Consensus in Personality Judgments: Moderating Effects of Target–Rater Acquaintanceship and Behavior Observability

Sampo V. Paunonen
University of Western Ontario, London, Ontario, Canada

Conducted a peer rating study (N = 111) to determine the effects of (a) level of acquaintanceship between rater and target and (b) degree of public observability of rated personality traits on peers' perceptions of target personality characteristics. As hypothesized, I found the agreement between peer ratings and target self-ratings to vary linearly and directly with acquaintanceship. In addition, acquaintanceship interacted with observability such that the public visibility of the behavior domain being judged was an important determinant of agreement for low to moderately acquainted peer dyads but not for highly acquainted dyads. Contrary to expectations, however, trait observability did not show a main effect with regard to self–peer agreement. The basis of the study is described with reference to the lens model of inferential behavior, and implications of the results are discussed with reference to past and future attempts at evaluating consensus and accuracy in person perception.

Person perception is a topic that persists as an area of research interest in psychology. This interest is maintained, in all likelihood, by an appreciation for the fact that observers' subjective evaluations of others play a large part in everyday social encounters. Observer judgments are also considered worthy of empirical study because they are often integral components of more formal interpersonal encounters, such as employment interviews and clinical psychodiagnostic assessments. It is, therefore, not surprising that much of the past research on person perception has dealt explicitly with the issue of accuracy. As a lay person, it is gratifying to think that one can be knowledgeable about the personal dispositions of close acquaintances, allowing one to predict their behavior in circumscribed situations with some degree of confidence. As a professional practitioner, one would like to believe that he or she can predict a job candidate's future job performance or a patient's responsiveness to a program of treatment for pathological behaviors.

A perusal of the recent empirical literature in the area of person perception indicates a significant change in research focus. Studies of accuracy have largely given way to studies of process; contemporary research in social judgment has dealt predominantly with judgment biases and the psychological mechanisms that facilitate errors of judgment. For instance, one of the more popular and impassioned areas of inquiry has been related to the issue of semantic determinants of distortion in personality inferences (e.g., see Block, Weiss, & Thorne, 1979; D'Andrade, 1974; Paunonen & Jackson, 1979; Shweder, 1975). This transition in research focus from accuracy to process has been interpreted as being due primarily to methodological criticisms of early accuracy studies, criticisms that began to appear in the mid-1950s (Funder, 1987). To this reason can be added the particularly discouraging fact that early studies of accuracy yielded mostly only modest or trivial support for the hypothesis that observers are good judges of others (Jackson & Paunonen, 1980; Mischel, 1968; Swann, 1984; Wiggins, 1973). The nature of this new wave of research, its evolution from early accuracy studies, and alternative interpretations of the resultant data have recently been analyzed in some detail in an insightful critical essay by Funder (1987).

The present study was designed in the belief that the popular intuition about observer accuracy, which seems to be contrary to many published findings, is not entirely incorrect; that numerous factors can operate to attenuate accuracy coefficients in experimental research; and that a study meeting certain preconditions can provide convincing empirical support for the hypothesis that people are capable judges of the personal characteristics of others. It is noted at this point that this article's principal concern is with observer perceptions of behavior tendencies in particular and personality dispositions in general. Person perception itself is a global term that includes not only personality inferences, but also inferences about emotions, attitudes, abilities, group membership, behavior intentions, and causes of behavior (Schneider, Hastorf, & Ellsworth, 1979).

Lessons Learned From Past Mistakes

Many of the early studies that have evaluated the accuracy of person perception have been criticized from both logical and psychometric perspectives. These criticisms have proven valuable because they have provided advice on appropriate empirical methods for maximizing obtained levels of judgment accuracy. Although it is beyond the scope of this article to describe the shortcomings of previous research exhaustively (see instead Funder, 1987; Jackson, 1982; Jackson & Paunonen, 1980; Swann, 1984; Wiggins, 1973), I review a few that are directly pertinent to the present study.

A highly profitable analysis of the person perception process in terms of maximizing accuracy begins with a consideration of Brunswick's (1955) lens model of inferential behavior. Although Brunswick developed the lens model in the context of
Figure 1. The lens model of inferential accuracy. (From "Analyzing the Components of Clinical Inference" by Kenneth R. Hammond, Carolyn J. Hursch, and Frederick J. Todd, Psychological Review, 1964, 71, p. 439. Copyright 1964 by the American Psychological Association. Adapted by permission.)
judge’s prediction might be based on a linear combination of the cues, whereas another judge’s prediction might follow some nonlinear combination. Their relative accuracies in this case depend on how the cues are empirically related, linearly or nonlinearly, to the criterion.

The use of cues by judges has recently been an intensely popular focus in the experimental study of biases in social cognition. This research has been predominantly concerned with factors affecting the perception, interpretation, and implicit weighting of cues. For example, it has been reported that people poorly encode target cues that are inconsistent with prior target information (e.g., Berman, Read, & Kenny, 1983; cf. Cantor & Mischel, 1979), that observers generally tend to overemphasize negative or undesirable target cues (e.g., Fiske, 1980; Kanouse & Hanson, 1972), that base-rate information implied by target cues is often overlooked by judges (e.g., Kahneman & Tversky, 1973), that target gender-related cues can markedly affect the interpretation of target behavior (e.g., Deaux & Emshwiller, 1974), and that target attractiveness and dress cues contribute to the accuracy of strangers’ ratings of certain personality traits (Albright, Kenny, & Malloy, 1988). The manner in which observers combine target cues configurally in formulating an overall impression or prediction has also been studied extensively with algebraic models, particularly in the area of clinical inference (e.g., Anderson, 1972; Goldberg, 1968, 1970).

