Self–Other Agreement in Personality and Affectivity: The Role of Acquaintanceship, Trait Visibility, and Assumed Similarity

David Watson, Brock Hubbard, and David Wiese
University of Iowa

Self- and other-ratings on the Big Five and a comprehensive inventory of trait affect were obtained from 74 married couples, 136 dating couples, and 279 friendship dyads. With the exception of Surprise, all scales showed significant self–other agreement in all 3 samples, thereby establishing their convergent validity. Consistent with the trait visibility effect, however, the Big Five consistently yielded higher agreement correlations than did the affectivity scales. Conversely, the affective traits consistently showed stronger evidence of assumed similarity (i.e., the tendency for judges to rate others as similar to themselves) than did the Big Five. Cross-sample comparisons indicated that agreement was significantly higher in the married sample than in the other 2 groups; however, analyses of 3 potential moderators in the dating and friendship samples failed to identify the source of this acquaintanceship effect.

Self–other agreement—that is, the convergence between self- and other-ratings of the same trait—is a centrally important issue in personality research. Personologists traditionally have used peer judgments as key evidence in the validation of self-ratings (e.g., Watson & Clark, 1991). Moreover, theorizing and research on this topic have played a crucial role in our understanding of personality itself. In 1934, for example, Mead proposed a model called "symbolic interactionism," arguing that "the individual experiences himself as such, not directly, but only indirectly, from the generalized standpoint of the social group as a whole to which he belongs" (Mead, 1934, p. 138). Mead’s model posited that a person’s self-concept emerges gradually in response to social feedback, and he clearly assumed a high level of self–other agreement.

However, early studies failed to support this assumption. In fact, in an influential review, Shrauger and Schoeneman (1979) concluded that the level of convergence between self- and other-raters was rather low. These discouraging results led to the argument that trait judgments essentially reflect "implicit personality theories" that have little basis in reality. Gradually, however, evidence from more carefully conducted studies established a strong level of convergence between both (a) self- and other-raters and (b) independent peer judges of the same target (e.g., Funder, 1995; Funder & Colvin, 1997; McCrae & Costa, 1987; Paunonen, 1989). Moreover, the data clearly demonstrated that trait judgments did not simply reflect internal cognitive representations but rather involve the systematic use of available information (Funder, 1995; Funder & Colvin, 1997). Consequently, evidence regarding self–other agreement played a crucial role in establishing the reality of traits (Kenrick & Funder, 1988).

It now is well established that substantial self–other agreement—reflecting correlations of .40 and higher—can be demonstrated across a broad range of traits. Much of the relevant research has focused on the prominent five-factor or "Big Five" model of personality, which consists of the general dimensions of Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness. Numerous studies have documented the existence of strong self–other agreement on all five of these traits; much has been learned, moreover, about the processes that influence the level of agreement (e.g., John & Robins, 1993; McCrae & Costa, 1987; Paulhus & Reynolds, 1995).

However, surprisingly few studies have examined self–other agreement for affective traits. This lack of research probably reflects the internal, subjective nature of affective experience. Given this subjective quality, most affect researchers have assumed that it would be quite difficult for peers to judge targets accurately. This view is not unreasonable. In fact, the accumulating data have established the existence of a trait visibility effect—that is, easily observable personality traits (i.e., those with clear, frequent behavioral manifestations) yield better interjudge agreement and higher self–other correlations than do more internal, subjective traits (e.g., Funder & Colvin, 1988, 1997; John & Robins, 1993; Kenrick & Stringfield, 1980; Watson & Clark, 1991). Nevertheless, a few studies have reported significant self–other convergence in judgments of trait affectivity. Much of the relevant research has examined global happiness and general well-being, obtaining agreement correlations in the .25 to .40 range (e.g., Hartmann, 1934; Kammann, Smith, Martin, & McQueen, 1984). In addition, a few investigators have examined agreement across multiple types of affect. For instance, McCrae (1982) obtained self- and spouse-ratings on the Anxiety, Hostility, Depression, and Positive Emotions facet scales of the NEO Personality Inventory (Costa & McCrae, 1992). These scales all yielded moderate to strong self-spouse agreement (rs ranged from .36 to .58). Watson and Clark (1991) found significant self-peer agreement on 8 different affect scales, with correlations ranging from .19 to .41. Finally, Diener, Smith, and Fujita (1995) obtained self- and other-
ratings on six specific types of emotion (e.g., Joy, Fear, Anger), as well as global positive and negative affect. They reported significant agreement on all 8 affect measures, with correlations ranging from .24 to .54.

The Current Study

Comprehensive Assessment of Affect

The current study was designed to examine self–other agreement in trait affectivity in greater depth. Our study extends the existing literature in several ways. First, our respondents completed the Expanded Form of the Positive and Negative Affect Schedule, or PANAS-X (Watson & Clark, 1994, 1997). The PANAS-X subsumes the earlier Positive and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988) and therefore includes scales assessing the broad dimensions of Positive and Negative Affect that repeatedly have emerged as higher order factors in mood research (Watson & Clark, 1997). In addition, the PANAS-X assesses 11 specific lower order affects: four different negative emotions (Fear, Sadness, Guilt, Hostility), three types of positive affect (Joviality, Self-Assurance, Attentiveness), and four other affective states (Shyness, Fatigue, Serenity, Surprise). Thus, this instrument enabled us to examine self–other agreement across a broader range of affective traits than have been studied previously.

Comparing Self–Other Agreement in Affectivity Versus the Big Five

As noted earlier, it is widely assumed that because of the trait visibility effect, affective traits should show lower levels of self–other convergence than other dispositional constructs. However, this assumption has never been examined empirically by assessing both affective and nonaffective traits in the same respondents. Accordingly, we tested it directly by obtaining self- and other-ratings on the Big Five in each of our samples.

In light of the trait visibility effect, we expected that affective traits would show lower self–other agreement than the Big Five. However, this prediction is complicated by the fact that four of the Big Five have strong links to trait affectivity (Watson & Clark, 1992). Neuroticism is strongly and broadly linked to negative affectivity. For instance, across combined samples with an overall N of 4,457, Watson, Wiese, Vaidya, and Tellegen (1999) obtained a correlation of .58 between Neuroticism and the trait form of the PANAS Negative Affect scale. Conversely, Extraversion is strongly associated with positive emotionality; Watson et al. (1999), for example, obtained a parallel correlation of .51 between Extraversion and trait Positive Affect. Furthermore, (low) Agreeableness and Conscientiousness are strongly correlated with the Hostility and Attentiveness scales, respectively, of the PANAS-X (Watson & Clark, 1992). In fact, among the Big Five, only Openness is weakly related to affectivity (Watson & Clark, 1992).

These strong links between the Big Five and affectivity raise some interesting questions. For instance, given that Neuroticism correlates strongly with negative affectivity—and, moreover, that it has been shown to be a relatively low-visibility dimension in relation to traits such as Extraversion (Funder & Colvin, 1988; Norman & Goldberg, 1966; Watson, 1989)—will it display levels of agreement that are similar to those of most affective traits? Or is Neuroticism a broader construct that yields higher levels of convergence? Our data permit a direct analysis of this issue by comparing self- and other-ratings of Neuroticism and trait affectivity in the same samples.

Further Examination of the Acquaintanceship Effect

We collected parallel sets of ratings in three different data sets, which were selected to explicate another robust finding in this literature—the acquaintanceship effect. Numerous studies have shown that self–other and interjudge agreement both improve with increasing levels of acquaintance; for instance, well-acquainted informants agree to a much greater extent with each other and with their targets than do relative strangers (Funder & Colvin, 1988, 1997; Funder, Kolar, & Blackman, 1995; Norman & Goldberg, 1966; Paulhus & Bruce, 1992; Paulhus & Reynolds, 1995).

