Identifying Causes of Disagreement Between Self-Reports and Spouse Ratings of Personality

Robert R. McCrae
Stephanie V. Stone
Gerontology Research Center, National Institute on Aging, National Institutes of Health
Peter J. Fagan
Johns Hopkins Medical Institutions
Paul T. Costa Jr.
Gerontology Research Center, National Institute on Aging, National Institutes of Health

ABSTRACT Self-reports and spouse ratings of personality traits typically show less-than-perfect agreement, but powerful moderators of agreement have not yet been identified. In Study 1, 47 married couples completed the Revised NEO Personality Inventory to describe themselves and their spouses. Extent of

Portions of this article were presented at the European Congress of Psychology, held in July 1995, at Athens, Greece, and at the American Psychological Association Convention, held in August 1996, in Toronto, Canada. We wish to thank Melissa Kittner-Triolo and Jeffrey H. Herbst for assistance in coordinating interviews and data management, and Robin Majeski for coding reasons for disagreement.

Correspondence concerning this article should be addressed to Robert R. McCrae, Personality, Stress and Coping Section, Gerontology Research Center, 4940 Eastern Avenue, Baltimore, Maryland 21224. Electronic mail may be sent via the Internet to JEFFM@MVX.GRC.NIA.NIH.GOV

Journal of Personality 66:3, June 1998.

agreement was not consistently moderated by response sets; the age, intelligence, or education of the respondent; or the length or quality of the relationship. In Study 2 these couples were interviewed about reasons for substantial disagreements, and an audiotape was content-analyzed. Sixteen reasons were reliably coded, including idiosyncratic understanding of items, reference to different time frames or roles, and unavailability of covert experience to the spouse. Faking good, assumed similarity, and other variables prominent in the psychometric literature were relatively unimportant. Findings (1) suggest that attempts to improve the validity of self-reports and ratings may need to be refocused and (2) underscore the desirability of routinely obtaining multiple sources of information on personality.

Not surprisingly, familiarity with a target generally increases the accuracy of observer ratings of personality (Funder & Colvin, 1988; Norman & Goldberg, 1966; Paulhus & Reynolds, 1995). Spouses are presumably the observers most intimately acquainted with their mates, and studies show substantial self/spouse agreement, with correlations typically ranging from .4 to .6 (McCrae, 1982; Mutén, 1991). Although somewhat higher than the usual level of agreement between self and single peer raters (e.g., Funder, Kolar, & Blackman, 1995), these correlations are far from unity even when corrected for unreliability. In some individuals and for some traits, the discrepancies may be very marked.

There are two major research traditions relevant to an understanding of these differences. Research on response sets and styles (Jackson & Messick, 1961; Schinka, Kinder, & Kremer, 1997) focuses on distortions introduced by the individual's approach to completing a questionnaire. Respondents may be careless, or misunderstand directions, or endorse only desirable statements, or agree indiscriminately with any item. Even if couples understood each other perfectly, response sets might lower observed agreement.

A second approach—social cognition (Fiske, 1993; Kenny, 1994)—focuses on the processes by which people come to understand each other, and suggests that agreement may be limited by imperfect social perception. External observers do not have access to covert information about private thoughts and feelings. Different observers may use different standards of comparison. Liking or disliking may bias perceptions of personality.

In principle, researchers should be able to control or compensate for these sources of disagreement and thus increase self/spouse correlations indefinitely. In practice, however, this has proven to be extremely

difficult. Great ingenuity and effort have gone into the design of validity scales to detect and correct response biases, but with limited success (Dicken, 1963; Alperin, Archer, & Coates, 1996). Indeed, a number of studies have shown that corrections for defensiveness or socially desirable responding sometimes have the effect of reducing rather than enhancing validity (see Barrick & Mount, 1996; McCrae & Costa, 1983; McCrae et al., 1989).

Similarly, attempts to moderate self/other agreement through variables thought to influence social perception have shown mixed results. In the short term, as information about the other is being acquired, some variables have been shown to operate as hypothesized. For example, Ickes, Stinson, Bissonnette, and Garcia (1990) reported that grade point average, presumably an index of intelligence, was related to accurate perceptions of a stranger's thoughts and feelings, and Paulhus and Reynolds (1995) showed that consensus increased with familiarity. But in the long term, when a stable perception of the other has been formed, the usefulness of moderators in explaining agreement or disagreement is more questionable. Funder (1995) noted that the search for characteristics that would identify good judges of personality "proved by any standard disappointing" (p. 654) as long ago as the 1950s (but see Vogt & Colvin, 1996, for some recent progress). Much the same can be said for relational variables. For example, in a study of peer ratings in a large sample of adults, McCrae (1994) found that neither length of acquaintance, nor liking for target, nor frequency of social interactions, nor perceived similarity, nor any of 28 other variables that characterized the rater or the relationship was consistently related to peer/self agreement.

Two possibilities remain: either the appropriate validity indices and moderator variables have not yet been identified, or reasons for disagreement are intrinsically unpredictable. Kenny (1994) notes that *uniqueness*—an effect in personality perception that can be attributed not to the perceiver or to the target, but to their relationship—is consistently strong, and concludes that "other-perception is quite idiosyncratic" (p. 204). Similarly, Dunning and McElwee (1995) point to idiosyncratic understandings of trait concepts as a source of invalidity in self-reports.

The present research was designed to explore causes of disagreements between self-reports and spouse ratings of personality traits. Study 1 examines two sets of hypothesized moderator variables, one related to response sets, the other to aspects of social perception. Study 2 adopts a qualitative approach: Participant couples were invited to consider and

comment on the most marked discrepancies between their self-reports and their spouses' ratings of them. A content analysis of their explanations was used to identify possible reasons for disagreement.

Agreement and Disagreement on Personality Traits

Consider an actual case. Figure 1 presents self-report (solid line) and spouse rating (broken line) profiles on the Revised NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992a). At the left of the figure are scores on the five major dimensions of personality: Neuroticism (N), Extraversion (E), Openness to Experience (O), Agreeableness (A), and Conscientiousness (C). Toward the right are five groups of specific traits, or facets, that are primary definers of these factors. Within-sex *T*-scores based on self-report and observer rating norms are plotted, and show that this man is average in N and E, and high in O. He considers himself high in A and low in C, whereas his wife places him in the average range on both factors. Turning to the facets, there is considerable scatter within the factors, but by and large the two profiles are similar. The largest disagreement appears to be on C4: Achievement Striving, in which he places himself in the very low range whereas his wife regards him as high.

There are several indices that might be used to quantify the extent of agreement between these two profiles. A simple Pearson correlation across profile elements is sometimes used, but that index is sensitive only to the shapes of the two profiles. Better indices take into account the actual distance between the scores, usually using as a measure of dissimilarity some version of D^2, the sum of the squared differences between corresponding profile elements (Cronbach & Gleser, 1953). Cattell's (1949) r_p transforms D^2 into a coefficient with a range from –1 to +1.

