Determinants of Interjudge Agreement on Personality Traits: The Big Five Domains, Observability, Evaluativeness, and the Unique Perspective of the Self

Oliver P. John
Richard W. Robins
University of California, Berkeley

ABSTRACT We examined several determinants of interjudge agreement on personality traits. The findings, which were cross-validated in two samples, suggest that agreement is a function of four factors: which Big Five content domain the trait represents, how observable relevant behaviors are, how evaluative the trait is, and whether the self is one of the judges. Agreement was highest for traits related to Extraversion and lowest for traits related to Agreeableness. More observable and less evaluative traits elicited higher interjudge

This research was supported by National Institute of Mental Health Grant MH49255; additional support came from Grants MH40662 and MH39077. The second author was supported by a National Science Foundation Graduate Fellowship. We are indebted to Lewis R. Goldberg, Zareen Farooqi, and Myron Rothbart, who helped us collect the data analyzed here, and to Robert R. McCrae and David C. Funder, who generously provided us with their data for the present reanalyses. Maureen Barckley contributed invaluable programming assistance. This article has benefited from thoughtful comments by William F. Chaplin, Kenneth H. Craik, David C. Funder, Lewis R. Goldberg, Robert Hogan, Delroy Paulhus, Robert R. McCrae, David Watson, and Stephen G. West. A portion of these findings were presented at the 1988 midwinter meetings of the Society for Personality Assessment in New Orleans, March 10–14, 1988, and at the annual meeting of the Society for Experimental Social Psychology, Madison, WI, October 20–23, 1988. Correspondence concerning this article should be addressed to Oliver P. John and Richard W. Robins, Department of Psychology, University of California, Berkeley, CA 94720.

Journal of Personality 61:4, December 1993. Copyright © 1993 by Duke University Press. CCC 0022-3506/93/$1.50

agreement. On average, self-peer agreement was lower than peer-peer agreement. However, this effect was limited to evaluative traits; for neutral traits, self-peer agreement was as high as peer-peer agreement. These findings suggest that self- and peer perception proceed through similar processes for neutral traits but not for highly evaluative traits, raising the possibility that self-perceptions become distorted when the trait is affectively charged.

Much of personality research makes use of judgments made either by the subjects themselves or by knowledgeable others such as peers, spouses, parents, and psychologists. Judgments by self and others are an indispensable methodological tool for researchers in the social sciences (e.g., Craik, 1986); or, as Kenny (1991) put it, "As biologists use electron microscopes and chemists use mass spectrometers, the most valued 'instrument' used by psychologists is the human observer" (p. 156). Central to the scientific use of such judgments is the demonstration that the perceptions of different judges agree with each other and reflect social reality.

In the context of the behavioral consistency debate in the 1970s, the convergence of self-reports with ratings by others became an issue of contention (e.g., Bem & Allen, 1974). For example, Shraugher and Schoeneman (1979) concluded that "there is no consistent agreement between people's self-perceptions and how they are actually perceived by others" (p. 549). However, their conclusions were criticized for a variety of reasons, and in the 1980s a flurry of studies showed that interjudge agreement on personality trait ratings is almost always statistically significant and often substantial in size, particularly when the judges are well-acquainted with the targets and the ratings are aggregated across items and judges (see Cheek, 1982; Funder, 1987; McCrae, 1982). McCrae and Costa (1989) reviewed 10 recent studies of interjudge agreement on personality traits and found a mean agreement correlation of .45. Similarly, Funder (1987) noted that agreement correlations "tend to be on the order of .30 to .60" (p. 84). Thus, the statistical significance and magnitude of interjudge agreement correlations are no longer at issue.

Subsequent research has begun to delineate the conditions under which agreement is relatively high or relatively low, and the present article builds on this research. In addition, we propose two new determinants of agreement, both based on the general idea that judgments about the self differ from judgments by others. In two empirical studies, we replicate determinants of agreement proposed in earlier research, examine the determinants newly proposed here, and test an integrative model specifying how these factors jointly influence interjudge agreement.

Previous Research on Factors Affecting Interjudge Agreement

Interjudge agreement (or consensus) reflects the degree to which judges agree in the relative ordering of target persons on a trait dimension. A number of recent studies have suggested that high agreement should not always be expected but rather depends on at least five factors: (*a*) the content domain of the trait judged; (*b*) the observability of trait-relevant behaviors; (*c*) the social desirability of the trait; (*d*) the level of acquaintanceship between judge and target; and (*e*) individual differences in the judgability of the target person.

First, interjudge agreement is higher for some content domains than for others. Norman and Goldberg (1966) examined interjudge agreement on the Big Five dimensions measured by 20 bipolar scales. In two samples, they found the highest agreement for Extraversion (e.g., talkative vs. silent) and the lowest agreement for Emotional Stability (calm vs. anxious) and Agreeableness (good-natured vs. irritable); Conscientiousness (responsible vs. undependable) and Intellect (intellectual vs. unreflective, narrow) fell in between. Several studies have since replicated these findings. In a study using the California Q set (Block, 1978), Funder and Dobroth (1987) found that self-peer and peer-peer agreement was highest for Q-sort items related to Extraversion (e.g., talkative, gregarious, socially poised), and lowest for items related to Neuroticism (e.g., thin-skinned, basically anxious, irritable), which is the low pole of the Emotional Stability dimension. Agreement between self-ratings and ratings by "strangers" varies across the Big Five domains in much the same way as agreement between well-acquainted individuals (Albright, Kenny, & Malloy, 1988; Funder & Colvin, 1988; Watson, 1989).

Second, several studies suggest that agreement is higher for observable traits than for less observable traits (Funder & Colvin, 1988; Funder & Dobroth, 1987; Paunonen, 1989). Using a measure they interpret as "apparent ease of observation or easy visibility" (p. 415), Funder and Dobroth (1987) found that more observable Q-sort items elicited higher levels of agreement than less observable items. Funder and Colvin (1988) replicated this effect with judges who were unacquainted with the target. Across the 100 Q-sort items, the correlations between interjudge agreement and observability ranged from .25 to .43.

Third, the effect of another trait property, social desirability, has received less attention. Funder and his colleagues (Funder, 1980; Funder & Colvin, 1988; Funder & Dobroth, 1987) have examined the relation between desirability (or favorability) and agreement in several studies of the 100 items of the California Q set. The correlation between the social desirability of a Q-sort item and self-other agreement on that item ranged from .30 to .43, suggesting a linear relation between social desirability and agreement. However, this effect did not generalize to peer-peer agreement; the correlation was .15 (Funder & Colvin, 1988, Table 2).

Fourth, well-acquainted individuals agree in their judgments to a greater extent than less well-acquainted individuals (Colvin & Funder, 1991; Funder & Colvin, 1988; Jackson, Neill, & Bevan, 1973; Norman & Goldberg, 1966; Paunonen, 1989; Watson, 1989). Norman and Goldberg (1966) were probably the first to demonstrate this "acquaintanceship effect." Self-peer agreement on the Big Five dimensions was much higher in a sample of Peace Corps volunteers who had trained together for 3 months than in a sample of undergraduate students who rated each other on their first day of class. Similarly, Funder and Colvin (1988) found that agreement correlations between self and close friends averaged .27 across the 100 Q items, whereas agreement between self and strangers, who had observed only a 5-minute videotape of the subject's behavior, averaged .05.

Fifth, agreement seems to be higher for some individuals than for others (Cheek, 1982; Colvin, 1993; Kenrick & Stringfield, 1980; Petersen, 1965). In an extension of Bem and Allen (1974), Kenrick and Stringfield (1980) had subjects rate their cross-situational consistency and reported that consistent subjects showed higher levels of agreement with ratings by their peers and parents (but see Chaplin & Goldberg, 1984). Colvin (1993) has examined the personality profile of the highly judgable person and found reliable individual differences in judgability that generalized across indices.

Much of the research reviewed above has been conducted with ratings of Q-sort items or with small sets of bipolar trait scales. Paunonen (1989) has argued that studies of observability and interjudge agreement need to take into account the bipolarity of trait dimensions because "the two poles of a trait may engender different levels of behavioral expression and visibility" (p. 826). One purpose of the present research, therefore, is to replicate the previous findings linking interjudge agreement to Big Five content domain, observability, and social desirability, using a large set of unipolar trait adjectives selected a priori to represent both the high and low poles of each of the Big Five domains.