The Lens Model and Peer Rating Accuracy

The foregoing discussion referred to the implications of the lens model for experimental research in particular; however, the same implications apply to naturalistic studies of person perception, such as those involving peer ratings. In casual encounters with strangers or friends, the target cues accessible to the observers are typically not intentionally manipulated or controlled by the experimenter. Instead, situational and social constraints operate to determine which cues are available. Nonetheless, for an observer’s perceptions to be accurate with regard to a criterion, the target information must have some empirical, not just theoretical, bearing on that criterion. In short, the cues must have demonstrable validity. Exchanging greetings and a handshake with a stranger might provide one with valid cues for predicting that person’s level of extraversion, but those same cues might be inadequate for predicting level of conscientiousness (see Albright et al., 1988).

The lens model structures on the number of target cues, cue redundancy, and cue utilization by observers are also relevant to peer ratings. One would expect that the accuracy of peer ratings should vary as a function of degree of target—rater acquaintance because the number of criterion-relevant cues available to the judge should vary as a function of the number and duration of interpersonal encounters. This expectation may not always be met, however, because situational or social constraints may not allow for a wide latitude in behavior, and hence target cues may be highly redundant. To illustrate, an acquaintance who is known only as a classmate may provide superfluous cues about his or her level of achievement motivation but few cues about level of autonomy. Other reasons why length of acquaintance or degree of familiarity itself may not be sufficient to guarantee accurate perceptions have been described by Cloyd (1977) and Taft (1966).

Which of the available target cues will a peer observer use in arriving at a judgment concerning a target? This is one area that has been, as mentioned previously, a popular focus of laboratory studies of bias in person perception. It has been less popular in naturalistic peer rating studies. A notable recent exception has been the work on behavior observability. Kenrick and Stringfield (1980) reasoned that a target’s more observable behavior cues should be more salient to judges and hence relied on extensively as a basis for judgment. They asked student targets to estimate “how publicly observable” their behavior on a trait dimension was and then divided the sample into high and low observability groups for the trait. Kenrick and Stringfield found, as they had expected, that peer judgments of the high observability targets were more predictive of target self-ratings than were judgments of the low observability group, across a variety of traits. Cheek (1982) and others have reported similar results (cf. Kammann, Smith, Martin, & McQueen, 1984). The observability variable merits further analysis at this point because it has been measured in different ways by different investigators. Some of these ways are more problematic than others.

Problems With Assessing Trait Observability

Although the notion that peer rating accuracy is moderated by behavior observability seems reasonable, the method often used to evaluate the hypothesis can be criticized. It has been argued logically (Paunonen & Jackson, 1985a), and shown empirically (Paunonen, 1988), that ratings of trait-behavior observability are substantially related to ratings of trait level. That is, people who are more extreme on a bipolar trait, high or low, are more likely to behave in a trait-consistent manner across occasions and situations than those who are moderate on the trait; those extreme people are the ones, therefore, for whom the trait behaviors are more likely to be observed by others. This analysis points to a problem with comparing trait rating accuracy for high versus low observability subjects—the high observability group may be characterized by an inflated range of trait scores (high and low on the trait), whereas the low observability group may be characterized by an attenuated range (moderate on the trait). Self-peer correlations for the trait will tend to be larger for the former group and smaller for the latter owing to restriction of range effects. One must, therefore, cautiously interpret the moderating effects of observability in some past studies, as those effects may be spurious. An exception to this caution is a recent study by Zuckerman et al. (1988), who reported some evidence for an observability moderator using moderated multiple regression. This statistical procedure is immune to the spurious outcomes due to the improper subgrouping of data described previously in this paragraph (see Paunonen & Jackson, 1985a). The caution also does not apply to analyses of observability reported by Funder (1980; Funder & Colvin, 1988; Funder & Dobroth, 1987), for reasons discussed in the following paragraph.

Whereas Kenrick and Stringfield (1980), Cheek (1982), and Kammann et al. (1984) measured individual differences in observability for a trait and then compared the accuracy of peer ratings for high and low observable subjects for that trait, Funder (1980) measured trait differences in observability and compared peer ratings on the more observable traits with ratings on the less observable traits. Funder found no relation,
however, between degree of self–peer agreement and level of trait observability \( (r = .15, p = .482) \). Kenrick and Stringfield (1980) also reported such an analysis in their article. However, in contrast to Funder's study, they did find a significant relation between self–peer agreement and observability across traits \( (r = .52, p = .98) \).