Indices based solely on D^2 make the assumption that the distance between two profile elements is equally important at all levels of the trait. McCrae (1993) argued, however, that disagreements are less serious when they occur in the extreme ranges of the trait than when they occur near the middle. *T*-scores of 60 and 80 both indicate high levels of a trait, even though they are two standard deviations apart. By contrast, *T*-scores of 40 and 60 yield the same D^2, but they have qualitatively different interpretations—low versus high standing on the trait. Similarly, close

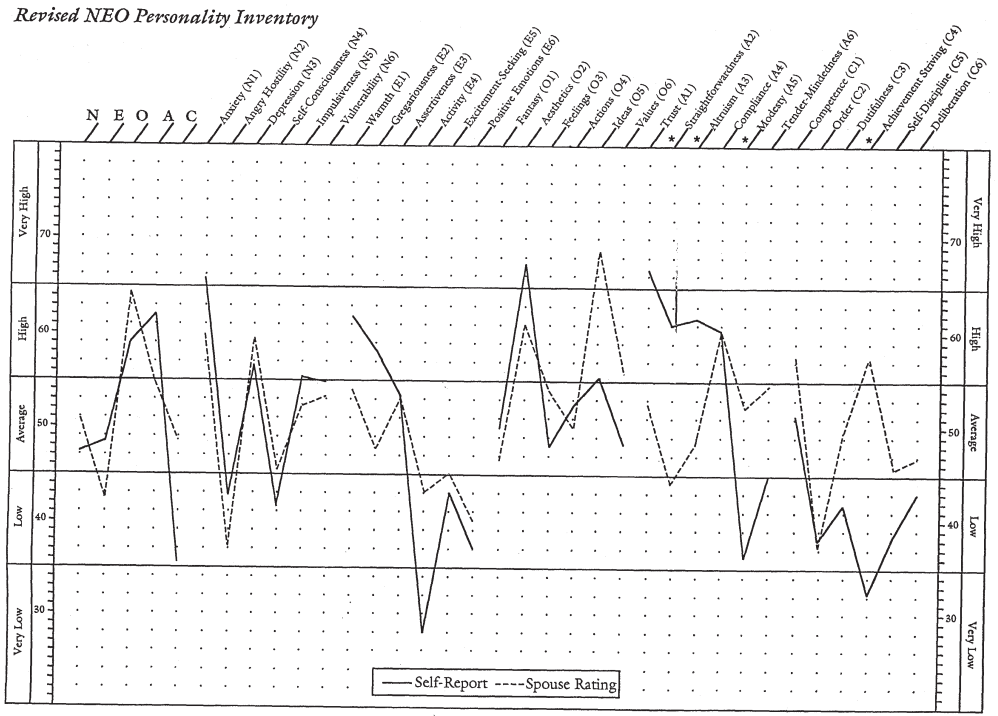

Figure 1

Revised NEO Personality Inventory profile for a typical case. Solid lines represent self-reports; broken lines represent spouse ratings. Profile form reproduced by special permission of the publisher from Revised NEO Personality Inventory. Copyright © 1978, 1985, 1989, 1992 by Psychological Assessment Resources, Inc. (PAR). Further reproduction is prohibited without permission of PAR.

agreement between two scores might easily happen by chance if both are near average, where most scores are distributed; but it would be very unusual for two very high or very low scores to agree by chance.

McCrae (1993) therefore proposed an Index of Profile Agreement (I_{pa}) that takes into account both the distance between profile elements and the extremeness of their mean. When both sets of ratings are expressed as *z*-scores, I_{pa} is defined as

$$\frac{k + 2\sum M^2 - \sum d^2}{\sqrt{10k}},$$

where k is the number of traits in the profile, M is the mean of the two ratings for the trait, and d is the difference between the two ratings. I_{pa} is highest when the profile is extreme and the two ratings are in close agreement; it is lowest when the difference between ratings is great and centered near average (e.g., *T*-scores of 30 and 70). One of the peculiarities of the index is that even perfect agreement does not result in high values of I_{pa} if all the scores in the profile are near average. However, empirical comparison showed that a coefficient based on I_{pa} was superior to r_p in distinguishing matched from mismatched pairs of profiles (McCrae, 1993).

For analyses of the NEO-PI-R, a total I_{pa} score is calculated from the five factor scores; the profile in Figure 1 has the median total I_{pa} value in the present sample, and thus represents a typical degree of agreement. By combining information from all five factors, total I_{pa} should be sensitive to variables that affect agreement in general. For example, McCrae (1993) showed that mean total I_{pa} was higher for spouse raters than for peer raters, presumably because spouses know their mates better than neighbors and coworkers do.

Individual I_{pa} values can also be calculated for each factor and facet. The latter are of particular use in identifying specific areas of substantial disagreement. For the purposes of this article, negative I_{pa} values are taken to indicate a significant difference between the self-report and spouse rating that is worth investigating. All negative I_{pa} values correspond to differences of at least one standard deviation. In Figure 1, significant disagreements are seen for A2: Straightforwardness, A3: Altruism, A5: Modesty, and C4: Achievement Striving. This article seeks to account for such disagreements.

Study 1: Nomothetic Moderators

Study 1 was intended to examine possible moderators of agreement in personality descriptions among married couples. The first set of moderators are measures of response sets that can be calculated in either self-reports or observer ratings. Acquiescence and extreme responding are easily computed from item response data. Because they believed that the use of social desirability and inconsistency scales was not strongly supported by the research literature, the authors of the NEO-PI-R did not include such scales as part of the inventory (Costa & McCrae, 1997). However, Schinka, Kinder, and Kremer (1997) created scales from NEO-PI-R items to measure positive and negative impression management and inconsistent responding. By the usual logic of validity scales, it can be hypothesized that high scores on any of these response style measures should be inversely related to accuracy, to agreement, and thus to I_{pa} scores.

We also examined three variables relevant to social perception that characterize individual participants: age, years of education, and general intelligence as estimated from vocabulary test scores. It might be hypothesized that maturity, formal education, and general intelligence would all facilitate insight into one's own personality and a more accurate understanding of others' (cf. Ickes et al., 1990).

Finally, we examined three variables that characterize the couple: years married, marital adjustment, and similarity in personality. If long acquaintance leads to better understanding, spouse ratings should be more accurate among long-married couples. The overall quality of the relationship might also be expected to show some relation to concordance on views of personality. Intimate communication, which should lead to better information about the spouse's thoughts, feelings, and values, is known to be related to marital adjustment (Dean & Carlson, 1984). In addition, serious problems in the relationship might motivate distortions in each partner's view of the other.

Kenny (1994) reviewed data on assumed similarity and concluded that people do indeed tend to rate others as they rate themselves, especially with respect to Agreeableness: nice people think others are nice, too. If self and spouse are in fact similar, an assumed similarity bias will enhance accuracy. In this study we operationalized similarity as I_{pa} between self-report of husband and self-report of wife, calculated across all five factors.