Acquaintanceship is likely to have a particularly important influence on low-visibility dimensions, such as affective traits. Previous research has shown that high-visibility traits such as Extraversion show substantial self–other agreement even when relatively unacquainted judges are used. In contrast, ratings of Neuroticism show little or no convergence when unacquainted judges are used, but they display far greater agreement with increasing acquaintanceship (Funder & Colvin, 1988; Norman & Goldberg, 1966; Watson, 1989). These data demonstrate that judges need valid trait-relevant information in order to produce accurate ratings (Funder, 1995) and that low-visibility traits require relatively high levels of acquaintanceship before this information is readily available.

To date, only one study has examined the effects of acquaintanceship on affectivity ratings (Watson & Clark, 1991). Participants in this study were asked to indicate how well they knew the targets they were rating on a single 5-point scale (1 = do not know at all, 5 = know very well). Watson and Clark (1991) then compared the ratings made by the best acquainted (M = 4.64) and least acquainted (M = 2.88) judges. As expected, the mean self–other agreement correlation was significantly greater for the best acquainted peers (.25) than for the least acquainted judges (.15).

In the current study, we first examined acquaintanceship on a cross-sample basis. Our samples reflect three different types of relationships: friendship pairs, dating couples, and married couples. Spouses generally are assumed to represent the summit of the acquaintanceship continuum (McCrae, 1982; McCrae, Stone, Fagan, & Costa, 1998). Moreover, the couples in our sample had been married, on average, approximately 17 years, and most of them belonged to an organization that promoted improved communication skills between spouses. Consequently, we expected to
find the highest level of agreement in this group. It is more difficult
to make a clear prediction regarding the other two samples. One
could argue that the increased intimacy associated with a romantic
relationship would lead to greater convergence in the dating cou-
pies. Alternatively, friends might be more honest and forthright
with one another than dating couples, who at times may be
motivated to create a deceptively favorable impression on their
partner (Keenan, Gallup, Goulet, & Kulkarni, 1997). Moreover,
the average length of acquaintanceship was virtually identical
(approximately 3 years) in these samples. Consequently, we made
no prediction regarding the relative level of agreement across these
two samples.

One significant limitation of this cross-sample analysis, how-
ever, is that it fails to clarify how the effect actually works. For
instance, what specific aspects of the relationship (e.g., length of
acquaintanceship, frequency and duration of contact) are essential
in producing greater self-other agreement? To address this issue,
it is necessary to use an alternative design in which specific
features of the relationship are analyzed on an intrasample basis.
We therefore conducted additional analyses of this type, using
moderated multiple regression in both our friendship and dating
samples.

Previous moderator analyses of acquaintanceship have yielded
mixed results. Paunonen (1989) asked judges to rate their familiar-
ity with the target on a single 9-point scale and found that
self-other agreement increased with greater acquaintanceship. As
noted earlier, Watson and Clark (1991) also reported positive
findings using a very similar acquaintanceship index. However,
other studies have obtained largely negative results. For instance,
McCrae (1994) found that neither length of acquaintanceship nor
liking for the target were consistently related to the level of
self-other agreement. Similarly, McCrae et al. (1998) reported that
neither length of marriage nor marital satisfaction were significant
moderators of self-spouse agreement.

Our study extends this limited evidence by examining multiple
indicators of acquaintanceship in both our friendship and dating
samples. For several reasons, our sample of friendship dyads
provides a particularly promising opportunity to identify potential
moderators. First, this sample is much larger (279 dyads) than
those examined in most of the earlier studies. Sample size is an
important consideration because moderator effects in personality
tend to be quite small, typically corresponding to correlations of
approximately .10 (Chaplin, 1991). Second, we examined a large
number of low-visibility traits. As discussed earlier, low-visibility
traits show stronger acquaintanceship-related effects than high-
visibility dimensions such as Extraversion. Third, in recruiting
participants, we did not require that they be "best friends" or even
be acquainted; rather, we simply indicated that participants had to be accompanied by a "friend." Consequently, the
actual range of acquaintanceship varied considerably in this
sample. Finally, we examined a broader range of potential mod-
erators; specifically, we examined three different aspects of the
relationship (length of acquaintanceship, perceived closeness,
number of shared activities) in both samples.

The Role of Similarity and Assumed Similarity

Finally, previous studies have not extensively investigated the
potential influence of similarity and assumed similarity on affect
ratings. Cronbach (1955) initially drew interest to these phenom-
ena by noting that they can generate significant self-other agree-
ment in the absence of any true understanding of the target.
Cronbach argued that well-acquainted judges can appear to rate
their target accurately if people tend to (a) associate with those
who are similar to themselves and (b) base their other-ratings on
their own self-perceptions. In other words, assuming that they
actually do associate with similar others, judges can achieve a
spurious accuracy simply by assuming that others are similar to
them.

Cronbach’s argument therefore rests on two basic assumptions,
both of which must be true for this confound to operate (see
Funder et al., 1995). First, judges must show a tendency to rate
other persons as similar to themselves, a process that Cronbach
termed assumed similarity. In our study, assumed similarity would
be demonstrated by establishing significant correlations between
the two sets of ratings made by the same judge (e.g., between a
husband’s own self-rated Extraversion and his assessment of his
wife’s Extraversion). Several studies have established the presence
of assumed similarity in ratings of various personality traits (e.g.,
Funder et al., 1995).

The second assumption is that people must form relationships
with similar others; in other words, "birds of a feather must flock
together." In the case of married couples, this is known as "assor-
tative mating." In our study, it would be demonstrated by obtaining
significant correlations between parallel sets of self-ratings (e.g.,
between a husband’s and wife’s self-rated Extraversion). Assortative
mating is of keen interest to behavior geneticists because if it
occurs, people actually will be more similar to their first-degree
biological relatives than the 50% that is traditionally assumed.
Generally speaking, however, there appears to be little or no
assortative mating on traits such as Neuroticism and Extraversion
(e.g., Eysenck, 1990; Finkel & McGue, 1997). Indeed, after re-
viewing the available evidence, Eysenck (1990) concluded that
"mating is essentially random for personality differences" (p. 252).
Similarly, studies of friendship dyads have failed to demonstrate
strong evidence of similarity (Funder et al., 1995). It should be
noted, however, that McCrae (1996) reported significant evidence
of assortative mating on characteristics related to Openness.

Very little evidence regarding either similarity or assumed simi-
larity has been reported in the affect literature. In fact, the only
relevant results were reported by Kammann et al. (1984), who
analyzed ratings of general well-being in several small data sets.
Their findings generally were consistent with the broader person-
ality literature; that is, they found substantial evidence of assumed
similarity (with correlations ranging from .09 to .54), but not of
actual similarity (correlations ranged from .01 to .21). In light of
this very limited evidence, another goal of our study was to
examine similarity and assumed similarity across multiple samples
and a broad range of affective traits.

Method

Participants

The married sample consisted of 74 heterosexual couples from the St.
Louis, Missouri, area. Forty-eight couples were members of the Gateway
Chapter of the Association for Couples in Marriage Enrichment (ACME),
a nonprofit organization dedicated to improving marriages through teach-
ing couples improved communication skills and creative use of conflict.
Letters were sent to members explaining that $20 would be donated to ACME for every couple that participated in the study. ACME leadership then followed up with phone calls encouraging participation. The remaining couples consisted of married friends of ACME members who were invited by members to participate. The mean age of the sample was 47.1 years (range = 26 to 81 years). The mean length of the marriage was 202.6 months, that is, slightly less than 17 years (range = 2 to 699 months).