In addition, we consider potential differences between two kinds of interjudge agreement: peer-peer and self-peer agreement. Although these two kinds of agreement are often analyzed and reported separately, theoretical and empirical analyses of the sources of differences between the two are rarely undertaken. However, important differences between self and other judges of personality may exist. The present article examines whether the self is a unique judge of personality by comparing self-peer to peer-peer agreement for a large set of personality traits.

Is the Self a Unique Judge of Personality?

Hypothesis 1: Self-peer agreement is generally lower than peer-peer agreement. Do people perceive themselves in the same way they perceive others, or is the self a unique judge of personality? The central thesis of the present research is that self- and other perceptions do not always proceed through the same processes. Previous research suggests that self-perceptions differ from perceptions of others in at least three fundamental ways. First, the self has information available from prior experiences and access to internal thoughts, intentions, and other "privileged" information, none of which are available to an external observer (Jones & Nisbett, 1971). Overall, then, the self has greater access to self-relevant information than others do. Second, the self does not have the same visual perspective as others; people typically do not observe their behavior from the perspective of an external observer (Storms, 1973; see also Robins & John, 1993). Consequently, different personality-relevant information may be available and salient to the self than to others. Third, individuals are more ego-involved in their self-evaluations than in their evaluations of others; consequently, self-perceptions may be influenced by motivational factors, such as self-esteem needs, that do not influence perceptions of others (Taylor & Brown, 1988). Each of these three differences points to the *general* prediction that agreement between self and a peer should be lower than agreement between two peers.

However, this prediction is complicated by other factors that influence interjudge agreement in general, such as acquaintanceship and information overlap (cf. Kenny, 1991). Self-peer agreement may exceed peer-peer agreement under certain conditions, such as when there is considerable overlap in the information available to self and each individual peer but little overlap in the information available to the peers (e.g., when the peers know the target in different contexts). In the present research context, however, the peer judges were well-acquainted with the target subjects and knew them in similar contexts. Thus, factors that influence interjudge agreement in general (e.g., information overlap) are less central for the present research than factors that make the self a unique judge of personality traits (e.g., ego involvement).

Hypothesis 2: The self is biased when judgments are evaluative. Our first hypothesis states that the self is a unique judge of personality, with both assets (e.g., more information) and limitations (e.g., motivational biases) that may serve to attenuate agreement with others. A stronger position, held by many psychologists, is that the self is generally biased and therefore less accurate than others. James (1890) commented on the "selective industry of the mind," Allport (1958) was suspicious about "the self-report of the subject, who is capable of self-deception" (p. 243), and Greenwald (1980) emphasized that the self distorts reality in the service of a "totalitarian ego." "Due to self-deception, selective inattention, repression, or whatever one wishes to call lack of enlightenment, self-views may be less accurate than are outsiders' views" (Thorne, 1989, p. 157).

Indeed, most self-concept theorists assume that people are motivated to maintain and enhance their self-esteem (e.g., Greenwald, 1980; James, 1890; Rogers, 1959; Tesser, 1988). Several studies have demonstrated that ego involvement increases self-serving attributional biases (e.g., Miller, 1976), suggesting that self-perceptions are more prone to distortion when the stimuli are ego-involving (i.e., relevant to feelings of self-worth). Similarly, some self-esteem maintenance processes operate only when the dimension being judged is important to the self (for a review, see Tesser, 1988). This research has important implications for the role of evaluative processes in self-other agreement on personality trait judgments.

Judging oneself on traits that are extremely evaluative (either desirable or undesirable) is more ego-involving than judging oneself on neutral traits. In contrast, making judgments about another person is typically less ego-involving than making judgments about the self, and therefore evaluativeness should have a weaker influence on peer judgments. Consequently, evaluative traits should produce more bias in self-perceptions than in peer perceptions, thus decreasing agreement between self and others. On the other hand, relatively neutral traits should not induce ego involvement, and therefore self-perceptions should derive from similar processes as peer perceptions, leading to higher levels of self-peer agreement.

Thus, we predict that self-peer agreement will be highest for neutral traits and will decrease as the evaluativeness of the trait being judged increases. That is, the relation between self-peer agreement and trait desirability should be curvilinear, as indicated by an inverse U-shaped function. This curvilinear relation should be weaker for peer-peer agreement. If the peers have no affective involvement with the target person, there should be no relation between evaluativeness and peer-peer agreement. However, because in most studies (including our own) the peers like the target person, we expect some relation between evaluativeness and peer-peer agreement. Thus, Hypothesis 2 states that both self-peer and peer-peer agreement will be related to evaluativeness, but this effect will be stronger for self-peer agreement.

Our second hypothesis specifies a potential boundary condition on Hypothesis 1. The prediction that self-peer and peer-peer agreement are differentially related to evaluativeness (Hypothesis 2) implies that the magnitude of the difference between self-peer and peer-peer agreement (Hypothesis 1) may depend on the evaluativeness of the trait. Thus, the difference between peer-peer and self-peer agreement should be most pronounced for extremely evaluative traits, whereas for neutral traits there should be little or no difference in agreement.

The present research investigates these two hypotheses. First, we examine the main effect of type of judge on interjudge agreement, predicting that self-peer agreement will be generally lower than peer-peer agreement. Second, we test whether this main effect is modified by the interaction between type of judge and the evaluativeness of the trait being judged; we predict that the difference between self-peer and peer-peer agreement will hold for evaluative traits but not for relatively neutral traits. Moreover, these two predictions will be tested in a model of interjudge agreement that incorporates the effects of three determinants from the previous literature—Big Five content domain, observability, and social desirability.

Study 1

METHOD

Subjects and Procedures

A total of 250 students (155 females and 95 males) from a large public university in the Northwest volunteered to participate. Subjects were recruited in groups of five and typically lived together (e.g., dormitories, cooperative housing). In each of the 50 groups, one subject (*self*) served as the target person and rated him or herself, and the other four subjects (*peers*) rated the target person. Both target and peer subjects completed their ratings in a university laboratory, and care was taken to ensure that subjects from the same group were not scheduled for the same session.

Subjects had known each other for at least one semester and were generally well-acquainted. Subjects reported a fairly high degree of familiarity with their peers (M = 3.8 on a 5-point familiarity scale, with 5 = very familiar, 3 = quite familiar, and 1 = slightly familiar), and generally liked each other (M = 4.6 on a 5-point likability scale, with 5 = like very much, 3 = neutral, and 1 = dislike strongly).

Measures

Self- and peer ratings. Personality ratings were obtained from both self and peers using a 9-step response scale, which ranged from "extremely uncharacteristic" to "extremely characteristic." We used a set of 100 unipolar trait adjectives; 80 of these were markers for the Big Five dimensions. To assess interjudge agreement separately for the high and low poles of each of the Big Five (e.g., Extraversion and Introversion), we selected 40 of Goldberg's (1983, 1992) bipolar scales (e.g., talkative-quiet) and administered them as 80 single traits (e.g., talkative). To disguise the bipolar structure of the stimulus set, we administered the 80 Big Five traits along with 20 filler items in a fixed random order, with the constraint that traits from the same bipolar scales were never presented adjacently. Thus, there were 16 unipolar adjective markers, 8 for the high pole and 8 for the low pole of each of the Big Five dimensions. Traits defining the high and the low pole of each domain included talkative and quiet for Extraversion (vs. Introversion), fair and unfair for Agreeableness (vs. Antagonism), well-organized and disorganized for Conscientiousness (vs. Lack of Direction), secure and insecure for Emotional Stability (vs. Neuroticism), and complex and simple for Intellect (vs. Simple-Mindedness).

The results of a factor analysis of the 80 adjectives using the 200 individual peer ratings were consistent with the a priori factor structure.[1] However, two

[1] The list of trait adjectives and the complete matrix of factor loadings are available from the authors. We use the label Intellect (rather than Openness to Experience) for

of Goldberg's bipolar scales (subjective-objective and selfless-selfish) did not retain their intended meanings when administered as single adjectives, and these four adjectives were omitted from the present analyses. This resulted in a set of 76 adjectives for the present analyses.