We hypothesized that all six of these social perception–related variables would be positively related to agreement.

METHOD

Participants

Questionnaires were completed by 94 married participants (47 couples) in the National Institute on Aging's Baltimore Longitudinal Study of Aging (BLSA; Shock et al., 1984). BLSA participants in general are healthy, well-educated, community-dwelling volunteers who visit the Gerontology Research Center every 2 years for biomedical and psychosocial testing. The subsample who consented to participate in the current study consisted of 47 men aged 28 to 85 (M = 63.5) and 47 women aged 26 to 82 (M = 61.6). All but one were high school graduates, and 64 (68%) held college or advanced degrees. Both spouses were in their first marriage in 83% of the cases, and couples had been married from 1 to 59 years (M = 35.5).

Measures

Personality was assessed through self-reports (Form S) and spouse ratings (Form R) on the Revised NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992a), a 240-item questionnaire designed to operationalize the five-factor model of personality (Digman, 1990). The NEO-PI-R measures 30 specific traits, or facets, that define the five factors; factor scores are calculated from scoring weights given in the manual (Costa & McCrae, 1992a, p. 8). Data on the reliability, validity, and longitudinal stability of the scales is detailed elsewhere (Costa & McCrae, 1992a). In the present sample, principal components analysis with varimax rotation of the 94 self-reports clearly replicated the structure of the normative sample, with factor congruence coefficients ranging from .93 to .98. Analysis of the 94 spouse ratings clearly replicated four of the factors (congruence coefficients = .92 to .97); the Openness factor was somewhat less clearly recovered (congruence coefficient = .83).

T-scores based on separate norms for Form S and Form R were calculated for all participants. In addition, an adjusted mean *T*-score was calculated to represent the best single estimate of personality scores (see Appendix A for the rationale and procedure for the adjustment).

In addition to the substantive scales, responses to the NEO-PI-R were scored for acquiescent responding (the number of "agree" and "strongly agree" responses) and extreme responding (the number of "strongly disagree" and "strongly agree" responses). These two response styles are reasonably reliable: the correlation between acquiescent responding in completing Form S and

Form R was .64; the cross-form correlation for extreme responding was also .64. Finally, three response style measures proposed by Schinka et al. (1997) were also scored. Positive Presentation Management and Negative Presentation Management scales consist of 10 items each that would likely be endorsed by someone wishing to make a positive or a negative impression. The Inconsistency scale consists of 10 pairs of items with similar content; the sum of differences in responses to these items is used as an index of random or careless responding. All five response style measures were scored both for self-reports and spouse ratings, because invalidity in either source could reduce agreement across methods.

Quality of the marital relationship was assessed by a slightly modified version of the Dyadic Adjustment Scale (Spanier, 1976) with subscales for satisfaction, cohesion (e.g., working together on a project), consensus (e.g., agreement on finances, goals), and affectional expression. Although these couples in general had stable long-term relationships, a fairly wide range of marital adjustment was represented in the sample, with descriptions of the relationship ranging from "fairly unhappy" to "perfect." Coefficient alpha for the total scale in the present sample was .94, and total adjustment was significantly related to low N, high C, and especially high A, *r*s = –.25, .22, and .41, respectively.

Data on years of education and scores on the Vocabulary subtest of the Revised Wechsler Adult Intelligence Test (Matarazzo, 1972) were available from the archives of the BLSA.

Procedure

Couples were mailed the Dyadic Adjustment Scale and the self-report form of the NEO-PI-R, along with an informed consent form. If both members of the couple agreed to join the study, they were subsequently sent the observer rating form of the NEO-PI-R to rate their spouse. Participants were asked to work independently, but were aware that their responses would be compared to those of their spouses.

RESULTS AND DISCUSSION

On average, men scored in the low range in self-reported C5: Self-Discipline ($T = 43.5$); otherwise, all mean NEO-PI-R facet scores in both self-reports and ratings were in the average range. The first column of Table 1 shows correlations between self-reports and spouse ratings for the five factors and 30 NEO-PI-R facets. The values are comparable to those seen in other normal (Costa & McCrae, 1992a) and clinical (Mutén, 1991) samples: they show significant and often substantial

Table 1
Simple and Disattenuated Correlations Between Self-Reports and Spouse Ratings of Personality, and Frequency of Disagreements on Facet Scales

NEO-PI-R	r	r/r_{tt}	Disagreements
Factors			
N: Neuroticism	.46	.55	
E: Extraversion	.74	.81	
O: Openness	.53	.60	
A: Agreeableness	.58	.67	
C: Conscientiousness	.49	.56	
Facets			
N1: Anxiety	.57	.74	13
N2: Angry Hostility	.57	.79	16
N3: Depression	.57	.79	15
N4: Self-Consciousness	.24	.33	31
N5: Impulsiveness	.48	.68	15
N6: Vulnerability	.37	.47	21
E1: Warmth	.59	.75	9
E2: Gregariousness	.58	.72	7
E3: Assertiveness	.52	.61	14
E4: Activity	.46	.55	20
E5: Excitement Seeking	.52	.63	17
E6: Positive Emotions	.52	.64	13
O1: Fantasy	.64	.85	12
O2: Aesthetics	.49	.57	20
O3: Feelings	.32	.43	20
O4: Actions	.49	.64	17
O5: Ideas	.48	.58	21
O6: Values	.51	.65	15
A1: Trust	.56	.73	15
A2: Straightforwardness	.39	.58	21
A3: Altruism	.33	.45	24
A4: Compliance	.55	.70	19
A5: Modesty	.45	.58	18
A6: Tender-Mindedness	.33	.43	21
C1: Competence	.33	.45	24
C2: Order	.66	.83	10
C3: Dutifulness	.36	.56	29
C4: Achievement Striving	.38	.50	21
C5: Self-Discipline	.51	.62	18
C6: Deliberation	.37	.47	22

Note: $N = 94$.

agreement, but are low enough to suggest meaningful differences in some instances.

In part, disagreement might be due simply to error of measurement. McCrae, Yik, Trapnell, Bond, and Paulhus (in press) reported 2-year reliability coefficients for NEO-PI-R domains and facets that are used in the second column of Table 1 to disattenuate the cross-observer correlations. The resulting correlations are higher, but still leave discrepancies to be explained.

The third column of Table 1 reports, for each facet, the number of instances in which there was significant disagreement by the negative I_{pa} criterion. These 538 disagreements represent 19.1% of the total, or about six facets per profile.

To test the hypotheses that response style measures would be negatively related, and personal attributes and characteristics of the relationship positively related to the extent of agreement between self-reports and ratings, these variables were correlated with indices of profile agreement. Table 2 reports correlations of I_{pa} for the total five-factor profile and for each of the factors separately with agreement moderator variables. Most of the correlations are quite small, with a median absolute magnitude of .09; more importantly, they do not show any consistent trend in support of the hypotheses. Of the 60 correlations between response set variables and I_{pa}s, exactly half (30) are in the hypothesized negative direction. Only a third (12) of the correlations with social perception variables are in the hypothesized direction. The one noteworthy correlation among these is the positive relation between couple similarity and agreement on the Agreeableness factor (r = .22, 95% CI = .02–.46)—a finding consistent with the assumed similarity hypothesis (Kenny, 1994).