The dating sample consisted of 136 heterosexual couples from the Iowa City area who participated in a study of “currently dating couples.” At least 1 member of each couple was a student enrolled in a psychology course at the University of Iowa; these students participated either (a) to fulfill a course research participation requirement or (b) for extra course credit. Nonstudents received $5 for their participation. The couples had known each other for an average of 36.0 months (range = 1 to 192 months) and had been dating for an average of 18.2 months (range = 1 to 180 months).

The friendship sample consisted of 279 dyads drawn from the Iowa City area who were asked to participate in a study examining “the nature and quality of friendship.” At least 1 member of each dyad was a student enrolled in a psychology course at the University of Iowa; these students again participated either (a) to fulfill a course research requirement or (b) for extra course credit. Nonstudents were not compensated for their participation. The participants had known each other for an average of 33.6 months, that is, slightly less than 3 years (range = 1 to 239 months). As noted earlier, recruitment posters indicated simply that students had to be accompanied by a “friend”; no further specification of the degree of acquaintanceship was provided.

Measures

Affectivity. Participants in all three samples completed both self- and other-rating versions of the 60-item PANAS-X (Watson & Clark, 1994, 1997). Self-raters were asked to indicate on a 5-point scale (1 = very slightly or not at all, 2 = a little, 3 = moderately, 4 = quite a bit, 5 = extremely) “to what extent you generally feel this way, that is, how you feel on average.” The format and instructions for the other-ratings were identical, except that respondents were told to “indicate to what extent the person you are rating feels or acts this way, that is, how they feel or act on average.”

As noted earlier, the PANAS-X subsumes the earlier PANAS (Watson et al., 1988) and includes 10-item scales assessing the general dimensions of Positive Affect (e.g., active, alert, interested) and Negative Affect (e.g., afraid, irritable, upset). In addition, the PANAS-X contains 11 factor-analytically derived scales that assess specific types of affect. Four scales measure specific negative emotions that are strong markers of the higher order Negative Affect dimension (Watson & Clark, 1994, 1997): Fear (6 items; e.g., scared, nervous), Sadness (5 items; e.g., blue, lonely), Guilt (6 items; e.g., ashamed, angry at self), and Hostility (6 items; e.g., angry, scornful). In addition, three scales assess positively valenced states that are strongly linked to the general Positive Affect factor: Joviality (8 items; e.g., happy, enthusiastic), Self-Assurance (6 items; e.g., proud, confident), and Attentiveness (4 items; e.g., alert, concentrating). Finally, four scales assess affects that are less strongly and consistently related to the higher order dimensions: Shyness (4 items; e.g., bashful, timid), Fatigue (4 items; e.g., ashamed, angry at self), and Hostility (6 items; e.g., angry, scornful). Extensive data establish the reliability and validity of these scales (Watson & Clark, 1991, 1994, 1997). For instance, Watson et al. (1988) reported coefficient alphas for the higher order scales ranging from .84 to .87 for Negative Affect, and from .86 to .90 for Positive Affect. Similar values have been obtained in subsequent samples (Watson & Clark, 1994). Watson and Clark (1997) presented internal consistency data for the 11 lower order scales computed across 11 samples with a combined N of 8,194; the median coefficient alphas ranged from .76 (Serenity) to .93 (Joviality).
Closeness score for the dating couples was 17.6 (SD = 2.8, range = 8–20), whereas that for the friendship dyads was 15.9 (SD = 3.4, range = 5–20).

Procedure

All participants were assessed in small-group sessions. Both members of the dyad were required to be present at the same session. At the beginning of the session, the members of each pair identified themselves; at this time, they were given a unique dyadic identification number that allowed us to link their responses back together. To ensure honest and independent responding, they then were physically separated and asked to sit in different parts of the room. At that point, each participant was given a battery of instruments that included basic demographic information, the self- and other-versions of the Big Five and affectivity scales, and the relationship questionnaire.

Results

Analyses of Similarity

We first consider the actual degree of similarity between the members constituting each dyad. These analyses are important for two reasons. First, as discussed earlier, similarity may be a spurious source of self–other agreement. Second, similarity produces statistical nonindependence in data analyzed at the individual level (Kashy & Snyder, 1995; Kenny, 1995; Kenny & Kashy, 1991). That is, if scores between members of a dyad are systematically interrelated, then the assumption of independent observations is violated and significance tests may be biased and misleading. Kenny, Kashy and their colleagues therefore recommended that dyadic researchers begin by determining the level of nonindependence in their data. Kenny (1995) offers specific interpretative guidelines, suggesting that if the level of nonindependence is rather small (corresponding to similarity correlations of less than .30), then "it is relatively safe to use the person as the unit" (p. 72). If stronger similarity exists, however, then it is inappropriate to include individual data from both dyadic members in the same analyses.

Accordingly, Table 1 presents similarity correlations for the Big Five and affectivity scales in each sample. Because dyadic members constituting heterosexual couples can be clearly distinguished by gender, the values shown in the dating and married samples are familiar Pearson product-moment correlations. In the friendship sample, however, dyadic members can be distinguished only on some arbitrary basis; consequently, we computed intraclass correlations in this case (Kashy & Snyder, 1995; Kenny, 1995).

2 To compute a similarity correlation, one first must create two variables (e.g., "X" and "Y") that represent the individual members of each dyad. In the married and dating samples, this is quite straightforward: For instance, the woman always can be designated as X, and the man invariably can be assigned to Y. Pearson ("intraclass") correlations therefore are appropriate in this instance. The problem in the friendship sample is that it is unclear who should be designated as X, and who should be assigned to Y. Note, moreover, that one will obtain different Pearson correlations depending on which member is arbitrarily designated as X and Y. Consequently, intraclass correlations must be computed in these data.

It is noteworthy that only 3 of the 54 correlations are negative; in other words, we see no evidence to support the popular notion that "opposites attract." More importantly, however, the correlations also tend to be quite small. In fact, only 9 coefficients (16.7%) are .20 or greater, and only 3 (5.6%) exceed .30. Thus, consistent with the broader personality literature, we see little evidence of dyadic similarity either in the Big Five or in trait affectivity. Note, moreover, that these results demonstrate that nonindependence does not represent a substantial problem in our trait scales. Consequently, we conduct our basic analyses of the personality ratings at the individual—rather than the dyadic—level. Nevertheless, because even a modest level of bias potentially can produce misleading results (Kenny, 1995), we computed adjusted p values and degrees of freedom in these data, using the correction program presented by Kenny (1995, Appendix C); these corrected p values are used in all relevant analyses.

We next consider the similarity correlations for our three acquaintanceship measures. Again, we computed Pearson correlations in the dating couples but intraclass correlations in the friendship dyads. As one would expect, these acquaintanceship indicators were strongly intercorrelated, indicating that the dyadic members viewed their relationship in largely similar terms. Not surprisingly, the similarity correlations approached +1.00 for Acquaintanceship Length (r = .96 and .93 in the friendship and dating samples, respectively), a variable that is highly objective and easy to rate. The coefficients for the other indicators (Number of Shared Activities, Relationship Closeness) were lower but still substantial, ranging from .46 to .68 across the two samples. Thus, dyadic members also tended to rate the scope and closeness of their relationship similarly.

It clearly would be inappropriate to include individual acquaintanceship data from both dyadic members in the same analyses. Therefore, in conducting our analyses of acquaintanceship, we
began by creating two subsamples—consisting of 1 member from each dyad—in both of these data sets. In the dating couples, these subsamples were nonarbitrarily formed along gender lines; that is, we conducted separate analyses of the male and female targets. This obviously was not possible in the friendship sample; here, we created two random subsamples by arbitrarily assigning 1 member of each dyad to each group.

Relations Between Affectivity and the Big Five

As discussed earlier, four of the Big Five traits have been found to be strongly correlated with affectivity (e.g., Watson & Clark, 1992). Before turning to our main analyses, it is worth examining whether similar associations emerged in the current data sets. To address this issue, we computed mean correlations between the self-rated Big Five and PANAS-X scales across the three samples; these values were calculated by (a) computing the individual correlations in each sample, (b) subjecting them to an $r$ to $z$ transformation, (c) averaging them, and then (d) converting them back to normal $r$s.