The problem of spurious moderator effects described earlier applies when measuring individual differences in observability but does not apply when measuring trait differences, because self–peer correlations are computed, with the latter method, across all subjects for each trait and not across subgroups of subjects. The procedure of estimating trait differences in observability can be criticized, however, if one fails to account for the bipolarity of trait domains, and only one pole of a trait continuum is scaled for observability. This criticism applies to the studies by Funder (1980), Kenrick and Stringfield (1980), and Zuckerman et al. (1988). A good illustration of the bipolarity problem with regard to the measurement of observability is with the trait dimension reserved versus outgoing. Quite obviously, outgoing behaviors are more publicly observable than are reserved behaviors. Which pole of the behavior domain is referred to by a particular predictor and criterion measure is arbitrary for the purposes of assessing peer rating accuracy on the trait, but it is not arbitrary for the purposes of assessing trait observability.

In two recent studies, Funder and Dobroth (1987) and Funder and Colvin (1988) correctly considered both poles of trait dimensions in assessing the "imaginability of [trait] confirming behaviors" (Funder & Dobroth, 1987, p. 412) and trait–disconfirming behaviors. Ratings of these two attributes for 100 trait behavior items correlated .59 in the Funder and Dobroth study (Table 2), indicating that the two poles of a trait may engender different levels of behavioral expression and visibility. (Rothbart & Park, 1986, reported a correlation of .58 for similar ratings of 150 trait adjectives.) Moreover, in both the Funder and Dobroth and Funder and Colvin studies a composite measure of trait observability, based on an average of the two imaginability measures and a few other indicators (e.g., ratings of the ease of judging a trait's level in a target), correlated positively with self–peer agreement across 100 trait items \( (r = .39, \text{Funder} \ & \text{Dobroth, 1987, Table 4}; \ r = .41, \text{Funder} \ & \text{Colvin, 1988, Table 2}) \).

Purpose of the Present Study

I designed the present study of person perception to evaluate two aspects of the lens model in a peer rating context. First, agreement between self-ratings and peer ratings of personality was evaluated as a function of the number of target cues available to peer judges by manipulating the degree of acquaintance–ship of target–rater dyads. Second, agreement as a function of cue utilization by peer raters was evaluated by examining agreement coefficients for trait domains varying in the observability of typical behavior referents. I hypothesized that (a) peer rating agreement would increase with increasing levels of target–rater acquaintance–ship, (b) agreement would also increase with increasing levels of trait behavior observability, and (c) acquaintance–ship and observability would interact in their effects on agreement, such that observability would become less of a determinant of agreement as acquaintance–ship increased. This last prediction was based on the belief that unobservable trait behaviors are generally not accurately rated by stranger judges, but, as acquaintance–ship increases, knowledge of a target's behavior predispositions broadens to include less salient trait domains.

Although the main effects of the acquaintance–ship and observability variables have been evaluated in past research, their interaction has not. Furthermore, the present study represents some improvements in methodology over past efforts. For example, both Norman and Goldberg (1966) and Paunonen et al. (1988) have reported increases in the accuracy of peer raters' personality inferences as a function of degree of target–rater acquaintance–ship. Norman and Goldberg, however, tested only two groups of subjects differing in acquaintanceship: Peace Corps volunteers who had trained together for 3 months and strangers meeting for the first time in an undergraduate psychology lab (see also Cloyd, 1977; Funder & Colvin, 1988; Taft, 1966). Although Paunonen et al. (1988, Study 3) evaluated student rater groups representing nine levels of acquaintance–ship, subjects at higher levels were extremely limited in number compared with those at lower levels, precluding certain statistical tests. Self-ratings and peer ratings in both studies were also limited because Norman and Goldberg used four-item adjective scales to measure personality traits, and Paunonen et al. used two-item adjective scales. Such measures have been shown generally to have inferior psychometric properties compared with longer, behavior-based measures (Paunonen & Jackson, 1985b, 1985c; cf. Burisch, 1985).

The present peer rating study addresses the previously mentioned concerns related to subject sampling, psychometric properties of rating scales, and the measurement of behavior observability. Student participants were assigned to target–rater dyads so that approximately equal numbers of subjects were included at each of nine levels of acquaintance–ship. Criterion and predictor measures of personality were composed of items from the Personality Research Form (PRF), an inventory with proven convergent and discriminant properties and freedom from desirability bias (Jackson, 1984). Finally, following Funder (1980), observability was scaled for each trait rather than for each subject on a trait. As reported by Funder and Dobroth (1987) and Funder and Colvin (1988), however, both poles of each of the bipolar trait dimensions under investigation were scaled for observability of referent behaviors and used to estimate the observability of the domain of behaviors as a whole.

Method

Subjects and Procedure

Undergraduate male and female students were solicited from a subject pool in return for course credit in introductory psychology. The study was described as investigating the accuracy of personality inferences, and volunteers were asked to sign up in pairs with a close friend. A total of 119 students agreed to participate in the study, 3 arriving without a partner.

Subjects were run in small groups ranging in number from 6 to 12 members per group. Participants were seated in a circle around a large table, each behind a plaque bearing a fictitious name. Subjects were instructed to refrain from talking during the course of the experiment, and the rating tasks were administered immediately. They were first asked to rate each other on an anchored 9-point degree of acquaintance-
ship scale (1 indicated that the ratee was known not at all and 9 meant that the person was known extremely well).