Interjudge agreement indices. For each of the 76 traits, we computed two agreement correlations, one representing the degree to which the peers agreed with each other about the target's personality and the other the degree to which the target's self-ratings agreed with the peer ratings. In previous studies of agreement, peer-peer agreement has been typically computed by correlating two individual judgments, whereas self-peer agreement has been typically computed by correlating the self-judgments with the aggregated (i.e., mean) peer judgments, which are more reliable if the peers show at least some consensus; thus, in these studies agreement between the self and the mean peer represents an overestimate of the actual agreement between the self and a single peer. To make the peer-peer and self-peer agreement indices comparable, we computed dyadic agreement correlations between pairs of judges. There were six possible pairwise agreement correlations among the four peers, which we averaged to form an overall index of peer-peer agreement. Similarly, there were four possible self-peer agreement correlations, which we averaged to form an overall index of self-peer agreement.[2]

Ratings of trait properties. Independent ratings of observability and social desirability were available for each of the 76 traits. For the observability ratings, judges were told that "some traits refer to behaviors that can be easily observed by an outside observer. Other traits refer to behaviors that can be observed only by the person himself or herself." The judges rated each trait on a 9-point scale, ranging from 1 (extremely difficult to observe by an outside observer) to 9 (extremely easy to observe). The composite ratings of 28 judges had an alpha reliability of .90. The mean observability value for our 76 traits was 5.4 (SD = .9); sociable and talkative were the most observable traits, and complex and uncreative were the least observable traits.

Social desirability ratings were available from 100 undergraduates who had rated each trait on a scale from 1 (extremely undesirable) through 5 (neutral)

the fifth Big Five domain because our traits were selected from Goldberg's (1983, 1992) research, which emphasizes intellectual aspects such as intelligent, perceptive, knowledgeable, and cultured (see also John, 1990; McCrae & Costa, 1987).

[2] A reviewer suggested that the peer-peer agreement index may be slightly more reliable than the self-peer agreement index because it is an aggregate of six (rather than four) pairwise correlations. Any differences in reliability would not bias the peer-peer and self-peer agreement estimates for each trait, but could produce minor changes in the level of statistical significance in analyses across traits because the within-cell variance would be greater for the less reliable index.

to 9 (extremely desirable) (see Hampson, Goldberg, & John, 1987). The composite ratings had an alpha of .99. The mean desirability value for our 76 traits was 5.2 ($SD = 2.2$); 41 traits were rated as desirable (i.e., above the neutral midpoint of 5.0) and 35 as undesirable (i.e., below the midpoint). Intelligent and conscientious were most desirable, and ignorant and undependable were least desirable.

The evaluativeness of a trait reflects the degree to which the trait is evaluatively extreme (i.e., highly desirable or highly undesirable) versus relatively neutral in desirability. Evaluativeness was measured by the absolute value of the distance of the trait's desirability value from the neutral midpoint of 5.0 on the 9-point desirability scale. Ignorant and conscientious were the most evaluative traits, and impulsive and talkative were the least evaluative (i.e., most neutral). The intercorrelations among desirability, evaluativeness, and observability, computed across the 76 traits, were all below .16, indicating that these three trait properties were essentially unrelated in our set of traits.

RESULTS AND DISCUSSION

In all analyses, we used interjudge agreement correlations as the data, transformed via Fisher's r' to z' formula. The unit of analysis was the trait (not the individual subject), and the sample size for our analyses was thus the 76 personality traits.[3]

Effects of Big Five and Factor Pole on Interjudge Agreement

We conducted an analysis of variance (ANOVA) on the dyadic agreement correlations using the 76 traits as the unit of analysis; Big Five content (the five domains) and pole (high vs. low) were between-traits factors. For both the peer-peer and the self-peer agreement indices, the ANOVAs indicated a main effect for Big Five content domain, $F(4, 66) = 5.4$ and $F(4, 66) = 6.2$, both ps $< .01$, no main effect for factor pole, and no interaction. Thus, agreement differed across the Big Five dimensions, but not across the pairs of traits that marked the high and low factor poles. Moreover, there were no pronounced asymmetries be-

[3] Because the 76 trait agreement correlations are derived from the same sample of subjects, the units of analysis are technically not independent and the distribution of the F statistic in our analyses may differ from formal assumptions. Thus, tests of statistical significance should be interpreted cautiously, and are less informative than the effect sizes we report.

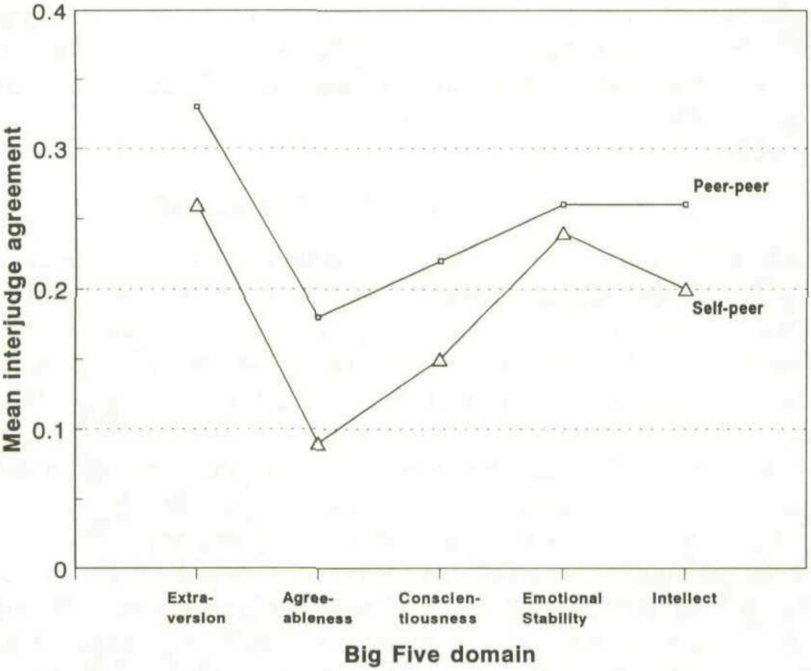

Figure 1
Peer-Peer and Self-Peer Agreement as a Function of the Big Five: Agreement Correlations Averaged across the Traits Representing Each of the Big Five Content Domains

tween the poles of any of the Big Five domains. The largest asymmetry was for Intellect; the average agreement correlation for traits from the high Intellect pole was .06 higher than the average for traits from the low pole.

The differences in agreement between the Big Five domains are illustrated in Figure 1, which presents the average self-peer and peer-peer agreement correlations for the traits representing each domain. Figure 1 shows that agreement was highest for traits related to Extraversion and lowest for traits related to Agreeableness, with traits related to Emotional Stability, Intellect, and Conscientiousness falling in between. Note that this ordering was identical for both the self-peer and the peer-peer agreement indices. Across the two indices, agreement for the average Extraversion trait was .29, whereas agreement for the average Agreeableness trait was .13. The difference between Extraversion and

Agreeableness was consistent across the individual traits; combining the two agreement indices, 81% of the Extraversion traits had agreement correlations exceeding that of the average trait, whereas this was true for only 14% of the Agreeableness traits.

Self-Peer versus Peer-Peer Agreement

Consistent with Hypothesis 1, Figure 1 shows that self-peer agreement was lower than peer-peer agreement for each of the Big Five domains. An ANOVA with type of judge (self-peer vs. peer-peer) as a within-traits factor and Big Five domain as a between-traits factor showed a significant main effect of type of judge, $F(1, 71) = 26.2$, $p < .001$, and no interaction with Big Five domain. Moreover, post hoc *t* tests showed that the difference between self-peer and peer-peer agreement was significant in every domain except for Emotional Stability.

The relation between peer-peer and self-peer agreement is illustrated further in Figure 2, which plots the agreement correlations for each of the 76 traits; peer-peer agreement is shown on the horizontal axis and self-peer agreement is shown on the vertical axis. Peer-peer agreement ranged from .02 to .50, with a mean of .25 ($SD = .10$). Self-peer agreement ranged from $-.07$ to .55, with a mean of .19 ($SD = .13$).

The dotted diagonal line from the lower left to the upper right of the figure is the unity line, representing the point at which peer-peer and self-peer agreement have the same value. In general, traits that elicited high peer-peer agreement also elicited high self-peer agreement, and vice versa. As one might expect from our Big Five agreement findings, many of the traits related to Extraversion (e.g., talkative, quiet, dominant, extraverted) are found in the upper-right corner of Figure 2, indicating high levels of both peer-peer and self-peer agreement. Across the 76 traits, the correlation between the peer-peer agreement index and the self-peer agreement index was .63 ($p < .01$).