In general, our results replicate those reported in earlier studies. Once again, Neuroticism was strongly and broadly correlated with negative emotionality, displaying particularly strong associations with general Negative Affect ($r = .60$), Guilt ($r = .57$), Fear ($r = .55$), Sadness ($r = .54$) and Serenity ($r = -.52$). Conversely, Extraversion was strongly linked to positive emotionality, correlating .52 with general Positive Affect, .62 with Joviality, and .51 with Self-Assurance; not surprisingly, it also was inversely related to Shyness ($r = -.48$). Consistent with previous research (Watson & Clark, 1992), Conscientiousness and Agreeableness showed more specific, lower order relations with affectivity: The former was strongly correlated with Attentiveness ($r = .51$), whereas the latter was inversely related to Hostility ($r = -.47$). Finally, as in earlier studies, Openness was not strongly correlated with any type of trait affect.

Parallel analyses of the other-ratings yielded very similar results. Indeed, the only real difference was that the correlations tended to be somewhat higher in these data. For instance, Neuroticism had mean correlations of .69 with general Negative Affect, .63 with Sadness, .62 with Fear, and .61 with Guilt. Similarly, Extraversion had average correlations of .59 with general Positive Affect, .68 with Joviality, and .53 with Self-Assurance. Finally, paralleling the self-report data, Conscientiousness and Agreeableness had mean correlations of .60 and -.59 with Attentiveness and Hostility, respectively.

Thus, our data again establish that with the sole exception of Openness, the Big Five are strongly linked to trait affectivity. The correlations between Neuroticism and negative affectivity are particularly strong, with numerous coefficients in the .50 to .70 range. As discussed previously, these strong links suggest that Neuroticism and trait negative affectivity might show very similar levels of self–other agreement. We turn now to an examination of this issue.

Correlational Analyses of Self–Other Agreement

Table 2 presents the self–other agreement correlations for the Big Five and affectivity scales in each sample. In addition to the

<table>
<thead>
<tr>
<th>Scale</th>
<th>Friendship dyads</th>
<th>Dating couples</th>
<th>Married couples</th>
<th>Mean r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big Five scales</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neuroticism</td>
<td>.37</td>
<td>.41</td>
<td>.59</td>
<td>.46</td>
</tr>
<tr>
<td>Extraversion</td>
<td>.48</td>
<td>.45</td>
<td>.61</td>
<td>.52</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>.39</td>
<td>.53</td>
<td>.49</td>
<td>.47</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>.34</td>
<td>.39</td>
<td>.53</td>
<td>.42</td>
</tr>
<tr>
<td>Openness</td>
<td>.44</td>
<td>.55</td>
<td>.58</td>
<td>.53</td>
</tr>
<tr>
<td>Mean r</td>
<td>.41</td>
<td>.47</td>
<td>.56</td>
<td></td>
</tr>
<tr>
<td>Negative affect scales</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General Negative Affect</td>
<td>.20</td>
<td>.22</td>
<td>.44</td>
<td>.29</td>
</tr>
<tr>
<td>Fear</td>
<td>.20</td>
<td>.20</td>
<td>.36</td>
<td>.26</td>
</tr>
<tr>
<td>Sadness</td>
<td>.31</td>
<td>.32</td>
<td>.47</td>
<td>.37</td>
</tr>
<tr>
<td>Guilt</td>
<td>.27</td>
<td>.26</td>
<td>.49</td>
<td>.34</td>
</tr>
<tr>
<td>Hostility</td>
<td>.21</td>
<td>.32</td>
<td>.50</td>
<td>.35</td>
</tr>
<tr>
<td>Mean r</td>
<td>.24</td>
<td>.26</td>
<td>.45</td>
<td></td>
</tr>
<tr>
<td>Positive affect scales</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General Positive Affect</td>
<td>.30</td>
<td>.33</td>
<td>.39</td>
<td>.34</td>
</tr>
<tr>
<td>Joviality</td>
<td>.38</td>
<td>.38</td>
<td>.51</td>
<td>.43</td>
</tr>
<tr>
<td>Self-Assurance</td>
<td>.36</td>
<td>.38</td>
<td>.52</td>
<td>.42</td>
</tr>
<tr>
<td>Attentiveness</td>
<td>.28</td>
<td>.29</td>
<td>.26</td>
<td>.28</td>
</tr>
<tr>
<td>Mean r</td>
<td>.33</td>
<td>.35</td>
<td>.43</td>
<td></td>
</tr>
<tr>
<td>Other affect scales</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shyness</td>
<td>.37</td>
<td>.28</td>
<td>.36</td>
<td>.34</td>
</tr>
<tr>
<td>Fatigue</td>
<td>.13</td>
<td>.17</td>
<td>.53</td>
<td>.29</td>
</tr>
<tr>
<td>Serenity</td>
<td>.21</td>
<td>.17</td>
<td>.38</td>
<td>.26</td>
</tr>
<tr>
<td>Surprise</td>
<td>.18</td>
<td>.17</td>
<td>.10</td>
<td>.15</td>
</tr>
<tr>
<td>Mean r</td>
<td>.22</td>
<td>.20</td>
<td>.35</td>
<td></td>
</tr>
</tbody>
</table>

Note. ns = 558 (friendship dyads), 272 (dating couples), 148 (married couples). With the exception of Surprise in the married couple sample, all rs are significant at $p < .01$, two-tailed.
individual coefficients. Table 2 also reports mean correlations computed (a) for each scale across all three samples and (b) for each type of scale (e.g., Big Five, negative affectivity measures) within each sample. These mean values were calculated by (a) subjecting the individual coefficients to an r to z transformation, (b) averaging them, and then (c) converting them back to normal rs.

Several aspects of these data are noteworthy. First, we again see clear evidence of the familiar acquaintanceship effect. As expected, self–other agreement tended to be substantially higher in the married couples than in the other two samples. This is corroborated by the results of paired-observation r tests, which indicated that the overall mean correlation in the married sample (.46) was significantly greater than the corresponding values in both the friendship dyads (.30), t(17) = 5.62, p < .01, and the dating couples (.33), t(17) = 5.02, p < .01; the latter two means, however, did not differ significantly from one another, t(17) = 1.68.

It also is clear, however, that some scales showed stronger acquaintanceship-related effects than others. In general, the biggest cross-sample differences were observed for measures of Neuroticism and negative affectivity. Ratings of Neuroticism showed moderate self–other agreement in both the friendship (r = .37) and dating (r = .41) samples but very strong agreement in the married couples (r = .59; both differences are significant at p < .01). Similarly, the agreement correlations for the negative affectivity scales were relatively modest in the friendship dyads (mean r = .24) and dating couples (mean r = .26) but substantially higher in the married couples (mean r = .45). In contrast, although measures of positive emotionality yielded somewhat better self–other agreement in the friendship and dating samples (mean rs = .33 and .35, respectively) than did the negative emotionality scales, they benefited less from the increased acquaintanceship of the married couples (mean r = .43).

The Table 2 data establish that a substantial level of acquaintanceship is required to judge affectivity accurately, particularly in the case of negative emotionality. It also must be emphasized, however, that once this level is attained (as in the case of the married couples), most affective traits tend to show moderate to strong levels of self–other agreement. In fact, the median correlation for the PANAS-X scales was .44 in the married sample, demonstrating a substantial level of agreement. Moreover, several scales yielded agreement correlations of .50 and greater, with the highest coefficients being observed for Fatigue (.53), Self-Assurance (.52), Joviality (.51), and Hostility (.50). At the other extreme, however, ratings of Surprise failed to show substantial agreement in any sample, including the married couples (r = .10). These negative findings are consistent with other data that challenge the construct validity of trait ratings of Surprise (Watson & Clark, 1994). Consequently, it seems reasonable to conclude that this affective state does not have a meaningful dispositional counterpart. The data for Surprise clearly represent an exception, however; more generally, our results offer particularly strong evidence that helps to establish the convergent validity of affective traits.