Subjects were next administered a battery of self-ratings. During that time, the experimenter tabulated the degree of acquaintanceship ratings and used this information to assign subjects into dyads for peer ratings. This assignment was made so that some target–rater dyads represented Acquaintanceship Level 1, others represented Level 2, and so on for all nine levels. Immediately after subjects completed their self-ratings, the experimenter administered a peer rating form. Peer ratings were collected on the same battery of measures used for self-ratings.

This method of assigning subjects into distinct acquaintanceship groups raised two minor problems. First, such ratings of acquaintance-ship are not entirely symmetric because, for example, the first member of a dyad may profess to know the other member better than the other claims to know the first. I therefore assigned subjects into dyads by selecting for a particular acquaintanceship rating, say 6, and then choosing a person who recorded a rating of 6 as the judge and the recipient of the rating as the target. The second problem was that, in the effort to represent each level of acquaintanceship equally, some targets were rated by two judges (n = 13) and one target was rated by three judges. I decided to treat these as independent dyads, and consequently these few targets were instated in more than one of the nine acquaintanceship cells.

Another potential problem is that the degree of acquaintanceship scale used may measure a rater’s perception of how accurately that person thinks he or she can judge the target rather than the duration of their acquaintanceship. A simple length of acquaintanceship measure, however, raises the possibility of some subjects receiving high scores on the scale even though they have a narrow knowledge of the target due to situational constraints tied to, say, classroom settings. The global measure of acquaintanceship used in this study was intended to capture, at once, an overall judgment of target–rater familiarity, intimacy, and extent of contact.

**Personality Rating Materials**

The measure of personality used in this study was a half-form of Jackson’s (1984) PRF Form E. This 176-item measure contains 8 behavior-based items for each of 20 bipolar content scales and two validity scales. The half-form was constructed by dividing the 16 items on each scale of the published form into 2, such that each half-scale has an equal number of true- and false-keyed items, and each is equivalent in mean item endorsement probability (Reed, 1976). The PRF has an established level of construct validity, proven to have utility in predicting criteria such as marital adjustment (Meyer & Pepper, 1977), interpersonal sociability (Gifford, 1981, 1982), employment preferences (Siess & Jackson, 1970), and behavior ratings of acquaintances (Harris, 1980; Paunonen & Jackson, 1985a).

In this study, both self-ratings and peer ratings were recorded for each of the 176 PRF items using anchored 9-point rating scales. For example, subjects read the statement “My personal papers are usually in a state of confusion” and then chose a number on the rating scale (1 represented an extremely uncharacteristic behavior of the person being rated [self or other], 9 indicated extremely characteristic, and 5 represented neutral).

**Observability Ratings**

An independent group of 7 undergraduates was asked to estimate the public observability of the 20 traits represented by the PRF content scales. However, because the observability of behaviors representing one pole of a trait continuum may differ from the observability of those at the other pole, both ends of each dimension were rated. Two adjectives for each PRF trait, used formerly by Paunonen (1988), were presented to judges. For example, the PRF scale of Abasement was represented by the adjectives meek and arrogant. Subjects then rated the 40 trait-descriptive adjectives (9-point scales) on “how publicly observable or visible to others are the behaviors related to the trait[s].”

**Results**

**Observability Ratings**

The seven judges who rated the trait adjectives on public observability agreed substantially in their estimates. I computed rater agreement by correcting the mean interjudge correlation with the Spearman-Brown formula. The reliability of ratings for the 20 trait adjectives representing the poles in the direction of which the PRF scales are scored was .77; the reliability of ratings for the 20 adjectives representing the other poles was .90; and the reliability computed across all 40 descriptors was .86. The correlation of mean ratings for one pole with mean ratings for the other pole was .54, indicating less than perfect correspondence between the observability of behaviors defining opposite ends of trait continua. As reported by Funder and Dobroth (1987), the two observability ratings for each trait were averaged for subsequent analyses.

**Rater Agreement and Acquaintanceship**

Self-ratings and peer ratings on the PRF items were first scored on their appropriate scales. One of these is the Infrequency scale, on which a high score indicates random responding, failure to comprehend written English, or deliberate falsification of responses (Jackson, 1984). Each target–rater dyad was checked for high Infrequency scale scores on both the target’s self-ratings (criterion) and the judge’s peer ratings (predictor). I decided to remove any dyad from the analysis if either predictor or criterion personality profile included a mean Infrequency scale score greater than 4.0 on the 9-point scale. This procedure resulted in the elimination of 8 dyads, leaving a total of 111 subjects. The final frequency distribution across the nine levels of acquaintanceship (from 1 to 9) was 13, 14, 12, 9, 10, 13, 13, 14, and 13 dyads.

I determined the correlation between self-ratings and peer ratings in two ways for each acquaintanceship subgroup. First, self-ratings were correlated with peer ratings across the subjects within a group separately for each PRF trait. Second, the self-report personality profiles were correlated with the corresponding peer estimated profiles across the 20 PRF traits separately for each dyad (see also Paunonen et al., 1988). The first method is more commonly reported in this type of research and has the advantage of allowing tests of agreement as a function of moderators using moderated multiple regression analysis. The second method, however, has the advantage of allowing certain specific rater or target factors to be statistically partitioned out of the agreement correlations.