In addition to showing the relation between the two indices, Figure 2 also shows which traits have relatively higher levels of peer-peer agreement (below the unity line) or self-peer agreement (above the unity line). Note that most traits fall below the unity line, reflecting the higher average peer-peer than self-peer agreement. Six traits even had negative self-peer agreement correlations, as indicated by their location below the dotted horizontal line.

These findings raise two questions: (*a*) Why do some traits generally

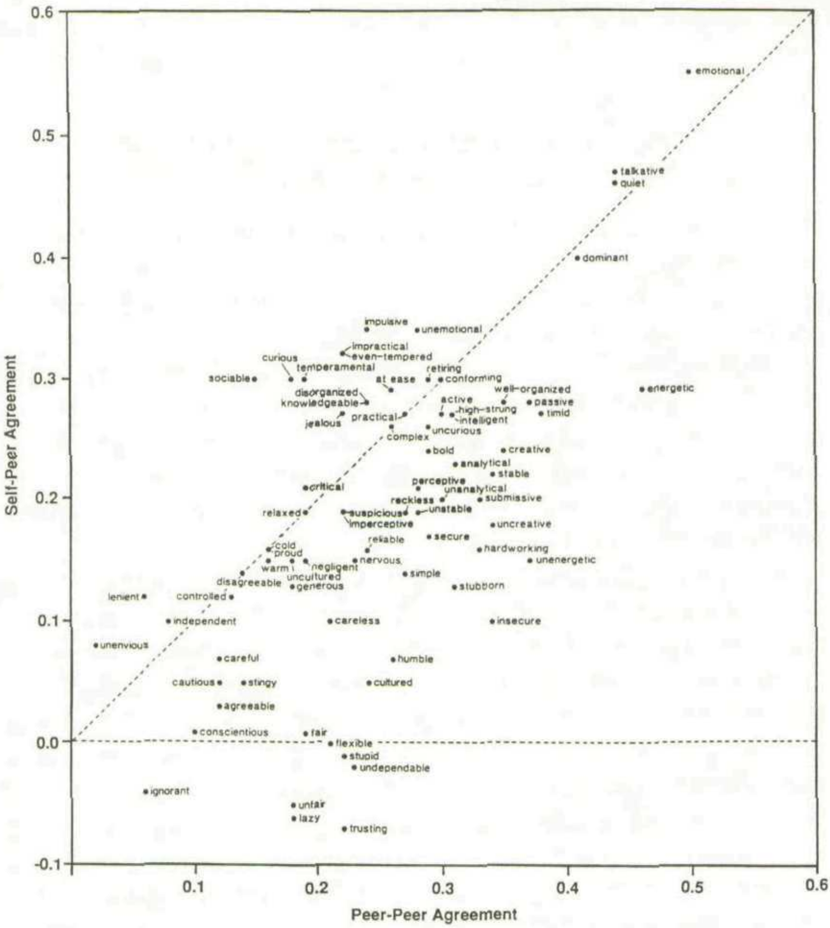

Figure 2

Relation of Peer-Peer Agreement (Horizontal Axis) to Self-Peer Agreement (Vertical Axis) for 76 Personality Traits

Note. The dotted diagonal line from the lower left to the upper right is the unity line, representing the point at which peer-peer and self-peer agreement have the same value. Traits falling below the unity line (two-thirds of the traits) have higher levels of peer-peer than self-peer agreement. The dotted horizontal line indicates the point at which self-peer agreement is zero.

elicit higher levels of agreement than others? and (*b*) Why does self-peer agreement appear to be generally lower than peer-peer agreement?

Effects of Observability, Social Desirability, and Evaluativeness on Agreement

Funder and Dobroth (1987) provided evidence linking the observability of a trait and its social desirability to interjudge agreement, and we predicted (Hypothesis 2) that evaluativeness will attenuate agreement, particularly when the self is one of the judges. To examine the degree to which agreement on a trait is influenced by its observability, social desirability, and evaluativeness, we correlated these three trait properties with self-peer and peer-peer agreement across the 76 traits. These correlations are shown in the first two columns of Table 1. (We emphasize that the values in Table 1 are *not* mean agreement correlations, which are shown in Figure 1.)

As expected, observability correlated positively with both peer-peer ($r = .36$) and self-peer agreement ($r = .38$); that is, agreement was higher on observable traits. This effect is illustrated more concretely by the difference in agreement between traits falling above versus below the median on observability. Peer-peer agreement averaged .27 for relatively observable traits, as contrasted with .22 for unobservable traits, and self-peer agreement averaged .22 for observable traits, as contrasted with .16 for unobservable traits.

As shown in Table 1, social desirability was not linearly related to either peer-peer or self-peer agreement. However, in support of Hypothesis 2, we did find evidence of a curvilinear relation. In particular, traits that were either highly desirable or highly undesirable elicited much lower agreement than the relatively neutral traits in the middle of the desirability continuum. This curvilinear effect is reflected in a negative correlation between evaluativeness and agreement. As predicted, the negative correlation between evaluativeness and agreement across the 76 traits was stronger for self-peer ($r = -.53$) than for peer-peer agreement ($r = -.35$), as shown by the paired-samples t test for the difference between correlations, $t(73) = 3.5, p < .01$.

Figure 3 illustrates the general curvilinear relation between social desirability and agreement, as well as the moderator effect of type of judge on this relation. The figure shows regression lines estimated from four separate regression equations, in which agreement on the 76 traits was predicted from the desirability values of the traits; these regressions

Table 1
Correlations of Peer-Peer and Self-Peer Agreement Indices with Observability, Desirability, Evaluativeness, and Each of the Big Five Domains Computed across Traits

	Study 1		Study 2	
	Peer-peer agreement	Self-peer agreement	Peer-peer agreement	Self-peer agreement
Trait property				
Observability	.36**	.38**	.37**	.50**
Desirability	−.04	.09	—	—
Evaluativeness	−.35**	−.53**	−.29*	−.53**
Big Five domain[a]				
Extraversion	.37**	.32**	.17	.18*
Agreeableness	−.34**	−.39**	−.19*	−.27*
Conscientiousness	−.17	−.14	−.06	−.16
Emotional Stability	.08	.22*	−.07	.13
Intellect/Openness[b]	.04	−.02	.09	.01
Multiple *R*	.48**	.51**	.29*	.44*
Self-peer agreement	.63**	—	.65**	—

Note. In Study 1, correlations were computed across 76 traits. In Study 2, a reanalysis of McCrae and Costa's (1987) data, correlations were computed across 80 bipolar trait scales, except those with observability and evaluativeness, which were computed across 40 bipolar scales.

a. Values are point-biserial correlations between agreement and each of the Big Five domains. Positive correlations indicate higher agreement for traits from that Big Five domain.

b. In Study 1, the fifth factor was represented primarily by traits related to Intellect; in Study 2, the fifth factor was represented primarily by traits related to Openness to Experience.

$*p < .05$

$**p < .01$.

were computed separately for the undesirable and the desirable traits, and for self-peer and peer-peer agreement.