Because the I_{pa} measure is unfamiliar, supplementary analyses were undertaken. In the first of these, the correlations in Table 2 were recalculated using D^2 in place of I_{pa}. Because D^2 is a measure of dissimilarity, the direction of all hypotheses was reversed in these analyses. As in Table 2, most of the correlations were very small (median absolute magnitude again was .09), although 67 of the 96 correlations were at least in the right direction. This trend in the data is consistent with the view that very small moderator effects (corresponding to rs of about .10) are operating as hypothesized. However, in this analysis the correlation of D^2 on the Agreeableness factor with

couple similarity was –.03, providing little support for the assumed similarity hypothesis.

The second set of supplementary analyses were moderated regression analyses (Chaplin, 1991). These analyses are commonly recommended to examine the effects of one variable on the association between two other variables. The approach cannot be employed on overall profile agreement, but can be used to examine agreement on each factor individually. We therefore conducted two sets of moderated regression analyses. In the first, spouse ratings on the five factors were used as criteria and self-reports on the corresponding factors were used as predictors. Age, years of education, WAIS-R Vocabulary

Table 2
Correlations of Indices of Profile Agreement with Hypothesized Moderator Variables

	Index of Profile Agreement					
Moderator Variable	Total	N	E	O	A	C
Acquiescence						
Self	–.03	.09	.02	–.09	.02	–.08
Spouse Rating	–.11	–.05	.06	–.05	.06	–.20
Extreme Responding						
Self	.13	.23	.02	.13	.15	–.11
Spouse Rating	.06	–.16	.07	–.06	.25	.03
Positive Presentation						
Self	–.25	–.09	–.16	–.20	.03	–.17
Spouse Rating	–.15	–.03	–.03	–.18	.11	–.20
Negative Presentation						
Self	.16	–.02	.06	.03	–.11	.32
Spouse Rating	.05	.09	.05	–.07	–.11	.12
Inconsistency						
Self	–.01	.20	.08	–.09	–.18	–.03
Spouse Rating	.13	–.04	.00	–.09	.00	.32
Age	–.18	.04	–.05	–.16	–.12	–.12
Education	–.13	–.05	.01	–.12	–.08	–.09
Vocabulary	.18	–.01	.13	.05	.14	.13
Years Married	–.16	.05	–.01	–.09	–.16	–.15
Marital Adjustment	–.01	–.05	–.02	.07	.17	–.13
Similarity	–.08	–.07	–.17	.00	.22	–.12

Note: $N = 94$.

scores, years married, marital adjustment, personality similarity, and the five response style measures from the self-report NEO-PI-R were examined as moderators. In the second set of analyses, self-reports were used as criteria and spouse ratings as predictors, and the five response style measures from the spouse rating NEO-PI-R were potential moderators.

Of these 80 analyses, only two showed moderator effects large enough to attain conventional levels of significance.[1] Both concerned extreme responding as a moderator of the prediction of Neuroticism, and they pointed in opposite directions, suggesting that extreme responding in self-reports enhances agreement, whereas extreme responding in spouse ratings reduces it. This puzzling finding is perhaps best regarded as a chance occurrence.

Neither the correlations with I_{pa} in Table 2 nor the moderated regressions provide strong support for the value of any of the 16 proposed moderators of self/spouse agreement. These data are consistent with earlier reports that failed to find moderating effects of characteristics of the relationship between self and rater (McCrae, 1994), and that questioned the utility of validity scales (Costa & McCrae, 1997). Degree of acquaintance is an important variable in the first few days or weeks (Paulhus & Reynolds, 1995), but all married couples are likely to have reached a ceiling on this effect. Similarly, it is doubtless true that some degree of intelligence is required to complete a personality questionnaire accurately, but variations in intelligence among high school graduates is apparently too restricted to affect accuracy.

It is, of course, possible that some of these variables would have a more pronounced effect in other circumstances. Impression management, for example, might operate more forcefully in situations (such as forensic or I/O settings) in which the individual has a clear motivation to distort responses. Acquiescence might be an important moderator of agreement on scales that—unlike the NEO-PI-R—have unbalanced keying. Marital dynamics might have proved important in a sample that included clearly dysfunctional couples.

1. With an *N* of 94, an alpha of .05, and an R^2 of .25 for the multiple regression as a whole (corresponding to the typical self/spouse correlation of .50), the power to detect a small moderator effect (corresponding to an *r* of .10) is only about .20 (see Cohen & Cohen, 1975). Even so, this means that we would expect 16 significant effects in 80 analyses if all the hypotheses were correct (and if the samples were independent). Power to detect a medium-sized moderator effect (corresponding to an *r* of .30) is greater than .90, so it is unlikely that major effects were missed.

In the present study, however, participants lacked any special incentive to distort responses—in fact, they knew that their answers would be compared to an external standard, a circumstance that is supposed to enhance accuracy (Roese & Jamieson, 1993). The inventory used was designed to reduce effects of acquiescence and other response biases. It is therefore all the more notable that, as Table 1 shows, agreement is far from perfect. What can account for these disagreements? In Study 2 we pointed out discrepancies to couples and asked them to offer explanations.

Study 2: Content Analyses

For both philosophical and scientific reasons, qualitative methods have become more prominent in the social sciences in recent years (see, e.g., Silverman, 1993). Philosophically, these approaches value the subjective experience of the participant, seeing scientific inquiry as a kind of collaboration between researcher and informant. Methodologically, qualitative methods can provide a fertile context of discovery, suggesting hypotheses that may not have appeared in the professional literature. Provided that they are conducted with reasonable regard for methodological rigor (Stiles, 1993), qualitative approaches can be a useful supplement to strictly quantitative, hypothesis-testing methods.

The present instance provided a research problem for which a qualitative approach seemed well suited. Years of hypothesis testing had made little progress in the search for moderators of self/other agreement, and we had at hand as potential collaborators participants in the BLSA, a group of well-educated and highly motivated volunteers. We decided to ask them why disagreements occurred.

METHOD

Participants

The couples from whom data were gathered in Study 1 were invited to discuss their questionnaire results with a clinical psychologist. Because of scheduling difficulties and unfinished interviews, no coded reasons for disagreement were available for 7 of the men and 9 of the women. However, *t*-tests showed that these 16 individuals did not differ from those who were interviewed in terms of personality factors, I_{pa} measures, or any of the moderator variables listed in Table 2.