**Trait correlations.** The mean self–peer trait correlations, averaged across 20 PRF dimensions, are presented in Figure 2 by degree of acquaintanceship. Overall, the mean agreement between self and peer was .22, but this varied as a function of degree of acquaintanceship, as seen in Figure 2, with higher agreements generally evident for higher levels of acquaintanceship. An F test evaluating agreement as a function of acquaintanceship, with traits as replications, yielded a significant main effect, F(8, 171) = 3.89, p < .0005. I also tested the individual trait correlations represented in Figure 2 for linear and qua-
component was highly significant ($p < .0005$), the quadratic component was not ($p > .05$).

I also evaluated self–peer rating agreement as a function of target–rater acquaintanceship using moderated multiple regression analysis. All 111 self-ratings on a trait scale were first regressed on corresponding peer ratings. Peer-rated acquaintanceship was then entered into the regression equation followed by the product Peer Trait Ratings $\times$ Acquaintanceship Ratings. It is this product term that carries the interaction of interest, and a moderator effect is revealed if the product adds significantly to criterion prediction beyond the linear contributions of the predictor variable (peer trait ratings) and the moderator variable (acquaintanceship ratings). A significant moderator effect, however, can refer to greater criterion–predictor correlation as the moderator increases in magnitude or to greater correlation as the moderator decreases in magnitude. The direction of the effect can be evaluated by examining the sign of the regression coefficient for the product term in the final regression equation. A more detailed exposition of moderated multiple regression analysis and its properties can be found in Cohen and Cohen (1983, chap. 8) and Paunonen and Jackson (1985a, 1988).

The results of 20 regressions evaluating degree of acquaintanceship as a moderator of rater agreement are presented in Table 1. The first column of numbers refers to the correlation between self-ratings and peer ratings on each PRF trait computed across all nine levels of acquaintanceship. The second column refers to increments in the criterion’s predictability due to the linear contribution of acquaintanceship ratings. (This contribution was negligible in all but 2 of 20 comparisons.) The third column of the table denotes increments in predictability due to the contribution of the interaction between peer trait ratings and acquaintanceship ratings. The last column lists the signs of the regression coefficients for the product terms in the regression equations. These signs should be positive if higher self–peer correlations are associated with higher levels of acquaintanceship.

I discovered significant moderator effects ($p < .05$) in the expected direction in 7 of 20 comparisons shown in Table 1. For those traits, agreement was found to be greater at greater levels of acquaintanceship. Furthermore, 16 of 20 tests, whether statistically significant or not, were in the predicted direction. Four comparisons were in the contrary direction, 1 of which was statistically significant; paradoxically, for the trait of cognitive structure, agreement was greater at lower levels of acquaintanceship.

The discovery of 7 in 20 significant moderator effects in the present analysis is considered noteworthy for the following reason. Reliable moderator and interaction effects have been notoriously elusive, even when proper multivariate procedures have been used to pursue them (e.g., Bobko, 1986; Sackett, Harris, & Orr, 1986). Indeed, some investigators have consequently questioned the assumed propriety of those procedures (e.g., Morris, Sherman, & Mansfield, 1986; but see Cronbach, 1987; Paunonen & Jackson, 1988). As a case in point, multiple evaluations of the popular hypothesis that behavior consistency moderates behavior predictability reported by Paunonen and Jackson (1985a) found only two significant moderators in 40 multivariate tests, one of which was in the unexpected direction. Also, of 11 evaluations for each of three putative moderators of predictability, Zuckerman et al. (1988) found six significant effects for a measure of trait relevance as a moderator, but only two effects for consistency and one for observability. (Supplementary analyses, however, led Zuckerman et al., 1988, to conclude that all three variables were viable moderators of behavior predictability.)

**Profile correlations.** I also assessed agreement among peers by correlating the targets’ 20-trait PRF profiles derived from self-ratings and from peer ratings. The mean self–peer PRF profile correlations, averaged across the raters within an acquaintance.
tanceship group (ns = 9–14 per group), are illustrated in the upper curve of Figure 3 (open circles). The overall mean profile correlation was .31, but again the value is observed to vary with target–rater acquaintanceship. An F test on the individuals’ agreement coefficients yielded a significant main effect for acquaintanceship, $F(8, 102) = 3.09, p < .005$. A multivariate regression analysis on those coefficients showed a significant linear component to the curve ($p < .0005$) but no quadratic component ($p > .05$).

It is possible that the generally positive correlations between the self-ratings and peer ratings of this study are due in part to a stereotype component pervading the judgments. To illustrate, an observer’s peer ratings could correlate substantially with a stranger’s self-ratings if (a) the observer guessed at each rating by correctly estimating the normative response for the item (say, the average male student’s response), and (b) the stranger target were similar to the group norm and, consequently, endorsed the normative responses. Note, however, that guessing in a person perception task by estimating normative responses or base-rate components of accuracy referred to by Cronbach, 1955, are not relevant to the correlation-based indexes used in this study.)