For the undesirable traits on the left side of Figure 3, self-peer agreement had a strong positive correlation with desirability ($r = .70$): Agreement *increases* as traits become more neutral and less undesirable. Conversely, for the desirable traits on the right side of the figure, self-peer agreement had a strong negative relation with desirability ($r = -.41$): Agreement *decreases* as traits become more desirable and less neutral. As shown in the figure, we found a similar curvilinear

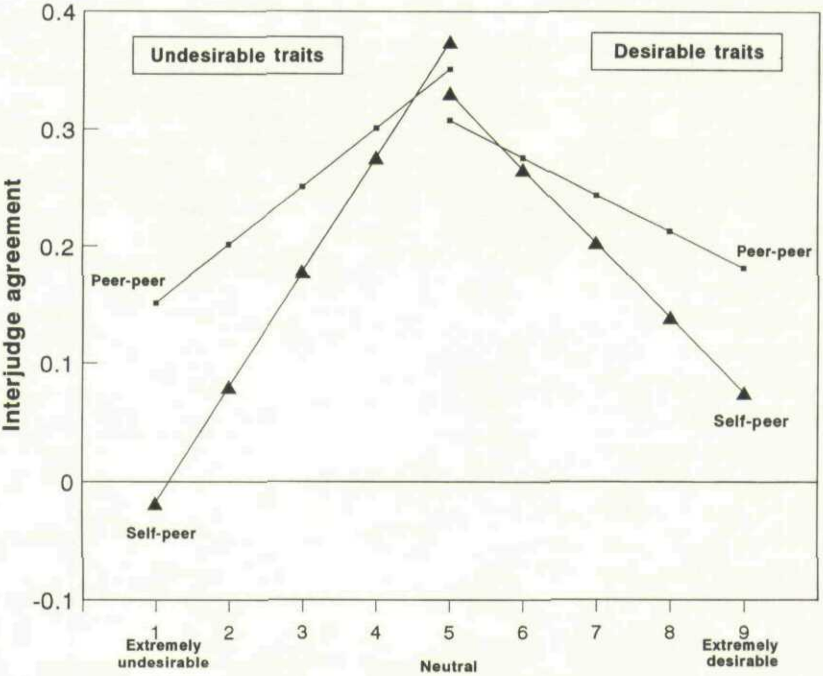

Figure 3
Interjudge Agreement as a Function of Social Desirability

Note. The relation of desirability to peer-peer and self-peer agreement is shown by regression lines computed separately for the undesirable traits (left of the neutral midpoint of the desirability continuum) and the desirable traits (right of the neutral midpoint).

effect for peer-peer agreement (r = .52 for the undesirable traits, and r = −.25 for the desirable traits), but this effect was less pronounced than for self-peer agreement. Finally, as the regression lines in Figure 3 suggest, peer-peer and self-peer agreement did not differ for relatively neutral traits. To further illustrate this point, we examined agreement for the 10 most neutral traits (i.e., those with desirability values between 4 and 6) and found that peer-peer and self-peer agreement were the same, both averaging .34. Thus, Hypothesis 1 holds for traits with evaluative implications but does not hold for neutral traits.

Finally, Table 1 also includes the point-biserial correlations between agreement and each of the Big Five domains across the 76 traits, providing an alternative way to represent the mean differences among the Big Five domains summarized in Figure 1. Positive correlations with a

Big Five domain indicate higher agreement for traits from that domain. As shown in Table 1, traits from the Extraversion domain elicited more agreement than traits from the other domains. Conversely, Agreeableness traits elicited less agreement. The overall effect size of content domain, expressed as the multiple correlation between agreement and all Big Five domains together, was .48 for peer-peer agreement and .51 for self-peer agreement.[4]

Thus, our findings replicate previous research linking the Big Five content domains to interjudge agreement. We also found that two general properties of personality traits—observability and evaluativeness —predicted agreement. How are these two sets of findings connected? Do the Big Five domains differ from each other in agreement because the traits defining the five domains differ in observability and evaluativeness? In other words, can the agreement differences among the Big Five be explained in terms of the two more general trait properties? To address these questions, we consider first how the Big Five differ in observability and evaluativeness, and then how these three predictors jointly influence agreement.

Big Five Differences in Observability, Desirability, and Evaluativeness

In Figure 4, we present the observability ratings as a function of Big Five content domain; the traits in the Extraversion domain ($M = 6.1$) were the most observable, and the traits from the Intellect domain ($M = 4.8$) were the least observable. A two-way ANOVA on the mean observability ratings, with Big Five domain and pole (high vs. low) as factors, showed a main effect of Big Five domain, $F(4, 66) = 5.3$, $p < .001$, no main effect of pole, and no interaction.

Next we examined social desirability and evaluativeness as a function of Big Five content domain. Figure 5 shows the mean desirability values for traits representing the high and low poles of each of the Big Five domains. When the two poles were combined, the Big Five domains were all close to the neutral line (i.e., a desirability of 5) and differed little in desirability. However, there were dramatic differences between the poles; for all five domains, the high pole (e.g., Extraversion) was

[4] In this multiple regression analysis, the Big Five content domains are represented by four independent dummy variables entered as a block; the fifth dummy variable would be redundant.

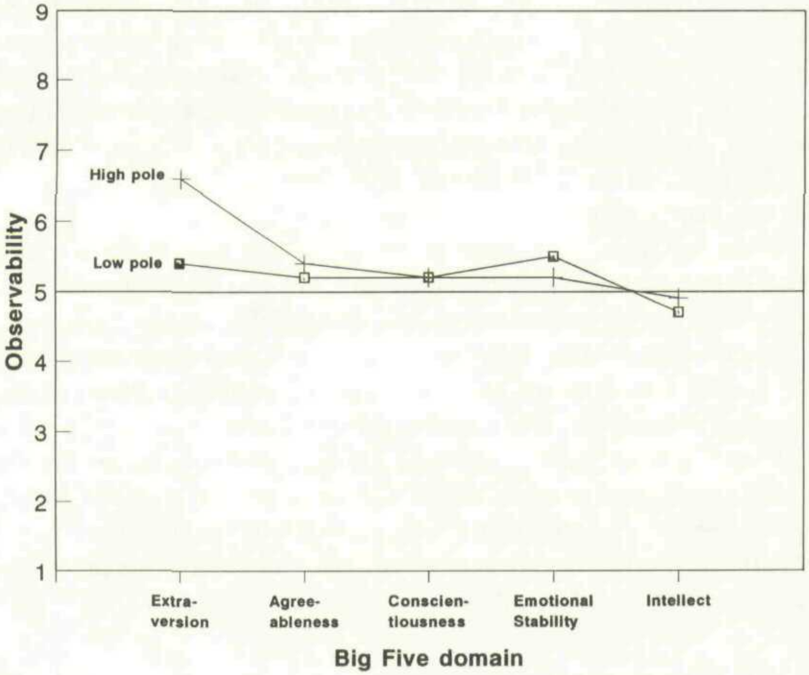

Figure 4
Observability as a Function of Big Five Content Domain: Mean Observability Values for Traits Representing the High and Low Poles of Each of the Big Five

substantially more desirable than the low pole (e.g., Introversion). The magnitude of this difference varied considerably across domains, and the desirability values for the low poles were essentially a mirror image of the values for the high poles, resulting in a fish-shaped figure with Extraversion at the mouth of the fish and Intellect at the tail.

These effects were confirmed by a two-way ANOVA, with Big Five domain and pole as factors. We found no main effect of Big Five domain, but the main effect of pole, $F(1, 66) = 230.0$, $p < .001$, and the interaction, $F(4, 66) = 3.7$, $p < .01$, were significant. The interaction effect is of particular importance because it reflects the differences among the Big Five domains in evaluativeness. In Figure 5, evaluativeness is indicated by the distance between the mean desirability values for the traits representing the high and the low pole of each domain. Extraversion was the least evaluative domain (i.e., the shortest

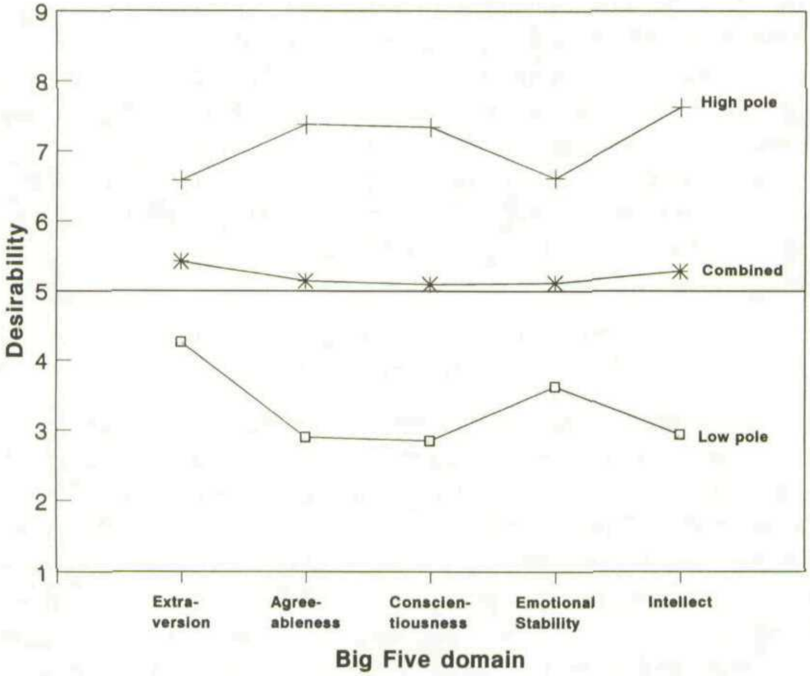

Figure 5

Social Desirability as a Function of Big Five Content Domain: Mean Desirability Values for Traits Representing the High and Low Poles of Each of the Big Five

