1 ($M = .33$). Thus, Hypothesis 4 was supported. This consistent increase in validity across three studies corroborates the results reported by Paulhus and Bruce (1992).

Differences across the Big Five. Slight differences in validity across the Big Five were observed, with Extraversion at .45 and Openness at .44, being the highest overall, and the others ranging between .35 and .40. Many previous studies have demonstrated that Extraversion has the highest validity of the Big Five factors (e.g., Kenrick & Stringfield, 1980; Paulhus & Bruce, 1992; Watson, 1989). The finding is partly explained by the fact that Extraversion consistently shows the highest consensus (Kenny et al., 1994). The high validity for Openness is less common in group settings but has been obtained when a variety of intellect- and imagination-related exercises are assigned to the discussion groups (John & Robins, 1993).

Self-report criteria. Some might criticize our use of self-report measures as criteria for the validity of peer ratings (e.g., Kenny, 1994, Chapter 7). The usual criticism that convergence results from participants communicating their personalities to others by self-descriptions has been undermined by studies showing that peers are influenced far more by the target's behavior than by the target's self-descriptions (Amabile & Kabat, 1982). More important in our studies, if self-descriptions during the group meetings were bringing about the convergence with peer ratings, then the latter should converge more with the self-ratings made in the group context than with the questionnaire measure administered before any meetings took place. In fact, our studies established firmly that the questionnaire measure outperformed the self-ratings in terms of predicting peer-ratings.

Admittedly, the present studies would have been improved by including criteria such as behavioral information (e.g., Borkenau & Liebler, 1992; Levesque & Kenny, 1993; Paulhus & Morgan, 1995) or ratings from close acquaintances (Funder & Colvin, 1988). As far as we know, however, there is no clear evidence that either of these criteria is superior to an established questionnaire measure such as the NEO–FFI as the ultimate criterion for personality constructs (see Ozer, 1989).

Conclusions

We had two goals in mind in exploiting the SRM for analyzing personality ratings in the group context. One was to redress

the impression from Kenny's (1994) review that the person is relatively unimportant in person perception. We showed that, even for short-term acquaintance, target variance could be boosted well beyond the 15% "rule" and that rater variance could be reduced well below the 20% rule: Thus, our first goal was clearly attained. In short, our person-perception methodology put more emphasis on the person and less on the perception. This study bridges the gap between the short-term and long-term acquaintance studies reviewed by Kenny et al. (1994). That is, we achieved target variance as high as that of long-term acquaintance studies within a short-term acquaintance paradigm. Higher target variance, in turn, allowed us to achieve substantial validities for all of the Big Five factors.

The inquiring reader might retort that Kenny's 15% rule is still a more typical figure for group rating studies. Rather than debating what a typical study is, we would simply warn that researchers should not expect high target variance and validity on personality factors that participants are not given an opportunity to manifest. Inclusion of appropriate exercise and requiring participation will rectify this. Undebatable is the conclusion that tighter methodology (disallowing ties, etc.) will yield higher target variance. As one reviewer pointed out: Although participant judges do not usually like to make fine distinctions, this doesn't mean they are unable to do so. In sum, future researchers should always be able to exceed the 15% rule on target variance.

Our second goal was to evaluate a number of rating effects that had not been emphasized in earlier SRM studies. This goal also was accomplished. Reciprocity was confirmed to be minimal in the group context. On the other hand, projection was found, but only on the Agreeableness factor. In estimating all the above effects, we tried to be as comprehensive and thorough as possible: We employed three modes of measurement across all the Big Five factors, and we replicated our procedure three times.

Perhaps most striking among our results is clear evidence for a consensus–accuracy paradox—that is, acquaintance increases accuracy without a corresponding increase in consensus. Pooled ratings do not converge, yet they become more accurate. This phenomenon was predicted earlier by Kenny's (1991) *Weighted-Average Model* of person perception. Kenny explained that high overlap of information available to raters, as in our discussion groups, yields a high initial target consensus that is unlikely to increase with acquaintance. After the first meeting, raters may all agree on stereotypic inferences to be drawn from that impoverished information. For example, initial stereotypes may be based on sex or race. With further acquaintance, the consensus shared by raters becomes gradually more aligned with the target's character. Stereotypes can be as consensual as informed impressions, but they are not as accurate.

References

Amabile, T. M., & Kabat, L. G. (1982). When self-description contradicts behavior: Actions do speak louder than words. *Social Cognition, 1,* 311–325.

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.