Interview Procedure

In preparation for the interviews, several documents were generated for each spouse from the self-report and spouse-rating NEO-PI-R data obtained in Study 1. First, for each domain and facet, we listed Form S and Form R *T*-scores, their adjusted means, and I_{pa} (McCrae, 1993), indicating significant differences where they occurred. Graphed personality profiles (see Figure 1) were also prepared. Second, for each facet on which there was a significant difference between Forms S and R, we listed the 8 items from that facet, along with the actual responses each spouse made to that item (ranging from "strongly disagree" to "strongly agree"). Third, for each domain, the NEO-PI-R computer Interpretive Report (Costa & McCrae, 1992a) was used to generate a description of the target spouse based on the adjusted means of self-report and rating scores.

Couples were interviewed together by a clinical psychologist with experience in couples therapy (PJF) during their regularly scheduled BLSA visit. Interviews were recorded on audiotape. The interviews (a) provided feedback on the results of the personality questionnaires, using the profile sheet and Interpretive Report, so that participants could knowledgeably discuss the facets they disagreed on; and (b) sought explanations from the participants for disagreements. The focus of the discussion alternated between spouses across all five personality domains.

An attempt was made to discuss every significant difference at the facet level. The interviewer first read the description of the facet from the NEO-PI-R manual. Couples were then encouraged to discuss their views on this trait and on individual items from the facet on which their responses differed. If possible, spouses were asked to resolve their differences, that is, to agree on the correct description of the individual's personality.

The interviewer allowed the spouses considerable latitude in discussing their differences: he clarified the meaning of the facets, ensured that the basic sequence of the interview was maintained, and offered interpretation when the discussion reached an impasse. Otherwise, there was minimal intervention by the interviewer. Most interviews lasted from 45 to 90 minutes.

Based on this discussion and his own observations, the interviewer provided a clinical judgment of the true score of the individual on each disputed facet. The absolute difference between these clinical judgment scores and the self-report, spouse rating, and adjusted mean NEO-PI-R scores gives a sense of the relative accuracy of the latter three scores. In 38% of the cases, the interviewer's impression was closest to the self-report; in 26% of the cases it was closest to the spouse rating; in the remaining cases, the adjusted mean best approximated the clinical judgment. This roughly even division suggests that reasons for disagreement should be sought in both the self-reports and the spouse ratings.

Content Coding

A content analysis of the audiotaped interviews was used to examine reasons for disagreement. A list of possible reasons for disagreement was initially derived from a review of the literature. One of the authors (SVS) then listened to five interviews and coded presence or absence of these reasons. Experience with these five interviews led to revisions in and elaborations of the list of reasons and the development of a code book that included the theoretical and operational definition for each of 28 reasons.

Coders were asked to consider the full context of the discussion and to code a reason as applicable if it was explicitly mentioned by the couple or if it could be reasonably inferred by the judge (as in the case of unconscious denial). Coders considered both the general discussion of the facet and any comments made about specific items, and every instance of a reason was coded; thus, multiple reasons for disagreement on one facet might be coded, and the same reason might be coded more than once if used to explain disagreements on different items. Coders also made a summary judgment of the primary reason for disagreement on the facet. Excerpts from one of the interviews, together with reasons coded, are given as examples in Appendix B.

The five taped interviews used in developing the coding system were also used to train a coder, a psychology graduate student with experience in content coding. Thereafter, SVS and the coder worked independently. The coder completed all the remaining interviews, yielding a total of 409 facets on which there were disagreements; SVS completed an additional 14 interviews to provide an estimate of interrater reliability. Analyses of reliability were undertaken at the level of facets. In the 14 interviews, disagreements on 129 different facets were discussed.

RESULTS AND DISCUSSION

Intraclass correlations between the two coders across these 129 facets are given in Table 3. Nine of the reasons suggested by the literature or by the first 5 interviews were never coded by one or both coders in the next 14 interviews, and thus interrater reliability could not be assessed. These included acquiescent responding, unconscious denial, and faking bad. Of the remaining 19 reasons, interrater agreement on 3 was so low that it might be attributed to chance. The other 16 reasons showed interrater reliabilities we judged to be acceptable, ranging from .19 to .82 (M = .57). It should be recalled that this was a difficult coding task, with 28 distinct reasons to consider and an open-ended number of codes that might be assigned to each facet.

Table 3
Coding System and Interrater Reliability for Reasons for Disagreement

Category/Reason	r^a
Clerical or reading error in self-report	.57
Clerical or reading error in spouse rating	.82
Extreme responding in self-report	.49
Extreme responding in spouse rating	—
Acquiescence vs. naysaying in self-report	—
Acquiescence vs. naysaying in spouse rating	—
Modesty (strict standards) in self-report	.68
Immodesty (lenient standards) in self-report	–.02
Unconscious denial in self-report	—
Conscious faking good in self-report	.33
Conscious faking bad in self-report	—
Leniency in spouse rating	.17
Animosity in spouse rating	.19
Assumed similarity of self to spouse in self-report	—
Perceived contrast of self with spouse in self-report	.62
Assumed similarity of spouse to self in spouse rating	.56
Perceived contrast of spouse with self in spouse rating	.76
Reporting spouse's view instead of own in self-report	—
Reporting spouse's view instead of own in spouse rating	—
Insufficient data for spouse rating	.54
Data available but not used in spouse rating	—
Spouse unaware of covert feelings, attitudes	.66
Different specific behaviors considered	.39
Different time frames adopted	.72
Different roles considered	.66
Different interpretation of words or items	.49
Different standards or reference groups used	.34
Unidentifiable source of difference	.15

Note: Dashes indicate that one or both coders never used the code. [a]Intraclass correlation between coders across 129 facets.

Experimental studies of an individual moderator variable (e.g., Chaplin & Panter, 1993) can document its effect, but they do not give a clear sense of its relative importance in personality assessment. The present study is chiefly intended to identify which sources of

disagreement are in fact most common. The frequency with which each of the 16 reliable reasons for disagreement was coded is given in Table 4, along with the coder's summary judgment on the primary reason for disagreement on a facet. The two indicators show a very similar pattern.

Differences in interpretation of the meaning of items is by far the most common reason for disagreements. This reason was coded when the discrepancy appeared to result from different interpretations of test items rather than from differences of opinion about the personality of the target. Different interpretations arose when spouses attended to different parts of the item, or understood qualifiers like *rarely* or *often* differently, or construed the meaning of words differently. This reason for disagreement and inaccuracy in personality assessments was noted as long ago as the

Table 4
Frequency of Reasons for Disagreements

Reason	Times Coded	Times Judged Primary Reason
Different interpretation of words or items	269	100
Different specific behaviors considered	155	38
Spouse unaware of covert feelings, attitudes	92	46
Perceived contrast of spouse with self in spouse rating	73	24
Different standards or reference groups used	66	20
Insufficient data for spouse rating	63	22
Clerical or reading error in self-report	59	17
Clerical or reading error in spouse rating	49	13
Different roles considered	47	19
Different time frames adopted	45	13
Modesty (strict standards) in self-report	28	18
Perceived contrast of self with spouse in self-report	21	3
Conscious faking good in self-report	17	8
Animosity in spouse rating	12	3
Extreme responding in self-report	6	1
Assumed similarity of spouse to self in spouse rating	3	0
Total	1,005[a]	345[b]

[a]Total is greater than 409 (the number of facet disagreements discussed) because multiple reasons could be coded for each facet. [b]Total is less than 409 because some reasons judged primary were from unreliable categories not reported here.