The base-rate problem is one that Cronbach (1955) addressed in his critique of accuracy studies and his designation of the distinction between differential and stereotype components of accuracy. A correlation computed between self- and peer ratings is an index of overall accuracy, which is composed of both differential and stereotype components. (The elevation components of accuracy referred to by Cronbach, 1955, are not relevant to the correlation-based indexes used in this study.) Whereas stereotype accuracy can accrue in the absence of any specific knowledge of a particular target’s characteristics, as illustrated in the example given, differential accuracy cannot; it is that portion of overall accuracy that is free of the stereotype component and represents the rater’s knowledge of how the target is distinct from the group norm.

I computed differential agreement coefficients for the judgments of the present study by subtracting the contribution of stereotype agreement from the overall agreement coefficients shown in the upper curve of Figure 3. I accomplished this by partialing university student base rates from the self-rated profiles and the peer-rated profiles before recomputing self–peer profile correlations (Paunonen et al., 1988). I estimated these base rates using published PRF student norms (Jackson, 1984, Table 4). Male subjects’ norms were applied to male targets’ ratings and female subjects’ norms to female targets. The resultant mean partial correlations, averaged across the dyads within each acquaintanceship group, are illustrated in the lower curve of Figure 3 (filled circles). An F test of the individual raters’ differential agreement coefficients produced a significant main effect of acquaintanceship, $F(8, 102) = 3.16, p < .005$. Trend analysis showed a significant linear component ($p < .0005$) but no quadratic component ($p > .05$).

The mean differential agreement computed across all subjects was .26 (lower curve of Figure 3), which compares with the mean overall agreement of .31 (upper curve). Clearly, there was a general stereotype component that contributed positively to self–peer rater agreement, although the amount was not large. Suppressors effects have been reported in this type of analysis (Paunonen et al., 1988) in which differential accuracy or agreement is greater than overall accuracy. This occurs when, as described earlier, base-rate judgments are inappropriate in rating a (normatively deviant) target. One would not expect these suppressors to appear when averaging accuracy or agreement coefficients across randomly sampled target–rater dyads, however, because the targets will in fact somewhat resemble group norms on average. The mean correlation between the self-rating personality profiles and the vector of PRF norms for the appropriate sex was .23 in this study (range = -.35–.76)

An inspection of Figure 3 shows that the two highest acquaintanceship groups had the greatest amounts of differential agreement compared with their overall agreement. That is, these subjects’ ratings were on the whole characterized by relatively small stereotype components. This observation would make sense theoretically because the most-acquainted subjects should best know how their peers differ from and are similar to group norms. This observation was not supported, however, by a statistical test of the correlation between the 111 dyads’ acquaintanceship ratings and a difference score computed by subtracting each rater’s differential agreement from his or her overall agreement ($r = -.085, p > .05$).

It is possible that overall accuracy or agreement in ratings could be in part due to a stereotype component that is less related to base rates of behavior and more related to the social desirability of behavior. That is, a rater could be partly accurate if he or she were to guess the socially desirable response option for an item and the target were to endorse the same option (which self-raters are wont to do). Funder (1980; Funder & Colvin, 1988) has reported finding some evidence of agreement as

![Figure 3. Mean self–peer profile correlations as a function of degree of acquaintanceship (open circles). (Mean partial correlations, with base rates removed, shown as filled circles.)](image-url)
Accordingly, I partialed the social desirability scale values of the PRF traits (Jackson, 1984, Table 16) from the self- and peer-rated profiles before recomputing self-peer agreement correlations.

Differential agreement coefficients with social desirability partialed out yielded results similar to those involving base rates as a covariate. The mean self-peer partial correlation across all subjects was .26, which is slightly lower than the mean overall agreement of .31 and equal to the mean differential agreement with base rates removed (Figure 3). Furthermore, I found a main effect of acquaintanceship on differential agreement, F(8, 102) = 3.43, p < .005, such that higher agreements were again associated with higher degrees of acquaintanceship. I found the linear trend to be strong (p < .0005) with no quadratic trend apparent (p > .05).

When both base rates and desirability were simultaneously partialed from the ratings, the mean differential agreement was .24, a value not much less than the .26 value found when either base rates or desirability alone were removed. Moreover, analysis of the main effect of acquaintanceship on agreement and evaluations of the trend components replicated those findings already reported for overall agreement, and for differential agreement when base rates or desirability alone were used as covariates.

**Rater Agreement and Observability**

I evaluated trait observability as a moderator of the correlation between judges' ratings and their peers' self-ratings. As reasoned earlier, observer ratings of a target's more visible behaviors should correspond more closely to self-ratings of those behaviors compared with ratings of less salient behaviors. I first computed overall agreement coefficients for each of the 20 traits across all 111 subject dyads. As shown in the first column of Table 1, these correlations ranged from .08 to .42 with a mean of .24. I then correlated these individual trait agreement coefficients with the mean observability ratings across the 20 traits. The resultant correlation yielded a value of 0 to two decimal places. It appears that, contrary to expectations, agreement among peer ratings was not a function of trait-behavior observability in the present data.