Burisch, M. (1984). Approaches to personality inventory construction: A comparison of merits. *American Psychologist, 39,* 214–227.

Byrne, D. (1971). *The attraction paradigm.* New York: Academic Press.

Campbell, D. T., Miller, N., Lubetsky, J., & O'Connell, E. J. (1964). Varieties of projection in factor attribution. *Psychological Monographs, 78,* 1–33.

Campbell, J. D. (1986). Similarity and uniqueness: The effects of attribute type, relevance, and individual differences in self-esteem and depression. *Journal of Personality and Social Psychology, 52,* 281–294.

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

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., & West, S. G. (1993). Consensus, self–other agreement, and accuracy in personality judgment: An introduction. *Journal of Personality, 61,* 457–476.

Graziano, W. G., & Eisenberg, N. H. (in press). Agreeableness: A dimension of personality. In R. Hogan, J. Johnson, & S. R. Briggs (Eds.), *Handbook of personality psychology.* San Diego, CA: Academic Press.

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.

John, O. P., & Robins, R. W. (1993). Determinants of interjudge agreement on personality traits: The Big Five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality, 61,* 521–551.

Kenny, D. A. (1990). *SOREMO: Version V.* Unpublished manual, University of Connecticut.

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

Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis.* New York: Guilford Press.

Kenny, D. A., Albright, L., Malloy, T. E., & Kashy, D. A. (1994). Consensus in interpersonal perception: Acquaintance and the Big Five. *Psychological Bulletin, 116,* 245–258.

Kenny, D. A., Horner, C., Kashy, D. A., & Chu, L. (1992). Consensus at zero acquaintance: Replication, behavioral cues, and stability. *Journal of Personality and Social Psychology, 62,* 88–97.

Kenny, D. A., & La Voie, L. (1984). The social relations model. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 141–182). New York: Academic Press.

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.

Kiesler, D. J. (1983). The 1982 interpersonal circle: A taxonomy for complementarity in human transactions. *Psychological Review, 90,* 185–214.

Kirk, R. E. (1968). *Experimental design: Procedures for the behavioral sciences.* Belmont, CA: Wadsworth.

Levesque, M. J., & Kenny, D. A. (1993). Accuracy of behavioral predictions at zero acquaintance: A social relations analysis. *Journal of Personality and Social Psychology, 65,* 1178–1189.

Malloy, T. E., & Albright, L. (1990). Interpersonal perception in a social context. *Journal of Personality and Social Psychology, 58,* 419–428.

Matlin, M. W., & Stang, D. J. (1978). *The Polyanna Principle: Selectivity in language, memory, and thought.* Cambridge, MA: Schenkman.

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor

model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52,* 81–90.

Montgomery, B. (1984). Individual differences and relational interdependences in social interaction. *Human Communication Research, 11,* 33–60.

Ozer, D. J. (1989). Construct validity in personality assessment. In D. M. Buss & N. Cantor (Eds.), *Personality psychology: Recent trends and emerging directions* (pp. 224–234). New York: Springer-Verlag.

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., & Bruce, M. N. (1992). The effect of acquaintanceship on the validity of personality impressions: A longitudinal study. *Journal of Personality and Social Psychology, 56,* 493–505.

Paulhus, D. L., Fridhandler, B., & Hayes, S. (in press). Psychological defense: Contemporary theory and research. In R. Hogan, J. Johnson, & S. R. Briggs (Eds.), *Handbook of personality psychology.* San Diego: Academic Press.

Paulhus, D. L., & Morgan, K. L. (1995). *Determinants of perceived intelligence in leaderless groups I: Shyness and acquaintance.* Manuscript submitted for publication.

Piedmont, R. L. (1994). Validation of the NEO PI–R observer form for college students: Toward a paradigm for studying personality development. *Assessment, 1,* 259–268.

Sherwood, G. G. (1980). Self-serving biases in person perception: A reexamination of projection as a mechanism of defense. *Psychological Bulletin, 90,* 445–459.

Trapnell, P. D. (1994). Openness versus intellect: A lexical left-turn. *European Journal of Personality, 8,* 273–290.

Trope, Y., & Higgins, E. T. (1993). Inferring personal dispositions from behavior. *Personality and Social Psychology Bulletin, 18,* (Special issue, Whole No. 5).

Warner, R. M., Kenny, D. A., & Stoto, M. (1979). A new round robin analysis of variance for social interaction data. *Journal of Personality and Social Psychology, 37,* 1742–1757.

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.

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

Received September 7, 1994
Revision received April 12, 1995
Accepted April 13, 1995 ■