1930s (Benton, 1935; Landis & Katz, 1934),[2] but has been given relatively little attention since.

Different interpretation of items was followed in frequency by a diverse set of other reasons, including differences in the specific behaviors, time frames, or roles considered when responding to items; insufficient information or insight into covert feelings and attitudes; and simple clerical errors. Faking good, extreme responding, and assumed similarity—response artifacts that have commanded enormous attention in the literature—were rarely considered responsible for disagreements.

Idiosyncratic interpretation of items and clerical errors would seem to add only random error to ratings, but some of the other reasons listed in Table 4 (such as unawareness of covert feelings, different standards used, and perceived contrast with others) might be enduring features of the judge's understanding and rating of the target. They might, in part, account for the presence of stable method variance, helping to explain why cross-observer correlations are usually smaller than within-observer stability coefficients, even after periods of many years (Costa & McCrae, 1992b).

In an effort to understand the source of these different kinds of disagreement, we examined correlations across the 409 facets between the 16 reasons for disagreement and person characteristics, including age, gender, years married, dyadic adjustment, couple similarity, years of education, vocabulary, and adjusted mean personality factor scores. Of these 192 correlations, only 2 exceeded .20 in absolute magnitude: more agreeable people were more likely to have made clerical or reading errors (or perhaps were more likely to admit such errors), and more introverted people were more likely to show discrepancies between overt behavior and covert feelings and attitudes—in this latter case it appears their need for privacy led them to be misunderstood. In general, however, it does not appear that characteristics of the person are strongly related to reasons for self/spouse disagreement.

Do different trait domains elicit different reasons for disagreement? To answer that question, we conducted analyses of variance on reasons for disagreement using trait domain as the classifying variable. Significant, and generally meaningful, effects were found for 6 of the 16 reliably coded reasons. Undue modesty in self-reports was most frequently

2. We wish to thank Eric S. Knowles for calling these and other relevant references to our attention.

associated with disagreements about the highly evaluative dimensions of C and A. Perceived contrast with spouse lead to inaccurate spouse ratings of E and O. Insufficient data for spouse ratings was most common in facets related to O, whereas consideration of different roles was seen chiefly on facets of E—perhaps reflecting differences in interpersonal behavior at home and at work. Different interpretations of words or items, the most common reason for disagreements, was particularly common in the E, O, and A domains. The largest effect of trait domain on reason concerned the contrast between overt behavior observed by spouses and covert feelings and attitudes accessible only through self-reports. Only 2% of the C facet disagreements elicited this explanation during the interview, whereas 22% of the O facet and 38% of the N facet disagreements did. Presumably this is because both the inner processing of feelings and ideas and psychological distress are largely a matter of private experience.

General Discussion

The results of Study 1 are broadly consistent with the literature on self/other agreement: neither response style measures nor characteristics of the respondent or relationship were consistently related to the extent of agreement on any of the five factors or the total personality profile. It is possible that other variables, such as self-monitoring (Snyder, 1974) would work more successfully (Tunnell, 1980); research to date, however, is not particularly encouraging. Chaplin (1991) concluded that even when they are found, "moderator effect sizes in personality can be expected to correspond to a correlation of about .10" (p. 143). And Kenny noted that "other-perception is quite idiosyncratic, yet researchers in interpersonal perception have little understanding about why there is uniqueness"[3] (1994, p. 204).

Study 2 provides some hints that may help explain the source of disagreements. A first finding of note is that several of the explanations that figure most prominently in the research literature were rarely encountered in participants' accounts of disagreement. Acquiescence, unconscious denial, and conscious faking bad were never coded as reasons

3. It should be noted that Kenny's (1994) concept of uniqueness applies not to consistent biases in the rater, but to idiosyncracies specific to the rater's view of a particular target. In the present study, where only one target was examined, the two effects are confounded.

in the reliability subsample, and conscious faking good, extreme responding, and assumed similarity were found in only a handful of instances in the full sample. It is, of course, possible that the reasons offered by participants were not in fact the operative ones—they may have preferred to attribute differences to misunderstandings rather than to their own biases or ignorance—and that the clinical judgment of the coders was not sufficiently sensitive to detect the true reasons. But on their face, the findings in Study 2 suggest that much of the literature on invalidity and disagreement may have addressed the wrong issues.

Table 4 points to a different list of explanations that may be more fruitful to pursue. Some beginnings in these directions have already been made. Dunning and colleagues (Dunning & Cohen, 1992; Dunning & McElwee, 1995) have conducted a series of studies on idiosyncratic interpretations of trait terms. Their work documents the pervasiveness of this problem, which was clearly the most common reason for disagreements in the present study. They have also begun to show that these differences are not purely arbitrary, but reflect in part motivated perceptions of the nature of a trait term. A person who forcefully complains about poor food in a restaurant may construe this action as assertiveness, whereas her less assertive spouse may see it as an expression of hostility (see Appendix B).

Many of the disagreements in the present study appeared to result from the use of different information: different roles, different time frames, and different specific behaviors. People behave differently in different settings, and observers may perceive only a single side of the individual's personality—in Kenny's (1994) terms, self and spouse have imperfect *overlap* of information. Funder and colleagues (1995) showed that there is stronger cross-observer consensus when observers have seen the individual in the same context. One implication here is that the fullest portrait of the individual may be obtained by aggregating not simply multiple observers, but observers from multiple contexts: for example, a spouse, a friend, and a coworker. A second implication is that context-specific agreement might be enhanced by focusing the attention of the self-reporter on the particular context known to the rater. This strategy has a practical application: Schmit, Ryan, Stierwalt, and Powell (1995) have suggested that the validity of personality assessment for particular purposes (e.g., employment) may be increased by specifying the role or context (e.g., work) that is to be used in making self-descriptions.

Chaplin and Buckner (1988) have called attention to the fact that individuals use different standards as the basis of self-reports of personality: some apparently use an idiothetic, some an ipsative, and some a normative standard. When comparing themselves to others, those who implicitly adopt a normative standard may use as their reference group everyone they know, or people of their own age and sex, or the members of their family. Personality assessors who leave the standard of comparison unspecified introduce an unknown amount of error into their measures. Table 4 shows that using different standards or reference groups was a common reason for disagreements.

Implications for Assessment

Funder (1995) has offered a sophisticated and systematic classification of moderator variables that might account for disagreements. He pointed out that characteristics of the judge, trait, target, and available information may all affect trait judgments, and that interactions among these are also possible. For example, one person may have expertise in judging Openness to Experience, but be a poor judge of Agreeableness. The finding in Table 2 that couples with similar personality profiles showed higher agreement with respect to the Agreeableness factor can be considered a Judge × Target × Trait interaction.