Note. Evaluativeness is indicated by the distance between the mean desirability values for the high and the low pole of each domain.

distance between the high and low poles), followed by Emotional Stability, Agreeableness and Conscientiousness, and Intellect, which was most evaluatively polarized.[5]

In summary, Extraversion was both the most observable and the least

[5] Our findings on observability and desirability are relevant to the debate about which pole should be used to name the "fourth" Big Five factor (see McCrae & John, 1992). Researchers in the lexical tradition have used Emotional Stability as the factor label because, as Figure 5 shows, this pole is more socially desirable. On the other hand, researchers in the questionnaire tradition prefer the label Neuroticism (or Negative Emotionality) because it represents the more observable and thus salient pole of this trait domain. Similarly, Goldberg (1992) noted that the English language includes many more trait adjectives referring to the neurotic than to the emotionally stable pole.

evaluative Big Five domain. Given that observability was positively related to agreement and negatively related to evaluativeness, the positive relation of agreement to Extraversion might be attributable to the high observability and low evaluativeness of the traits in this domain. Agreeableness, on the other hand, did not differ from the other trait domains on either of these two trait properties. Thus, as we show below, the negative relation between Agreeableness and agreement cannot be explained by observability and evaluativeness.

Determinants of Interjudge Agreement: An Integration

Figure 6 brings together the different determinants of interjudge agreement examined in the present article. To test their independent effects in a joint analysis, we conducted a multiple regression using five predictors: (*a*) Big Five content domain (represented by four independent dummy variables entered as a block), (*b*) observability, (*c*) evaluativeness, (*d*) type of judge (represented by a dummy variable coded −1 for self-peer and +1 for peer-peer agreement), and (*e*) an interaction term representing the moderator effect of judge type on the relation between evaluativeness and agreement. Following Aiken and West's (1991) recommendations, we (*a*) standardized the criterion and each of the first four predictors, (*b*) computed the interaction term as the product of the standardized variables (i.e., Judge Type × Evaluativeness), and (*c*) interpreted the beta weights from the unstandardized (Friedrich) solution (see Aiken & West, 1991, pp. 42–44). This procedure makes the interaction term independent of the predictors from which it was formed, thus reducing multicollinearity problems and permitting direct interpretation of both the higher order and the lower order regression coefficients in Figure 6.

As shown in Figure 6, all five predictors had significant independent effects, and together they accounted for a substantial portion of the variance in interjudge agreement (multiple $R = .69$). The finding that the Big Five content domains as well as observability and evaluativeness had independent effects suggests that some of the Big Five differences in agreement cannot be reduced to differences in these two more general trait properties. More specifically, although Extraversion did not have an effect on agreement when observability and evaluativeness were taken into account, Agreeableness still had a significant effect. That is, the higher agreement for Extraversion traits can be attributed to their

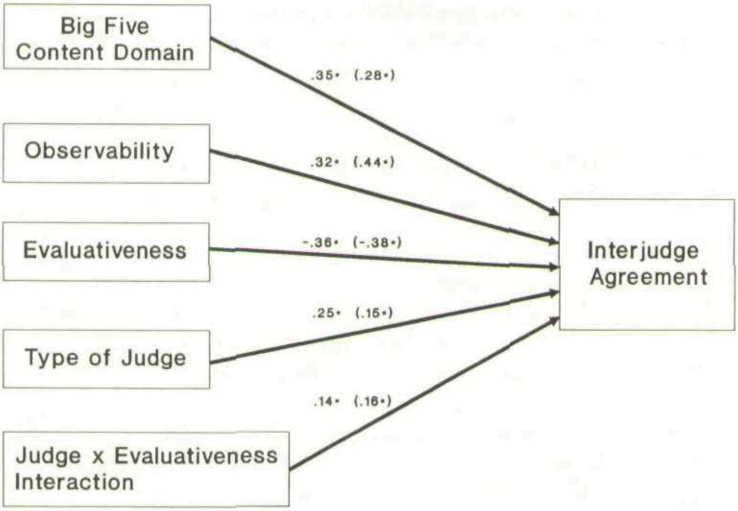

Figure 6
Determinants of Interjudge Agreement

Note. All values are beta weights from a multiple regression with all five predictors. The betas from Study 1 are given first, and the betas from Study 2 (our reanalysis of McCrae and Costa's [1987] data) are given in parentheses.

high observability and their low evaluativeness, but the lower agreement for Agreeableness traits could not be explained by these two trait properties.

Overall, evaluativeness was the strongest predictor, indicating that when people make personality judgments they have more difficulty agreeing with others about evaluative traits than about neutral traits (see Figure 3). The significant interaction between evaluativeness and type of judge further indicates that this effect is even more pronounced when people make judgments about themselves. These results confirm our earlier finding that the evaluativeness effect is stronger for self-peer than for peer-peer agreement, and they suggest that self- and peer judges differ more in their responses to evaluative traits than in their responses to neutral traits. Finally, in addition to the interaction between judge type and evaluativeness, the effect of judge type was also significant, supporting our earlier observation that agreement between two peers generally exceeded agreement between the self and a peer.

Although the results of Study 1 were theoretically meaningful and statistically significant, they need to be replicated in a different sample of subjects and traits. We tested the generalizability of our findings using interjudge agreement data collected by McCrae and Costa (1987) on the participants in the Augmented Baltimore Longitudinal Study of Aging.

Study 2: A Replication Using Data from McCrae and Costa (1987)

METHOD

For 218 subjects ranging in age from 30 to 90 (with roughly equal numbers in each decade), McCrae and Costa (1987) obtained self-ratings and ratings from two peers ($n = 72$), three peers ($n = 85$), or four peers ($n = 61$). The peers were extremely well-acquainted with the target subjects, having known them for an average of 18 years in a wide variety of situations and life contexts. These data provide us with the opportunity to assess the replicability of our findings in a sample of subjects who are much older than our college students, and a sample of peers who were much more familiar with the target subjects than in Study 1.

Personality ratings were made on 80 bipolar trait scales (e.g., talkative-quiet). These trait scales included 40 Big Five marker scales from Goldberg (1983), plus an additional 40 scales designed by McCrae and Costa (1987) to measure the Big Five domains. For each of these 80 bipolar scales, Robert R. McCrae provided us with overall indices of self-peer and peer-peer agreement, reflecting dyadic agreement correlations between all possible combinations of judges.[6]

To index Big Five content domain, we used each trait scale's absolute factor loading on each of the five peer rating factors reported in Table 3 of McCrae and Costa's (1987) original report. The 80 unipolar adjectives comprising Goldberg's 40 bipolar scales had been included in the observability and evaluativeness ratings obtained for our first study.[7] For each bipolar scale, the observability value was the rating of the more observable of the two unipolar traits comprising the scale. For example, for the bipolar scale talkative-quiet, we used the observability rating of talkative (7.7) as the observability value rather than the rating of quiet (6.6). For evaluativeness, we used the squared distance between the desirability values of the two unipolar traits comprising the bipolar scale. For example, talkative-quiet was low in evaluativeness

[6] The agreement correlations from McCrae and Costa (1987) are intraclass correlations, whereas we used Pearson correlations among randomly assigned judges. The similarity of the findings across the two studies suggests that the procedure used to compute agreement made little difference.

[7] Seventy-two of our 76 unipolar traits corresponded to 36 of the 80 bipolar scales used by McCrae and Costa (1987). Our additional four traits (reckless, cautious, impulsive, controlled) were all from the Conscientiousness domain.

(.04) because the desirability values of talkative (5.3) and quiet (5.5) were very close.

RESULTS AND DISCUSSION

As in Study 1, we used interjudge agreement correlations as the data. Again the unit of analysis was the trait, with the 80 bipolar trait scales serving as observations. Table 1 presents the correlations of the agreement indices with observability, evaluativeness, and each of the Big Five domains, computed across the bipolar trait scales. The correlations on the right-hand side of Table 1 show that both the observability and the evaluativeness effects on peer-peer and self-peer agreement were replicated.[8] Moreover, the evaluativeness effect was again stronger for self-peer ($r = -.53$) than for peer-peer agreement ($r = -.29$), as shown by a paired-samples test for the difference between correlations, $t(37) = 3.5, p < .01$.