Having said that, however, we also must acknowledge another important aspect of these results, namely, that—consistent with the familiar trait visibility effect—the agreement correlations tended to be substantially higher for the Big Five than for trait affectivity. In fact, although the PANAS-X scales clearly benefited from increasing acquaintance, the Big Five traits nevertheless showed greater self–other agreement in all three samples, with mean correlations of .41 (vs. .26) in the friendship dyads, .47 (vs. .27) in the dating couples, and .56 (vs. .41) in the married couples.

This pattern is particularly striking when one recalls the strong associations between the Big Five and trait affectivity that we discussed earlier. Most notably, we again obtained very strong correlations (in the .60 to .70 range) between Neuroticism and the general Negative Affect scale of the PANAS-X in our samples. Nevertheless, Neuroticism showed higher self–other agreement (with rs ranging from .37 to .59) than general Negative Affect (rs ranged from .20 to .44) in all three samples; moreover, these agreement correlations differed significantly from one another in both the friendship (z = 3.56, p < .01) and dating (z = 2.94, p < .01) samples. Similarly, Extraversion displayed significantly greater agreement than general Positive Affect in both the friendship dyads (.48 vs. .30; z = 3.70, p < .01) and married couples (.61 vs. .39; z = 2.78, p < .01).

These correlational differences cannot be attributed to differential reliability across scales. For instance, coefficient alphas for the NEO-FFI Neuroticism scale were .85 (self-ratings) and .85 (partner-ratings) in the dating couples; parallel values for the PANAS-X Negative Affect scale were .85 and .88, respectively. Accordingly, if the correlations shown in Table 2 were corrected for attenuation due to unreliability, the agreement coefficient for Neuroticism (.48) in the dating sample still would be significantly higher than that for general Negative Affect (.25). Similar results emerged in the other two samples. It therefore appears that although measures of Neuroticism and trait Negative Affect are strongly interrelated, the former subsume a somewhat broader range of content; this expanded coverage presumably includes nonaffective content that is somewhat more visible and, consequently, easier to judge in others.

More generally, the Table 2 results demonstrate that certain traits consistently show substantially better agreement than others. Indeed, one unique strength of our cross-sample design is that it allows us to demonstrate the consistency of these agreement correlations across different types of relationships. To quantify this consistency, we computed correlations between the 18 agreement coefficients that were obtained in each sample; for instance, we correlated the 18 coefficients in the friendship sample with the corresponding values in the dating sample. These correlations are reported in Table 3; they range from .52 (married couples vs. friendship dyads) to .88 (dating couples vs. friendship dyads) and clearly establish that certain traits consistently are easier to rate than others. In our data, the best overall agreement was observed for the Big Five traits of Openness (mean r = .53) and Extraversion (mean r = .52); at the other extreme is the aforementioned Surprise, which produced a mean agreement correlation of only .15.

3 In a related vein, one might wonder whether the cross-sample comparisons in Table 2 were influenced by the differential reliabilities of the BFI (used in the friendship sample) versus the NEO-FFI (used in the dating and married couples). To examine this issue, we corrected all of the Big Five agreement correlations for attenuation due to unreliability. These analyses revealed a slightly stronger acquaintanceship effect: Specifically, across the Big Five traits, the corrected mean correlations were .69 (married couples), .57 (dating couples), and .48 (friendship dyads).
Table 3
Correlations Among the Correlational Arrays Presented in Tables 2 (Agreement Correlations) and 4 (Assumed Similarity Correlations)

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agreement correlations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Friendship dyads</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Dating couples</td>
<td>.88**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Married couples</td>
<td>.52*</td>
<td>.61**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assumed similarity correlations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Friendship dyads</td>
<td>-.77**</td>
<td>-.65**</td>
<td>-.40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Dating couples</td>
<td>-.32</td>
<td>-.29</td>
<td>-.30</td>
<td>.66**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Married couples</td>
<td>-.70**</td>
<td>-.68**</td>
<td>-.52*</td>
<td>.78**</td>
<td>.66**</td>
<td></td>
</tr>
</tbody>
</table>

Note. n = 18.
*p < .05, two-tailed. **p < .01, two-tailed.

Analyses of Assumed Similarity

We next consider the issue of assumed similarity, that is, whether judges tend to rate others as similar to themselves. We should note that even if it exists, assumed similarity is unlikely to be a significant source of spurious accuracy in our data. This is, of course, because we found little evidence of actual dyadic similarity in our samples (see Table 1).

Table 4 presents the assumed similarity correlations—that is, correlations between a judge’s own self-rating and his or her other-rating on the same trait—for the Big Five and affectivity scales in each sample. As in Table 2, Table 4 also reports mean correlations computed for (a) each scale across all three samples and (b) each type of scale within each sample. One interesting aspect of these data is that the cross-sample pattern is different here than in the agreement correlations. Overall, the assumed similarity correlations tended to be highest in the dating couples (mean r = .38); in fact, paired t tests indicated that the mean correlation in this sample was significantly greater than that in the married couples, mean r = .29; t(17) = 3.28, p < .01, and was marginally greater than that in the friendship dyads, mean r = .32; t(17) = 2.05, p < .06. Thus, assumed similarity shows no clear relation to acquaintanceship, at least when assessed on a cross-sample basis.

Two other aspects of these data are noteworthy. First, parallelizing the pattern observed with self–other agreement, we again find

Table 4
Assumed Similarity Correlations in the Three Samples

<table>
<thead>
<tr>
<th>Scale</th>
<th>Friendship dyads</th>
<th>Dating couples</th>
<th>Married couples</th>
<th>Mean r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big Five scales</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neuroticism</td>
<td>.13**</td>
<td>.19**</td>
<td>.20*</td>
<td>.17</td>
</tr>
<tr>
<td>Extraversion</td>
<td>.08</td>
<td>.27*</td>
<td>.18</td>
<td>.18</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>.17*</td>
<td>.11</td>
<td>.11</td>
<td>.13</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>.28**</td>
<td>.18**</td>
<td>.18*</td>
<td>.21</td>
</tr>
<tr>
<td>Openness</td>
<td>.23**</td>
<td>.50**</td>
<td>.18*</td>
<td>.31</td>
</tr>
<tr>
<td>Mean r</td>
<td>.18</td>
<td>.26</td>
<td>.17</td>
<td></td>
</tr>
<tr>
<td>Negative affect scales</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General Negative Affect</td>
<td>.48**</td>
<td>.45**</td>
<td>.42**</td>
<td>.45</td>
</tr>
<tr>
<td>Fear</td>
<td>.45**</td>
<td>.29**</td>
<td>.27**</td>
<td>.34</td>
</tr>
<tr>
<td>Sadness</td>
<td>.34**</td>
<td>.37**</td>
<td>.46**</td>
<td>.39</td>
</tr>
<tr>
<td>Guilt</td>
<td>.42**</td>
<td>.38**</td>
<td>.36**</td>
<td>.39</td>
</tr>
<tr>
<td>Hostility</td>
<td>.40**</td>
<td>.50**</td>
<td>.40**</td>
<td>.43</td>
</tr>
<tr>
<td>Mean r</td>
<td>.42</td>
<td>.40</td>
<td>.38</td>
<td></td>
</tr>
<tr>
<td>Positive affect scales</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General Positive Affect</td>
<td>.39**</td>
<td>.57**</td>
<td>.34**</td>
<td>.44</td>
</tr>
<tr>
<td>Joviality</td>
<td>.42**</td>
<td>.54**</td>
<td>.26**</td>
<td>.41</td>
</tr>
<tr>
<td>Self-Assurance</td>
<td>.22**</td>
<td>.31**</td>
<td>.15</td>
<td>.23</td>
</tr>
<tr>
<td>Attentiveness</td>
<td>.23**</td>
<td>.35**</td>
<td>.27**</td>
<td>.28</td>
</tr>
<tr>
<td>Mean r</td>
<td>.32</td>
<td>.45</td>
<td>.26</td>
<td></td>
</tr>
<tr>
<td>Other affect scales</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shyness</td>
<td>.18**</td>
<td>.34**</td>
<td>.21**</td>
<td>.24</td>
</tr>
<tr>
<td>Fatigue</td>
<td>.47**</td>
<td>.45**</td>
<td>.35**</td>
<td>.42</td>
</tr>
<tr>
<td>Serenity</td>
<td>.25**</td>
<td>.26**</td>
<td>.25**</td>
<td>.25</td>
</tr>
<tr>
<td>Surprise</td>
<td>.51**</td>
<td>.56**</td>
<td>.54**</td>
<td>.54</td>
</tr>
<tr>
<td>Mean r</td>
<td>.36</td>
<td>.41</td>
<td>.34</td>
<td></td>
</tr>
</tbody>
</table>