Although a useful framework for research, Funder's model has somewhat limited applications in personality assessment. Clinicians and researchers cannot normally restrict their assessments to easily judged traits or targets, so the model is perhaps most useful in directing attention to characteristics of good judges (see Vogt & Colvin, 1996)—people whose skills in interpersonal perception and whose history of interaction with the target give them a good basis for making personality ratings.

Where self-reports are used, or where only a single rater (e.g., a spouse) is available, improving assessment depends on modifying the rating task. An examination of Table 4 provides some guidance here. Idiosyncratic interpretation of words is a major problem; the most obvious solution is to use clear and simple language, and to increase the number and diversity of items. Clerical or reading errors were relatively common even in this well-educated and cooperative group; increasing items would presumably attenuate such random errors (although at the possible expense of fatigue and carelessness). These are hardly novel suggestions; they have been part of the lore of test construction for

decades. Personality psychologists might also benefit from the recent work of survey researchers (e.g., Tanur, 1992) who have intensively studied the determinants of responses to single items.

It may be wise to attend more closely to instructions that focus the respondent on particular roles or parts of the lifespan. Such instructions might be intentionally broad ("Consider all your thoughts, feelings, and actions from your childhood on to the present") or emphasize a narrow aspect of the individual's total personality ("Consider only your thoughts, feelings, and actions when at work"). Similarly, instructions might specify the desired reference group: for example, all other adults, adults of the same sex, or other members of one's profession. Research on the effects of these variations in instructions would be welcome.

More fundamentally, the fact that there are idiosyncratic differences in personality perceptions underscores the need for multiple sources of information. Psychologists who reply on a single self-report (or a single observer rating) on a well-validated personality measure know that, in general, they have useful information. But sometimes the information will be wrong, and they have no way of knowing when. Validity scales have been devised to identify responses distorted by social desirability and defensiveness, but they have not worked well (see Alperin, Archer, & Coates, 1996; Ones, Viswesvaran, & Reiss, 1996), perhaps because they have addressed the wrong sources of distortion. No validity scale has been designed to reveal whether the respondent has attended to the full range of relevant behaviors, or used the appropriate reference group—yet these are more much more common reasons for discrepancies than faking good or extreme responding.

Obtaining two or more sources of information on a personality profile increases accuracy; the simple expedient of aggregating raters almost invariably increases validity of ratings (e.g., Kolar, Funder, & Colvin, 1996). In addition, substantial disagreements between informants call attention to particular traits that may be problematic. Consensus does not guarantee accuracy: it may instead reflect shared stereotypes or a folie à deux. But a substantial lack of consensus must mean that at least one rating is wrong, or that the two refer to different aspects of the person. Ideally, one would be able to gather additional information to adjudicate the disagreement, as the interviewer attempted to do in the present study; at a minimum, one should treat both estimates with appropriate caution.

APPENDIX A

Adjusting Means of Two Ratings

Personality scores are normally interpreted in comparison to normative data; separate norms are provided for Form S and Form R of the NEO-PI-R. Means and standard deviations in these norms are based on single raters. When scores from two sources are averaged, the appropriate norms would consist of data from averaged pairs of raters. Because such data are currently unavailable for the NEO-PI-R, an adjustment is used to estimate the appropriate *T*-score.

When two imperfectly correlated variables are averaged, the standard deviation of the average is less than that of the original variables, and must be adjusted if the same metric (e.g., *T*-score) is to be used. If *a* and *b* are expressed as *z*-scores, the variance of their sum is given by

$$\begin{aligned} \mathrm{VAR}(a + b) &= \mathrm{VAR}(a) + \mathrm{VAR}(b) + 2{*}\mathrm{COV}(ab), \\ &= 1 + 1 + 2{*}r_{ab}, \end{aligned}$$

and the variance of the mean is $(2 + 2{*}r_{ab})/4$ or $(1 + r_{ab})/2$. The standard deviation of the mean is $\sqrt{\left[\left(1 + r_{ab}\right) / 2\right]}$.

The correlation, r_{ab}, between self-reports and spouse ratings is typically about .50, so the standard deviation of their mean is estimated as $\sqrt{\left[(1 + .50) / 2\right]} = .866$. Adjusted means in this study are expressed as *z*-scores, divided by .866, and converted to *T*-scores. For example, if the *T*-score for the self-report is 56 and the *T*-score for the spouse rating is 64, the mean *T*-score is 60 but the adjusted mean *T*-score is 61.5. This adjustment is routinely used in computer interpretation of combined NEO-PI-R profiles.

APPENDIX B

Transcribed Excerpts: Disagreement on N2, Angry Hostility

Participant described herself as low on this facet whereas her husband rated her as high. There were substantial differences on responses to four of the eight items. In the following excerpts, they discuss these responses:

Item: "I often get angry at the way people treat me." (Her response: *Strongly disagree.* His response: *Neutral.*)

Husband: She sometimes gets upset with people she works with, and tells me about it.

Wife: I don't think I do. It takes a lot to even make me say anything back to somebody.

Interviewer: Feel it though?

Wife: I feel some anger at times, but not often.[a]

Item: "I'm an even-tempered person." (Her response: *Strongly agree.* His response: *Disagree.*)

Husband: I'm probably thinking more compared to me. She varies in it more than I do . . . less even-tempered.

Wife: That's probably true. But he holds a lot of his anger inside sometimes, and I let it go.[b, c]

Item: "I'm known as hot-blooded and quick-tempered." (Her response: *Strongly disagree.* His response: *Neutral.*)

Husband: Maybe some of the same reasoning as the previous one.

Interviewer: Examples from life?

Wife: I can't recall a great deal of anger from recent years, but I can recall one time I was very angry at him, but that was years ago. A lot of things have taken place since then, and I really don't see myself as hot-blooded.

Interviewer: Feeling anger.

Wife: Right.

Interviewer: You think she does?

Husband: For one thing, I'll probably end up more in the middle on everything. I don't go to extremes. I wasn't necessarily thinking of the two of us. I was thinking of others. Sometimes in a restaurant you express yourself to the waiter.

Wife: I don't see anything wrong with saying you're dissatisfied with the meal.

Interviewer: Feel angry?

Wife: No. That's not anger, it's taking care of yourself.
Husband: Maybe the difference's between assertive and aggressive, and I tend to see it more aggressive, and she sees it more assertive.[a]

Item: "It takes a lot to get me mad." (Her response: *Strongly agree*. His response: *Disagree*.)

Husband: Well, it seems to me that she gets upset over things that I don't think I would. I suppose that's what I'm looking at, the difference between myself and her. And, as she says, she expresses it much more than I would, so I hear it more.
Interviewer: Compared to general women, not yourself?
Husband: Maybe more than middle.
Wife: I think I see myself as less than other women.
Interviewer: Same situation, they experience more anger?
Wife: Uh huh.[b, d]

---

[a]Coded as: Different interpretation of words or items.
[b]Coded as: Rater's perceived contrast of spouse with self.
[c]Coded as: Self-reporter's perceived contrast of self with spouse.
[d]Coded as: Different standards or reference group.