With respect to the Big Five domains, we also found a similar pattern of correlations across the two studies; traits from the Agreeableness domain again elicited significantly less interjudge agreement than traits from the other domains, and traits from the Extraversion domain elicited somewhat higher levels of agreement. Overall, the multiple correlation of the Big Five domains was .44 with self-peer agreement and .29 with peer-peer agreement. This close replication of our Big Five effects with a set of traits representing McCrae and Costa's (1987) variant of the Big Five domains shows that our findings are not specific to a particular selection of traits. For example, the fifth factor was not related to agreement, whether it was measured by traits related to Intellect or to Openness.

We also replicated the difference between the two types of agreement by comparing peer-peer and self-peer agreement for the 80 bipolar trait scales.[9] Averaged across all 80 scales, peer-peer agreement was

[8] In Study 2, all p values for replicated effects were based on one-tailed significance tests.

[9] The difference between peer-peer and self-peer agreement was somewhat smaller in McCrae and Costa's (1987) data than in our own data (Study 1), probably because of the much greater length of acquaintanceship among their subjects. Because the peers were long-time friends of the subjects, they probably shared more information with them (cf. Kenny, 1991) and were also more emotionally involved, making their judgments more similar to self-judgments. Both factors would tend to reduce the difference between self-peer and peer-peer agreement. Importantly, however, this difference remained significant and none of our other findings changed appreciably (see Figure 6).

.22 (*SD* = .08), whereas self-peer agreement averaged .20 (*SD* = .08). Although this difference is small in magnitude (one-quarter of a standard deviation), it was statistically significant as shown by a paired-samples *t* test across the 80 trait scales, $t(79) = 2.7, p < .01$. Moreover, this difference held for both the original 40 Goldberg scales, $t(39) = 2.1, p < .05$, and for the 40 scales McCrae and Costa added, $t(39) = 1.8, p < .05$.

Finally, we tested our integrative model of the determinants of interjudge agreement. As in Study 1, we conducted a regression analysis with all five predictors entered simultaneously, using the Friedrich solution described in Aiken and West (1991). The regression weights from this analysis are given in parentheses in Figure 6. Again, all five determinants contributed independently to the prediction of agreement, and the multiple correlation was .67, similar to the .69 value obtained in Study 1. Moreover, both the direction and the magnitude of the regression weights were replicated.

To test the generalizability of the model more formally, we conducted a double cross-validation analysis. The multiple correlation was .65 when we applied the regression equation obtained in McCrae and Costa's sample to our sample in Study 1; conversely, when we applied the equation from our sample to theirs, the multiple correlation was .62. These results reveal relatively little shrinkage in the cross-validation samples and thus provide impressive evidence for the generalizability of the model presented in Figure 6.

GENERAL DISCUSSION

We have examined several determinants of interjudge agreement on personality traits in two independent samples. The findings were similar across the two studies and provide a clear replication. Our model held whether the subjects were college students or adults, whether the peers had known the subjects for 1 year or for 20 years, whether the trait ratings were made on unipolar adjectives or bipolar scales, and whether the Big Five representation was derived from the lexical tradition (see John, 1990) or from the questionnaire tradition (see McCrae & Costa, 1987). In both data sets, we were able to explain almost half of the total variance in agreement, and probably most of the reliable variance. Our findings suggest that agreement on personality trait adjectives is largely a function of four factors: which Big Five content domain the trait represents, how observable relevant behaviors are, how evaluative the trait is, and whether the self is one of the judges.

Both our data set and McCrae and Costa's (1987) used trait adjectives selected to be representative of the Big Five. How do these findings compare with those from studies using other stimulus sets and procedures? With regard to content domain, our finding that Extraversion traits elicited the most agreement and Agreeableness traits the least agreement is consistent with the previous literature. Our regression analyses further showed that the Extraversion effect can be explained by the high observability and low evaluativeness of the traits comprising that domain. The Agreeableness effect, in contrast, could not be explained by these two trait properties; future research needs to clarify why traits from the Agreeableness domain elicit less interjudge agreement.

Several earlier studies (Cheek, 1982; Funder & Dobroth, 1987; Norman & Goldberg, 1966; Paulhus & Bruce, 1992) have also found relatively low agreement for Emotional Stability (vs. Neuroticism), a finding we did not obtain in the present studies. This difference may be due to the observability and evaluativeness of the items used to represent Emotional Stability; in our research, the Emotional Stability traits were not particularly unobservable or evaluative, which may account in part for the relatively high agreement we found for this domain.

With regard to observability, our findings closely replicate Funder and Dobroth's (1987) results obtained with Q-sort items. Similarly, our evaluativeness effect is consistent with Park and Judd's (1989) finding that agreement on a factor consisting primarily of neutral traits (their Factor 1) was higher than on a factor consisting primarily of highly evaluative traits (their Factor 2).

In addition to the three trait properties of content domain, observability, and evaluativeness, our findings suggest that interjudge agreement also depends on whether the self serves as one of the judges; specifically, agreement between two peers generally exceeded agreement between the self and a peer. To further examine this difference, we reanalyzed interjudge agreement correlations from Funder and Dobroth (1987), which were provided by David C. Funder. Using the 100 Q-sort items as the observations, mean peer-peer agreement (.26) was significantly higher than mean self-peer agreement (.22), even though their self-peer agreement index used the average of two peers and may thus have boosted its size relative to dyadic-level agreement. The Q-sort data provided a less clear-cut replication of the evaluativeness effect. As in our research, evaluative items elicited less agreement than neutral items, but this effect was limited to negatively evaluative (i.e., socially undesirable) Q-sort items. This difference may arise because the Q-sort items are considerably longer and more complex than single trait adjectives and may therefore have less clear-cut evaluative implications. In fact, the distribution of desirability values is bimodal for English trait adjectives (i.e., most traits are either clearly desirable or clearly undesirable) (see Goldberg, 1982), whereas the distribution is unimodal for the 100 Q-sort items. In summary, our findings seem to be consistent with previous research, although differences between trait adjectives and more complex personality items may provide a boundary condition on some of our effects.

Implications for Personality Assessment

The view that self-perceptions correspond with perceptions by others has served as a theoretical basis for the use of self-reports as data in psychological research (e.g., McCrae, 1982). Our model of the determinants of interjudge agreement suggests several conditions under which such correspondence can be expected to be relatively high and relatively low. In particular, researchers constructing personality measures would be well-advised to avoid highly evaluative and unobservable items because such items may reduce interjudge reliability and convergent validity between self and peers.

In interpreting the overall level of interjudge agreement in the present research, it is important to note that we examined agreement between self and a single peer for individual items. Nonetheless, even without any aggregation, the self-reports generally converged with the peer reports, providing evidence for the validity of both types of judgments. Note that self-peer agreement on the Big Five dimensions was much more substantial when the judgments were aggregated across individual trait adjectives and across peers, averaging .43 in Study 1 and .48 in McCrae and Costa's (1987, Table 6) data.

However, convergent validity seldom approaches the boundaries imposed by reliability, suggesting that method-specific factors influence personality ratings. The nature of method-specific variance is not yet well understood, and psychological analyses of method effects are badly needed (Ozer, 1989). The present findings suggest that self-ratings of evaluative traits contain more method-specific variance than self-ratings of neutral traits. This is an important finding because most personality traits are at least somewhat evaluative. More generally, we hope that our findings will help elucidate the psychological mechanisms and processes underlying method variance in self-reports.

Implications for the Processes Involved in Self- and Other Perception

Perhaps our most intriguing findings involve the differences between self- and peer judgments. If, as some theorists have suggested (e.g., Bem, 1972; Mead, 1934), self-perception and peer perception follow similar processes, then the determinants and the level of self-peer and peer-peer agreement should not differ. Indeed, we found that two determinants of interjudge agreement—observability and Big Five content domain—had the same effect on both self-peer and peer-peer agreement. However, we also found an important difference: self-peer agreement was lower than peer-peer agreement when the trait being judged was evaluative. These findings lead us to two speculations. Determinants of interjudge agreement that implicate cognitive-informational processes (e.g., observability and content domain of the trait being judged) may have the same effect on self-peer and peer-peer agreement. On the other hand, determinants of interjudge agreement that implicate motivational factors such as self-enhancement needs (e.g., evaluativeness) may differentially affect self-peer and peer-peer agreement.