Note. ns = 558 (friendship dyads), 272 (dating couples), 148 (married couples).
*p < .05, two-tailed. **p < .01, two-tailed.
that certain traits showed consistently stronger evidence of assumed similarity than others. To quantify this effect, we computed correlations between the three correlational arrays displayed in Table 4; for instance, we calculated correlations between the 18 coefficients in the dating sample with the corresponding values in the married couples. These correlations also are presented in Table 3, and they establish a high degree of consistency across the three samples (rs ranged from .66 to .78). Thus, certain traits consistently show greater evidence of "projection" than others. In our data, assumed similarity was greatest for ratings of Surprise (mean r = .54), general Negative Affect (mean r = .45), general Positive Affect (mean r = .44), and Hostility (mean r = .43) and was weakest for the Big Five traits of Conscientiousness (mean r = .13), Neuroticism (mean r = .17), and Extraversion (mean r = .18).

Second, assumed similarity tends to be greatest for those traits that showed relatively weak self–other agreement. The Big Five showed substantially better self–other agreement than the affective traits; here, we see the opposite pattern, as assumed similarity tends to be greatest in ratings of negative affectivity (with mean rs ranging from .38 to .42 across the three samples) and weakest for the Big Five (with mean rs ranging from .17 to .26). To quantify this effect, we correlated the three arrays of agreement coefficients (shown in Table 2) with their counterparts in Table 4; these correlations also are displayed in Table 3. It is noteworthy that all of these correlations are negative and that those in the friendship dyads (−.77) and dating couples (−.52) reached significance. Thus, we see a general tendency for assumed similarity to be greatest when self–other agreement is low. The most obvious manifestation of this trend is Surprise, which displayed both the highest overall level of assumed similarity (mean r = .54) and the lowest level of self–other agreement (mean r = .15). More generally, it is striking that many of the PANAS-X scales yielded higher assumed similarity correlations than agreement correlations in the friendship and dating samples. This effect is most pronounced for measures of negative affectivity, which generated mean assumed similarity correlations of .42 and .40 in the friendship and dating samples, respectively; in contrast, the corresponding values for the agreement correlations were only .24 and .26, respectively.

These results support the suggestion made by earlier writers (Dawes, 1990; Funder et al., 1995) that assumed similarity may, in part, represent a heuristic that judges use when they lack adequate trait-relevant information. When judges know their targets very well—or when they are rating a high-visibility characteristic such as Extraversion—they easily can make accurate and distinctive judgments based on their direct knowledge of the person’s standing on the trait. But what can judges do when they are asked to rate low-visibility traits (e.g., Surprise or Fear) in targets whom they do not know intimately? In this context of limited information, one understandable—perhaps even rational—strategy is to use their own self-rating as a starting point and to assume that the targets resemble them to some extent. Because the actual degree of similarity is slight (see Table 1), this heuristic is not particularly effective; nevertheless, it seems to represent a realistic response to a difficult judgmental situation that offers no good alternative. As Funder et al. (1995) put it, "strangers rely heavily on their self-judgments because they must" (p. 669).

**Moderator Analyses of Acquaintanceship**

Thus far, we have examined acquaintanceship only on a cross-sample basis. We now investigate this important issue on an intrasample basis, analyzing our three potential moderator variables—Acquaintanceship Length, Relationship Closeness, and Number of Shared Activities—in the friendship and dating samples. Because of the nonindependence problem that was discussed previously, we conducted separate analyses in two subsamples—consisting of 1 member from each dyad—in each of these data sets. In the dating couples, this simply involved separate analyses of the male and female raters. In the friendship dyads, however, we created two random subsamples by arbitrarily assigning 1 member of each dyad to each group.

Our initial analyses indicated that Relationship Closeness and the Number of Shared Activities were moderately to strongly interrelated, with correlations ranging from .33 to .60 across the various subsamples. In contrast, Acquaintanceship Length was only modestly correlated with the other relationship variables, with rs ranging from .18 to .34. Clearly, the temporal duration of a relationship is a relatively poor predictor of its scope and intimacy. More fundamentally, these results demonstrate that acquaintance-ship is a complex, multifaceted phenomenon that cannot easily be reduced to a single item, or even a single dimension (see also Colvin & Funder, 1991). We therefore analyzed these variables separately in our moderated multiple regressions.

Do any of these acquaintanceship variables influence the level of self–other agreement? In considering this issue, it again is important to emphasize that both the dating and friendship samples—but especially the latter—were characterized by broad variation in acquaintanceship. For instance, although more than a quarter (25.6%) of the friendship pairs had known each other for 6 months or less, approximately one sixth (16.6%) had known each other for at least 5 years. Overall, the participants in this sample had known each other for as little as 1 month, and as long as 20 years.

To investigate whether acquaintanceship influenced the level of self–other agreement, we conducted a series of moderated multiple regressions, using the other-ratings as predictors and the self-ratings as the criteria. In light of the exploratory nature of these analyses—and to reduce the overall number of analyses that were conducted—we restricted these moderated regressions to 7 trait scales: the Big Five and the general Negative and Positive Affect scales of the PANAS-X.

In these moderated regressions, the two main effects were entered as a block in Step 1, followed by the interaction term in Step 2. Following the recommendation of Aiken and West (1991), all interaction terms were centered to reduce collinearity. As an example, consider the analysis in which Relationship Closeness was used as a potential moderator of Neuroticism. In this analysis, self-rated Neuroticism served as the criterion variable. The two main effects—other-rated Neuroticism and Relationship Closeness—were entered as a block in Step 1, followed by the centered interaction term (the product of the two main effects) in Step 2. Note that we actually ran four parallel analyses (conducted on the two subsamples in each data set) in each case. This allowed us to investigate the replicability of our results on both an intra- and intersample basis.
Overall, these results were quite disappointing. Only 8 of the 84 individual regressions (9.5%) yielded a significant moderator effect. Moreover, none of the significant effects replicated, even within the second subsample in the same data set; indeed, no effect was significant in more than one analysis. When we ignored statistical significance and simply examined the size of the interaction effect, we found that no moderator consistently accounted for at least 1% of the variance across all four analyses; in fact, in only 2 of 42 cases (i.e., using the Number of Shared Activities to predict Conscientiousness in the friendship dyads; using Relationship Closeness to predict Agreeableness in the dating couples) did a moderator account for at least 1% of the variance in both subsamples of the same data set. Clearly, none of these acquaintanceship indicators emerged as a strong, consistent moderator of self–other agreement in our data.