REFERENCES

Alperin, J. J., Archer, R. P., & Coates, G. D. (1996). Development and effects of an MMPI-A *K*-correction procedure. *Journal of Personality Assessment,* **67,** 155–168.

Barrick, M. R., & Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. *Journal of Applied Psychology,* **81,** 261–272.

Benton, A. L. (1935). The interpretation of questionnaire items in a personality schedule. *Archives of Psychology* (Whole No. 190), 5–38.

Cattell, R. B. (1949). r_p and other coefficients of pattern similarity. *Psychometrica,* **14,** 279–298.

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

Chaplin, W. F., & Buckner, K. E. (1988). Self-ratings of personality: A naturalistic comparison of normative, ipsative, and idiothetic standards. *Journal of Personality,* **56,** 509–530.

Chaplin, W. F., & Panter, A. T. (1993). Shared meaning and the convergence among observers' personality descriptions. *Journal of Personality,* **61,** 553–585.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Costa, P. T. Jr., & McCrae, R. R. (1992a). *The Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T. Jr., & McCrae, R. R. (1992b). Trait psychology comes of age. In T. B. Sonderegger (Ed.), *Nebraska Symposium on Motivation: Psychology and aging* (pp. 169–204). Lincoln: University of Nebraska Press.

Costa, P. T. Jr., & McCrae, R. R. (1997). Stability and change in personality assessment: The Revised NEO Personality Inventory in the Year 2000. *Journal of Personality Assessment,* **68,** 86–94.

Dean, D. G., & Carlson, R. S. (1984). Definitions of life situation and marital adjustment. *Journal of Comparative Family Studies,* **15,** 441–448.

Dicken, C. (1963). Good impression, social desirability, and acquiescence as suppressor variables. *Educational and Psychological Measurement,* **23,** 699–720.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology,* **41,** 417–440.

Dunning, D., & Cohen, G. L. (1992). Egocentric definitions of traits and abilities in social judgment. *Journal of Personality and Social Psychology,* **63,** 341–355.

Dunning, D., & McElwee, R. O. (1995). Idiosyncratic trait definitions: Implications for self-descriptions and social judgment. *Journal of Personality and Social Psychology,* **68,** 936–946.

Fiske, S. T. (1993). Social cognition and social perception. *Annual Review of Psychology,* **44,** 155–194.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review,* **102,** 652–670.

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., Kolar, D. C., & Blackman, M. C. (1995). Agreement among judges of personality: Interpersonal relations, similarity, and acquaintanceship. *Journal of Personality and Social Psychology,* **69,** 656–672.

Ickes, W., Stinson, L., Bissonnette, V., & Garcia, S. (1990). Naturalistic social cognition: Empathic accuracy in mixed-sex dyads. *Journal of Personality and Social Psychology,* **59,** 730–742.

Jackson, D. N., & Messick, S. (1961). Acquiescence and desirability as response determinants on the MMPI. *Educational and Psychological Measurement,* **21,** 771–790.

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

Kolar, D. W., Funder, D. C., & Colvin, C. R. (1996). Comparing the accuracy of personality judgments by the self and knowledgeable others. *Journal of Personality,* **64,** 311–337.

Landis, C., & Katz, S. E. (1934). The validity of certain questions which purport to measure neurotic tendencies. *Journal of Applied Psychology,* **18,** 343–356.

Matarazzo, J. D. (1972). *Wechsler's measurements and appraisal of adult intelligence.* Baltimore: Williams and Wilkins.

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. (1993). Agreement of personality profiles across observers. *Multivariate Behavioral Research,* **28,** 13–28.

McCrae, R. R. (1994). The counterpoint of personality assessment: Self-reports and observer ratings. *Assessment,* **1,** 159–172.

McCrae, R. R., & Costa, P. T. Jr. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology,* **51,** 882–888.

McCrae, R. R., Costa, P. T. Jr., Dahlstrom, W. G., Barefoot, J. C., Siegler, I. C., & Williams, R. B. Jr. (1989). A caution on the use of the MMPI K-correction in research on psychosomatic medicine. *Psychosomatic Medicine,* **51,** 58–65.

McCrae, R. R., Yik, M.S.M., Trapnell, P. D., Bond, M. H., & Paulhus, D. L. (in press). Interpreting personality profiles across cultures: Bilingual, acculturation, and peer rating studies of Chinese undergraduates. *Journal of Personality and Social Psychology.*

Mutén, E. (1991). Self-reports, spouse ratings, and psychophysiological assessment in a behavioral medicine program: An application of the five-factor model. *Journal of Personality Assessment,* **57,** 449–464.

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

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology,* **81,** 660–679.

Paulhus, D. L., & Reynolds, S. (1995). Enhancing target variance in personality impressions: Highlighting the person in person perception. *Journal of Personality and Social Psychology,* **69,** 1233–1242.

Roese, N. J., & Jamieson, D. W. (1993). Twenty years of bogus pipeline research: A critical review and meta-analysis. *Psychological Bulletin,* **114,** 363–375.

Schinka, J., Kinder, B., & Kremer, T. (1997). Research validity scales for the NEO-PI-R: Development and initial validation. *Journal of Personality Assessment,* **68,** 127–138.

Schmit, M. J., Ryan, A. M., Stierwalt, S. L., & Powell, A. B. (1995). Frame-of-reference effects on personality scale scores and criterion-related validity. *Journal of Applied Psychology,* **80,** 607–620.

Shock, N. W., Greulich, R. C., Andres, R., Arenberg, D., Costa, P. T. Jr., Lakatta, E. G., & Tobin, J. D. (1984). *Normal human aging: The Baltimore Longitudinal Study of Aging* (NIH Publication No. 84-2450). Bethesda, MD: National Institutes of Health.

Silverman, D. (1993). *Interpreting qualitative data: Methods for analyzing talk, text, and interaction.* Thousand Oaks, CA: Sage.

Snyder, M. (1974). Self-monitoring of expressive behavior. *Journal of Personality and Social Psychology,* **30,** 526–537.

Spanier, G. B. (1976). Measuring dyadic adjustment: New scales for assessing the quality of marriage and similar dyads. *Journal of Marriage and the Family,* **38,** 15–28.

Stiles, W. B. (1993). Quality control in qualitative research. *Clinical Psychology Review,* **13,** 593–618.

Tanur, J. M. (Ed.). (1992). *Questions about questions: Inquiries into the cognitive basis of surveys.* New York: Russell Sage Foundation.

Tunnell, G. (1980). Intraindividual consistency in personality assessment: The effect of self-monitoring. *Journal of Personality,* **48,** 220–232.

Vogt, D. S., & Colvin, C. R. (1996, August). *Individual differences in judgmental ability.* Paper presented at the annual convention of the American Psychological Association, Toronto.