More generally, our findings are consistent with the idea that differences between self- and peer perception may stem, in part, from a differential response to the evaluativeness of the attribute judged. Thus, self- and peer perception may indeed proceed through similar processes for judgments that are not ego-involving, but the process is altered and self-perceptions may become distorted when the trait is affectively charged. This interpretation points to the importance of motivational biases activated by ego involvement. Ego involvement may trigger affective and defensive processes that influence our self-perceptions to a greater extent than our perceptions of most others. When evaluating others, we typically do not experience threats to self-worth.[10] However, when evaluating ourselves on extremely evaluative traits such as lazy, honest, or stupid, many of us experience a threat to self-worth and engage in the "selective industry of the mind" and possibly distort reality in the service of our "totalitarian ego."

[10] Exceptions may occur when the judges are emotionally invested in their perceptions of the other person. For example, parents' perceptions of their child's personality may involve much the same psychological processes as their self-perceptions. Thus, the factors underlying differences between self- and other judgments (e.g., level of emotional involvement) are perhaps more critical than the simple distinction between self and other.

This difference in the process of self- and other perception will lead to lower self-peer agreement on evaluative traits only when individuals differ in their responsiveness to ego-involving stimuli. That is, our findings imply that evaluative traits elicit self-enhancement biases for some individuals but not for others. Previous research supports the notion that individual differences in self-enhancement biases are systematic and psychologically meaningful (e.g., Lockard & Paulhus, 1988). One possible personality variable that may account for our finding that self-peer agreement is lower on evaluative traits is the construct of narcissism (John & Robins, in press). When judging themselves on evaluative traits, narcissistic individuals may experience a threat to their self-worth and bolster their self-image by perceiving themselves more positively than they are seen by others, whereas this should not be true for relatively modest, nonnarcissistic individuals.

The motivational explanation we have proposed for self-judgments of evaluative traits implies that under conditions of ego involvement self-judgments may be less accurate than the judgments of a well-informed other. Although agreement and accuracy are related, agreement does not ensure accuracy (Funder, 1987; Kenny, 1991). Thus, although we have shown that self-judgments agree less with peer judgments when the trait is evaluative, the present research cannot provide conclusive evidence that the self is biased. Such evidence would require research comparing self-judgments to an accuracy criterion that can be justified on logical or empirical grounds.

In general, our findings contribute to the burgeoning literature on the conditions under which human observers agree with one another. Consistent with the view that self-judgments are influenced by motivational factors, we have delineated a condition under which self-judgments differ from judgments by others. The effects of other factors on self-other agreement, such as prior information and the unique visual perspective of the self, remain to be explored and provide important avenues for future research.

REFERENCES

Allport, G. W. (1958). What units shall we employ? In G. Lindzey (Ed.), *Assessment of human motives* (pp. 238–260). New York: Reinhart.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

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

Bem, D. J. (1972). Self-perception theory. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 6, pp. 1–62). New York: Academic Press.

Bem, D. J., & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review*, **81**, 506–520.

Block, J. (1978). *The Q-sort method in personality assessment and psychiatric research*. Palo Alto, CA: Consulting Psychologists Press.

Chaplin, W. F., & Goldberg, L. R. (1984). A failure to replicate Bem and Allen's study of individual differences in cross-situational consistency. *Journal of Personality and Social Psychology*, **47**, 1074–1090.

Cheek, J. M. (1982). Aggregation, moderator variables, and the validity of personality tests: A peer-rating study. *Journal of Personality and Social Psychology*, **43**, 1254–1269.

Colvin, C. R. (1993). "Judgable" people: Personality, behavior, and competing explanations. *Journal of Personality and Social Psychology*, **64**, 861–873.

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

Craik, K. H. (1986). Personality research methods: An historical perspective. *Journal of Personality*, **54**, 18–51.

Funder, D. C. (1980). On seeing ourselves as others see us: Self-other agreement and discrepancy in personality ratings. *Journal of Personality and Social Psychology*, **48**, 473–493.

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

Funder, D. C., & Colvin, C. R. (1988). Friends and strangers: Acquaintanceship, agreement, and the accuracy of personality judgment. *Journal of Personality and Social Psychology*, **55**, 149–158.

Funder, D. C., & Dobroth, K. M. (1987). Differences between traits: Properties associated with interjudge agreement. *Journal of Personality and Social Psychology*, **52**, 409–418.

Goldberg, L. R. (1982). From Ace to Zombie: Some explorations in the language of personality. In C. D. Spielberger & J. N. Butcher (Eds.), *Advances in personality assessment* (Vol. 1, pp. 203–234). Hillsdale, NJ: Lawrence Erlbaum.

Goldberg, L. R. (1983, June). *The magical number five, plus or minus two: Some considerations on the dimensionality of personality descriptors*. Paper presented at the Gerontology Research Center, NIA/NIH, Baltimore.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, **4**, 26–42.

Greenwald, A. G. (1980). The totalitarian ego: Fabrication and revision of personal history. *American Psychologist*, **35**, 603–618.

Hampson, S. E., Goldberg, L. R., & John, O. P. (1987). Category-breadth and social-desirability values for 573 personality terms. *European Journal of Personality*, **1**, 241–258.

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

James, W. (1890). *Principles of psychology*. New York: Holt.

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.

John, O. P., & Robins, R. W. (in press). Accuracy and bias in self-perception: Individual differences in self-enhancement and the role of narcissism. *Journal of Personality and Social Psychology*, **66**.

Jones, E. E., & Nisbett, R. E. (1971). *The actor and the observer: Divergent perceptions of the causes of behavior*. Morristown, NJ: General Learning Press.

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

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

Lockard J. S., & Paulhus, D. L. (1988). *Self-deception: An adaptive mechanism*. Englewood-Cliffs, NJ: Prentice-Hall.

McCrae, R. R. (1982). Consensual validation of personality traits: Evidence from self-reports and ratings. *Journal of Personality and Social Psychology*, **43**, 293–303.

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

McCrae, R. R., & Costa, P. T., Jr. (1989). Different points of view: Self-reports and ratings in the assessment of personality. In J. P. Forgas & J. M. Innes (Eds.), *Recent advances in social psychology: An international perspective* (pp. 429–439). North-Holland, The Netherlands: Elsevier Science.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, **60**, 175–215.

Mead, G. H. (1934). *Mind, self, and society*. Chicago: University of Chicago Press.

Miller, D. T. (1976). Ego involvement and attributions for success and failure. *Journal of Personality and Social Psychology*, **34**, 901–906.

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

Ozer, D. (1989). Construct validity in personality assessment. In D. M. Buss & N. Cantor (Eds.), *Personality psychology: Recent trends and emerging issues* (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*, **63**, 816–824.

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

Petersen, P. G. (1965). *Reliability of judgments of personality as a function of subjects* *and traits being judged*. Unpublished doctoral dissertation, University of California, Berkeley.

Robins, R. W., & John, O. P. (1993, August). *Factors underlying accuracy and bias in self-perception: The role of visual perspective*. Poster presented at the 101st annual convention of the American Psychological Association, Toronto.

Rogers, C. R. (1959). A theory of therapy, personality, and interpersonal relations, developed in the client-centered framework. In S. Koch (Ed.), *Psychology: A study of a science* (Vol. 3, pp. 185–256). New York: McGraw-Hill.

Shraugher, J. S., & Schoeneman, T. J. (1979). Symbolic interactionist view of self-concept: Through the looking glass darkly. *Psychological Bulletin*, **86**, 549–573.

Storms, M. D. (1973). Videotape and the attribution process: Reversing actors' and observers' points of view. *Journal of Personality and Social Psychology*, **27**, 165–175.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, **103**, 193–210.

Tesser, A. (1988). Toward a self-evaluation maintenance model of social behavior. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 21, pp. 181–227). New York: Academic Press.

Thorne, A. (1989). Conditional patterns, transference, and the coherence of personality across time. In D. M. Buss & N. Cantor (Eds.), *Personality psychology: Recent trends and emerging directions* (pp. 149–159). New York: Springer-Verlag.

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.

Manuscript received October 21, 1992; revised March 26, 1993.