**Agreement and the Interaction of Observability and Acquaintanceship**

Recall that I hypothesized that the acquaintanceship and observability variables would interact in their effects on peer rating accuracy and agreement. I expected this because observability should be more of a determinant of rater accuracy at low levels of acquaintanceship and less at high levels. Increasing acquaintanceship should lead to increasing familiarity with a target's traits that are inconspicuous but nonetheless characteristic. To test for this possibility, I first computed agreement coefficients for each trait in each acquaintanceship group separately. I then attempted to predict these coefficients with level of acquaintanceship and observability in a moderated multiple regression analysis. Thus, each of the 180 agreement coefficients (9 acquaintanceship cells × 20 traits) was associated with the level of acquaintanceship under which it was computed and the mean observability rating of the particular trait dimension. I then entered these variables into a hierarchical regression in which agreement was regressed on acquaintanceship, observability, and the product of Acquaintanceship × Observability.

As described earlier, the test of interest in a moderated regression analysis is that which involves the increment in criterion predictability owing to the inclusion of the product term in the regression equation. In the present analysis, that term did contribute significantly to prediction, F(1, 176) = 3.75, p < .05. The empirical response surface for predicting self-peer agreement with acquaintanceship and observability, as computed for the data of this study, is shown in Figure 4. Such figures are instructive because they help to clarify the nature of a moderator's operation in a particular data set. They are also instructive because they show how the method of moderated multiple regression works to distort, geometrically, the usual flat three-dimensional plane of a normal three-variable multiple regression. The lines on the surface in Figure 4 are parallel and straight.

Notice the direction of the moderator effect in Figure 4. At low levels of acquaintanceship, agreement improves as trait observability increases. The effect gradually diminishes, however, as target–rater acquaintanceship is enhanced. For those raters falling above the Acquaintanceship scale's midpoint of 5, for example, their agreement coefficients correlated −.06 (p > .05) with the observability of the rated domain. The corresponding correlation was .26 (p < .01) for raters falling below 5 on the Acquaintanceship scale.

The observation that agreement is attenuated somewhat on the more observable traits at high levels of acquaintanceship may largely be due to ratings on one trait: cognitive structure. In the moderated regressions reported in Table 1, it was seen that self- and peer ratings on this generally unobservable domain of behaviors (having to do with planfulness and the need for certainty) were significantly (and paradoxically) more in agreement in the lower acquaintanceship groups than in the higher acquaintanceship groups. Another possible reason for the apparent decline in agreement as a function of observability
for the high acquaintanceship group is presented in the next section of this article.

Discussion

I outlined three research hypotheses in the introduction to this article, based on derivations from the lens model of inferential accuracy. I expected that, in a person perception context, the correlation between a peer’s ratings of a target person’s personality traits and the target’s self-ratings would be a function of (a) level of target–rater acquaintanceship, (b) observability of the behavior domain being rated, and (c) the interaction of acquaintanceship and observability. Only the first and last expectation received empirical support in the data reported herein. The implications of these findings for past and future research on accuracy and consensus in person perception are briefly described.

Accuracy and Acquaintanceship

The fact that self–peer correlations were found to vary directly with acquaintanceship is encouraging for at least two reasons. First, although intuitive, there is no guarantee that friends are better judges of each other than are casual acquaintances. Empirical tests of the hypothesis have sometimes produced counterintuitive results (e.g., Gormly & Edelberg, 1974; Stemachers & McHugh, 1964). Also, rational arguments have presented reasonable grounds for expecting friends to be relatively misguided judges. Cloyd (1977) and Taft (1966) have both cited reasons why close acquaintances may at times go beyond available target cues and make inferences that are both unwarranted and inaccurate. This effect has been explained as a tendency for close acquaintances to overemphasize irrelevant target cues or to invoke overly desirable stereotypes in making their judgments. These rational arguments, however, were not supported by this study’s data (with one exception to be discussed later). The arguments imply no relation between accuracy and acquaintanceship, or perhaps a curvilinear inverted-U relation, with accuracy and agreement being maximal at some moderate level of acquaintanceship. Several tests of the nature of the relation across the nine levels of acquaintanceship revealed only positive linear functions in the present data. Furthermore, social desirability was found not to be a large component in the ratings of friends or of strangers.

There is a second reason for finding encouragement in the observation that rater agreement varies positively with acquaintanceship. This concerns the widely held belief that personality ratings are subject to distortions due to the semantic or conceptual similarity of the words incorporated in the rating scales used to record such judgments (e.g., D’Andrade, 1974; Shweder, 1975). The nature of this hypothesis is that ratings of close friends and strangers alike are largely determined by a perceived network of trait relations. This network is organized not according to how traits actually co-occur in people but according to universally shared elements of linguistic meaning that are inherent to trait descriptive labels. A cogent discussion of early studies giving rise to the semantic hypothesis can be found in Schneider (1973) and in Wiggins (1973, chap. 8).

One of the major problems with the semantic similarity hypothesis is that the network of word meanings is thought to be mostly independent of the network of actual trait co-occurrences, but no explanation is provided for how the former is established in the minds of observers (Jackson & Paunonen, 1980). The linguistic associations thought to characterize the semantic network cannot be due to inductive generalizations of real-world events because if they were the semantic network would perforce correspond to the actual network of trait associations. A second problem with the semantic hypothesis becomes apparent when evaluating this study’s results. That hypothesis has no plausible means for explaining why greater degrees of target–rater acquaintanceship should lead to greater degrees of self–peer correlation. Proponents of the theory would have to argue that, somehow, increasing acquaintanceship causes a peer rater increasingly to ascribe to the target the same linguistically or conceptually related traits as does the target him- or herself for self-ratings. A more plausible interpretation of the data is that increasing acquaintanceship indeed causes the peer to ascribe similar traits to the target as compared with self-ratings, but that those traits are veridical in the sense that they actually characterize the behavioral predispositions of the target.