Discussion

In Search of the Acquaintanceship Effect

Our data augment the existing literature on acquaintanceship in several ways. First, our results support the argument made by earlier writers (Colvin & Funder, 1991; Funder, 1995) that acquaintanceship is a multifaceted phenomenon that cannot be reduced to a single dimension. It is unfortunate that few researchers actually have paid much attention to the conceptualization and assessment of acquaintanceship; typically, in fact, it either (a) has been assessed using a single item (e.g., Paunonen, 1989; Watson & Clark, 1991) or else (b) simply has been inferred from cross-sample comparisons involving different types of relationships (e.g., Norman & Goldberg, 1966; Watson, 1989). The approach used in this study represents a substantial improvement, in that we examined multiple indicators of acquaintanceship in two different types of relationships. Our approach is hardly definitive, however, and it is important that future studies of this issue devote greater attention to the measurement of acquaintanceship.

In addition, our cross-sample analyses (see Table 2) again demonstrated that self–other agreement differs systematically across different types of relationships. As expected, our sample of married couples yielded a significantly higher level of self–other agreement (mean r = .46) than either the dating couples (mean r = .33) or the friendship dyads (mean r = .30). Generally speaking, measures of Neuroticism and negative emotionality showed the largest cross-sample differences, demonstrating that this general domain of personality is particularly susceptible to acquaintanceship-related effects.

Having said this, however, we must acknowledge a limitation of our design, namely, that these cross-sample differences actually may not be attributable to acquaintanceship per se but perhaps may reflect some other variable. Our married couples were significantly older and more experienced than the participants in the other samples, and it is entirely possible that they possessed other characteristics (e.g., greater perceptiveness and psychological insight, increased empathy) that made them better, more accurate judges of their spouses (see Funder, 1995). Note, moreover, that many of them belonged to an organization that promoted improved communication skills; this also could have played a significant role in their enhanced level of agreement. On the other hand, previous studies have shown that increasing acquaintanceship is associated with greater self–other agreement (a) when the same judges rate different targets (Funder & Colvin, 1988) and (b) in the same individuals over time (Paulhus & Bruce, 1992; Paulhus & Reynolds, 1995). Consequently, it seems reasonable to conclude that our cross-sample differences are at least partly a function of increased acquaintanceship.

Unfortunately, our moderator analyses failed to clarify how this effect works. The results from our friendship sample were particularly discouraging because we attempted to create exceptionally favorable conditions for identifying potential moderators; that is, we examined (a) several low-visibility traits (b) across various relationship indicators (c) in a very large sample that (d) included friendship dyads at widely varying levels of acquaintanceship.

To illustrate this point further, consider the results for general Negative Affect. As shown in Table 2, the overall self–other agreement correlation for this scale was .20 in the friendship dyads, indicating a significant—but rather modest—level of convergence. The corresponding correlation in the married couples was .44, suggesting that this trait is highly susceptible to acquaintanceship-related effects. Moreover, as discussed previously, our indicators of acquaintanceship varied enormously within the friendship sample. We already have noted that these friendship pairs had known each other for as little as 1 month, and as long as 20 years. Similarly, scores on Relationship Closeness varied from 5 to 20 (possible range = 4–20) in this sample, whereas those on the Number of Shared Activities ranged from 1 to 25 (possible range = 0–25). Despite these exceptionally favorable conditions, we found no significant moderator effects on this trait.

It seems reasonable to assume that the acquaintanceship effect arises because judges acquire more trait-relevant information as they come to know the target better (Funder, 1995; Funder & Colvin, 1997; Funder et al., 1995). However, the effect itself remains somewhat enigmatic, and we still know very little about the specific processes that make it work. Clearly, it is important that future research be directed toward explicating the mechanisms underlying this effect. In this regard, we agree with earlier writers (e.g., Funder, 1995) that well-controlled longitudinal studies can play a particularly important role in clarifying these mechanisms. To date, only Paulhus and his colleagues (Paulhus & Bruce, 1992; Paulhus & Reynolds, 1995) have examined the effects of increasing acquaintanceship in the same groups over time. Longitudinal research is particularly important because acquaintanceship ultimately may be more a function of the nature of the relationship that exists between the judge and target, rather than specific quantitative indicators such as length of acquaintanceship. For instance, agreement may increase systematically as one moves across different relationship stages, that is, as one moves from being a "stranger" to being a "casual acquaintance" and then to being a "friend." Once a true friendship is formed, however, further quantitative increases in acquaintanceship may prove to be relatively unimportant. This would help to explain why acquaintanceship effects appear to be easier to establish on a cross-sample than a within-sample basis.

Trait Visibility: Clarifying the Qualities of a "Good Trait"

Previous research has shown that certain traits are easier to judge than others (Funder, 1995; Funder & Colvin, 1988, 1997). To date, however, few investigators actually have investigated the consistency of these differences across different samples and con-
dations. One notable exception is Funder and Colvin (1988), who reported a .44 correlation between arrays of agreement correlations (reflecting 100 individual items) representing two different types of relationships (self-informant vs. self-stranger).

We were able to extend this limited evidence by examining agreement correlations across three different types of relationships. Our results established that certain traits consistently show better agreement than others. Indeed, agreement exhibited a strong level of "cross-situational consistency" in our data, with cross-sample correlations ranging from .52 (married couples vs. friendship dyads) to .88 (dating couples vs. friendship dyads).

What specific attributes produce this phenomenon of the "good trait" (Funder, 1995), that is, a trait that is relatively easy to judge in others? As discussed earlier, the commonplace explanation is that it is a function of visibility; that is, easily observable dimensions (i.e., those with clear, frequent behavioral manifestations) yield better agreement than do more internal, subjective traits (e.g., Funder, 1995; Funder & Colvin, 1988, 1997; Watson & Clark, 1991). Funder (1995) described this phenomenon in more basic psychological terms, arguing that "good traits" are those that provide the rater with more frequent and informative cues regarding the target's standing on the dimension.

Our data are entirely consistent with this argument. Still, it would be helpful to have a clearer idea of exactly what distinguishes a "good trait" from a "bad trait." One interesting aspect of our design is that it permitted the first direct comparison between affective traits and strongly correlated dimensions of the Big Five. The most striking comparison involved Neuroticism and the general Negative Affect scale of the PANAS-X. Replicating earlier studies (e.g., Watson & Clark, 1992), we obtained average correlations of .60 (self-ratings) and .69 (other-ratings) between these scales. Nevertheless, the Neuroticism scales yielded significantly better self–other agreement in two of our three samples; overall, the mean agreement correlation for Neuroticism (.46) was substantially higher than that for general Negative Affect (.29). Moreover, these differences emerged despite the fact that our measures of these constructs were virtually identical in length (ranging from 9 to 12 items) and in reliability.

How can we explain these strong, consistent differences in agreement across constructs that clearly overlap to a substantial degree? Earlier, we suggested that they reflect the fact that—in addition to assessing variance related to negative emotionality—Neuroticism scales contain nonaffective content that is somewhat more visible and, consequently, easier to judge in others. Although this explanation seems plausible, it is difficult to identify specific items in either the NEO-FFI or BFI Neuroticism scales that are completely nonaffective in character. The BFI items (e.g., "is depressed, blue"); "can be tense"); "worries a lot"); "gets nervous easily") show a particularly strong resemblance to the PANAS-X, with regard to both format and content.

With regard to the NEO-FFI, however, we were able to identify four Neuroticism items that tapped cognitive manifestations of depression and low self-esteem (i.e., feeling inferior to others; feeling worthless, helpless, and discouraged) and that showed relatively high levels of self–other agreement; in contrast, 6-items containing PANAS descriptors (i.e., feeling blue, sad, fearful, jittery, angry, ashamed) tended to show more modest convergence. When they were combined into scales, the depression–low self-esteem items showed significantly greater agreement (r = .41) in the dating couples than did the PANAS items (r = .25; z = 2.41, p < .05, two-tailed); analyses of the married couples showed a similar pattern (agreement rs = .57 vs..45, respectively), but this difference failed to reach significance (z = 1.53). These results suggest that relatively subtle differences in item content (e.g., assessing feelings of worthlessness and inferiority vs. feelings of sadness and depression) may have a significant effect on the visibility of trait scales. This is an important issue that merits greater attention in the future.

**Similarity and Assumed Similarity**

Do "birds of a feather flock together"? Or, alternatively, do "opposites attract"? Somewhat surprisingly, with the possible exception of traits related to Openness (McCrae, 1996), the accumulating evidence indicates that personality similarity–dissimilarity actually plays little role in the formation of close relationships (Eysenck, 1990; Finkel & McGue, 1997; Funder et al., 1995). Our data extend this literature by examining similarity in measures of trait affectivity. Note, moreover, that our cross-sample design provided us with an unusual opportunity to analyze similarity correlations in the same traits across three different types of relationships. Consistent with this larger literature, our analyses of the PANAS-X scales revealed very little evidence of dyadic similarity: Only 6 of the 39 similarity correlations were .20 or greater, and only 2 exceeded .30 (see Table 1). These findings have important implications for the growing literature on self–other agreement, in that they establish that similarity is unlikely to represent a significant problem in ratings of personality and affectivity (see also Funder et al., 1995).

In contrast, we found strong evidence of assumed similarity in the ratings of many traits. Two findings are especially noteworthy. First, because of our unique three-sample design, we were able to establish that certain traits consistently display stronger evidence of assumed similarity than others. Indeed, the data reported in Table 3 establish that assumed similarity is a property with a very high degree of "cross-situational consistency," with cross-sample rs ranging from .66 to .78. Second, our results indicated that assumed similarity was highest for those traits that showed relatively low levels of self–other agreement. Many of the affective traits displayed particularly strong evidence of assumed similarity, with most scales yielding mean correlations in the .30 to .50 range. In contrast, assumed similarity correlations tended to be quite low for the Big Five scales; indeed, only 1 of the 15 coefficients exceeded .30.

Thus, our data support the suggestion that assumed similarity represents, in part, a rating strategy—or heuristic—that judges use in the absence of adequate trait-relevant information (Dawes, 1990; Funder et al., 1995). This further suggests that assumed similarity does not exert a substantial influence on the accuracy of trait ratings. It is not a source of accuracy because there is little actual similarity between dyadic members. It is not a source of inaccuracy because judges tend to avoid using this heuristic when they have ready access to valid, trait-relevant information. To test this idea directly, we recomputed the agreement correlations in all three samples, partialing out the influence of the judges' own self-rating (thereby eliminating the assumed similarity effect). As expected, these partial correlations tended to be quite similar to the
simple correlations presented in Table 2 (typically, they were .01 to .03 lower than the zero-order coefficients).

Although the basic phenomenon of assumed similarity was identified decades ago (Cronbach, 1955), it still has not been widely studied by trait psychologists. We believe that it warrants greater attention in the trait literature and that analyses of assumed similarity offer a useful complement to the more traditional focus on self–other agreement. For instance, we obtained low agreement correlations (ranging from .10 to .18) for ratings of Surprise across our three samples. How should these correlations be interpreted? In answering this question, one should consider that Surprise also generated the highest assumed similarity correlations in our data, with rs ranging from .51 to .56. Note, moreover, that the assumed similarity correlation was substantially higher than the agreement correlation in each sample. In other words, other-ratings of Surprise were a much better predictor of the judge’s own self-rating than that of the person he or she supposedly was rating. Putting all of this evidence together, the most parsimonious interpretation is that these judges lacked sufficient trait-relevant information to make accurate ratings. The fact that even long-married spouses showed the same pattern strongly suggests that it is impossible to rate others accurately on this dimension.

In a related vein, various researchers have argued that other-ratings offer an informative complement to the more typical self-ratings and, therefore, should play a much more prominent role in personality assessment (e.g., Funder, 1991; McCrae et al., 1998). Funder (1991) took this a step further, arguing that other-ratings represent the best single method of trait assessment. However, he added one important—and entirely appropriate—qualification, namely, that these judges must have had “ample opportunity to observe the target’s behavior in daily life” (p. 35, emphasis in original). If this is not the case (e.g., when a stranger rates a target on Neuroticism), then other-ratings obviously may be less valid than self-ratings; indeed, they may have little validity at all.

But how does one know whether judges have acquired sufficient information to provide valid, trustworthy ratings of the target? One possible approach would be to compare the agreement and assumed similarity correlations for a trait under different conditions, such as varying levels of acquaintanceship. It seems reasonable to propose that other-ratings should be viewed with caution if assumed similarity correlations far surpass the corresponding agreement correlations. For instance, the mean agreement correlations for the PANAS-X negative affectivity scales were .24 and .26 in the friendship and dating samples, respectively; the corresponding values for assumed similarity were much higher (.42 and .40, respectively). We therefore suspect that the other-raters lacked sufficient trait-relevant information to render highly valid judgments at this level of acquaintanceship. Accordingly, in cases of disagreement, it seems prudent to assign greater weight to the self-ratings than to the other-ratings. Clearly, however, the situation is different in the married sample: Here, the mean agreement correlation for these scales (.45) exceeds the average assumed similarity correlation (.38); at this level of acquaintanceship, it makes sense to assign equal or greater weight to the other-ratings. More generally, analyses of this sort can be used to determine the conditions necessary to yield highly valid other-judgments.

On the Validity of Trait Affectivity

Earlier, we suggested that affect researchers had largely ignored the issue of self–other agreement because they assumed it would be quite difficult for judges to rate targets with any degree of accuracy. Our data demonstrate that this concern was quite well founded. Table 2 indicates that most of the PANAS-X scales yielded only low to moderate levels of self–other convergence (with rs generally falling in the .15 to .35 range) in both the friendship and dating samples. Note, moreover, that the raters in both samples had known each other, on average, approximately 3 years. Under identical conditions, we obtained much greater agreement for the Big Five traits, with coefficients generally falling in the .35 to .55 range. It is difficult for judges to rate affective traits in others.

Nevertheless, our data also add significantly to a small body of evidence indicating that substantial levels of agreement can be attained under the right circumstances. Specifically, it is difficult to establish strong self–other convergence in affectivity unless one uses (a) multiple judges or (b) raters who are very highly acquainted with the targets. As an example of the former approach, Diener et al. (1995) obtained approximately four other-judgments per target from friends and family members. They reported a median self–other agreement correlation of .43 across eight different affect measures, with two coefficients surpassing .50 (.54 for Love, .53 for general Positive Affect). As an example of the latter approach, our analyses of the married couples yielded a median correlation of .44 across the 13 PANAS-X scales, with four coefficients surpassing .50 (see Table 2). Taken together, data such as these (see also Watson & Clark, 1991) clearly establish the convergent validity of trait affect ratings.

Two decades ago, it was widely assumed that self–other agreement generally was poor and that trait judgments had little basis in empirical reality (e.g., Shrauger and Schoeneman, 1979). Fortunately, as trait psychologists learned more about the processes (such as acquaintanceship and trait visibility) that influence agreement, they were able to create the necessary conditions that clearly established the reality of traits such as the Big Five. Although research in affectivity has lagged somewhat behind—and although the task has proven to be somewhat more difficult—we now have sufficient evidence to reach the same conclusion regarding the reality of affective traits.

References


**Received April 6, 1999**

**Revision received August 23, 1999**

**Accepted August 25, 1999**