Accuracy and Observability

This study found no evidence for the hypothesis that trait observability by itself moderates self–peer correlations, in contrast to data reported elsewhere. In commenting on this discrepancy, it is important to note that some other studies also did not find the expected moderator effect (Funder, 1980; Kammann et al., 1984). Moreover, some of those that did were criticized in the introduction to this article either for using inappropriate measures of the observability of bipolar traits (Zuckerman et al., 1988) or for using inappropriate methods of evaluating moderator effects (Kenrick & Stringfield, 1980). The studies by Funder and Dobroth (1987) and Funder and Colvin (1988), however, are the nearest approximations to the present study in terms of evaluating observability as a moderator variable. Those studies did find the expected moderator effect, reporting correlations of .39 and .41, respectively, between a composite trait observability measure and self–peer agreement coefficients.

The difference in the present data and those of Funder and his colleagues may be due to differences in the levels of observability for the traits used. The PRF trait with the lowest mean observability rating on the 9-point scale in this study was change (M = 3.29), which refers to a dislike for routine. However, Funder’s traits represented 100 modified California Q-set items, some of which referred to such highly unobservable dispositions as having “insight into own motives and behavior” and being “uncomfortable with uncertainty.” It is possible that the PRF traits used in this study represent a restricted range on the observability continuum, compared with the larger set of traits used by Funder, and that this restriction of range produced an attenuated correlation between observability and intrarater agreement. It is also possible that the relatively small number of traits used in this investigation (N = 20) made for a relatively weak statistical test of the hypothesis (however, cf. Kenrick & Stringfield, 1980, Table 8).

The difference in the present observability findings and those of other researchers may also reside in the Acquaintanceship X
Observability interaction found in this study. Recall that agreement was seen to vary as a function of trait observability for low to moderately high levels of acquaintanceship (Figure 4). However, at higher levels of acquaintanceship, agreement was not related to observability, extending to an apparent ceiling across the length of the observability continuum. In fact, at the highest levels, agreement seemed to decrease somewhat with observability. It is conceivable that the subjects in Funder's studies were not as highly acquainted as were the closest acquaintances in this study. This would account for their having found a positive correlation between agreement and trait observability, a finding corroborated by the present data for low to moderately acquainted target–rater dyads. Funder and Dobroth (1987) reported their subjects to be "roommates or close friends" (p. 412) but provided no quantitative data on the level of acquaintanceship that characterized their sample. Funder and Colvin (1988) used similar subjects, reporting a mean length of acquaintanceship of 18.5 months, but 56% of their subjects described themselves simply as "friends" (p. 151). As stated earlier, a long period of acquaintanceship with a person is no guarantee of a broad knowledge of that individual's personality characteristics.

The interactive effect of acquaintanceship and observability on agreement merits further discussion. I predicted this effect because it seemed reasonable that close acquaintances should be privy to more covert and latent behavioral tendencies of others than should more distant acquaintances. Such a broadening of knowledge derives from increased experience with the behaviors of the target person. If the moderator effect reported in this study proves reliable, then it has obvious implications for the study of observer judgments. Past studies that have failed to provide demonstrable support for the accuracy of personality judgments may have used subjects who were less than optimally acquainted and behavior domains that were less than optimally salient. Unless one condition or the other is reversed, accuracy coefficients may be severely attenuated. Another implication of the interaction of acquaintanceship and observability on person perception accuracy pertains to formal (and informal) interviewing practices. As most interviews are conducted with a stranger or a new acquaintance, it would seem reasonable to structure verbal inquiries around personality dimensions for which relevant behaviors are not likely to occur normally in the presence of the interviewer. In an employment context, for example, extant cues about the interviewee's congeniality, neatness, argumentativeness, and defensiveness might be sufficient to obviate a lengthy structured scrutiny of these traits. The interview time would be better spent on less obvious (job-relevant) personal characteristics.

A final word is in order regarding the finding that the most highly acquainted judges were most in agreement on the least observable traits and somewhat less so on the more observable traits. One explanation for this seemingly paradoxical outcome is the possibility that people may recognize a public persona in their close friends. Such perceivers may believe that their peers' apparent standpoints on publicly observable behavior domains are partly facades adopted by the actor to maintain a desired self-image, public posture, or social appearance. In this sense, the perceivers could go beyond the target cues and make inferences about personal characteristics that may or may not be warranted (Cloyd, 1977; Taft, 1966). If the criterion measure of personality is able to penetrate a target's facade, then an astute perceiver's personality inferences will manifest some measure of empirical accuracy. If the actor is able to maintain the public persona in his or her self-ratings, however, the empirical accuracy of the peer's judgments will necessarily be attenuated. This explanation is speculation at this point and is based on a result of this study that requires replication.

References


Received July 7, 1988
Revision received November 7, 1988
Accepted November 16, 1988