

Personality Development in Emerging Adulthood: Integrating Evidence From Self-Ratings and Spouse Ratings

David Watson and John Humrichouse
University of Iowa

The authors examined self-ratings and spouse ratings in a young adult newlywed sample across a 2-year interval. Rank-order stability correlations were consistently high and did not differ across the 2 types of ratings. As expected, self-ratings showed significant increases in conscientiousness and agreeableness—and declines in neuroticism/negative affectivity—over time. Spouse ratings yielded a very different pattern, however, showing significant decreases in conscientiousness, agreeableness, extraversion, and openness across the study interval. Spouse ratings also showed evidence of a “honeymoon effect,” such that they tended to be more positive than self-ratings at Time 1. This effect had dissipated by the 2nd assessment; in fact, the spouse ratings tended to be more negative at Time 2. Analyses of individual-level change revealed little convergence between self- and spouse-rated change, using both raw change scores and reliable change index scores. Finally, correlational and regression analyses indicated that changes in spouse ratings were significantly associated with changes in marital satisfaction; in contrast, changes in self-ratings essentially were unrelated to marital satisfaction. These results highlight the value of collecting multimethod data in studies of adult personality development.

Keywords: trait stability, mean-level change, personality development, marital satisfaction

Classic models of trait psychology promoted a relatively static view of personality. These traditional models acknowledged that change was prevalent—indeed, even the norm—during childhood and adolescence. Once individuals reached adulthood, however, traits were viewed as essentially being set like plaster and highly resistant to change (see Costa & McCrae, 1994; McCrae & Costa, 1990; Srivastava, John, Gosling, & Potter, 2003). As evidence has accumulated, however, it has become clear that a simple “plaster” model fails to capture the complexities of personality development across the life span. In fact, recent findings have established that personality traits are not static constructs but rather show meaningful change well into middle age (Caspi, Roberts, & Shiner, 2005; Clark & Watson, 1999; Fraley & Roberts, 2004; Roberts, Robins, Trzesniewski, & Caspi, 2003; Roberts, Walton, & Viechtbauer, 2006).

Stability and Change in Adult Personality

Rank-Order Stability

In reviewing the prior literature on this topic, we consider three basic types of evidence. The first type of evidence concerns

rank-order stability, that is, the extent to which individuals maintain their relative position on the trait continuum over time. In other words, rank-order stability data establish the extent to which individual differences persist over time (e.g., whether a highly conscientious job applicant is likely to remain a highly conscientious employee several years after being hired). The data on this issue are highly consistent and yield several clear conclusions. First, stability correlations for personality traits are moderate to strong in magnitude, even when assessed in childhood and adolescence (Caspi et al., 2005; Roberts & DelVecchio, 2000).

Second, stability correlations decline in magnitude as the elapsed time interval increases (Caspi et al., 2005; Roberts & DelVecchio, 2000; Watson, 2004). This consistent pattern has helped to establish the existence of true change in personality, given that change is more and more likely to occur with increasing retest intervals (see Watson, 2004). It must be emphasized, however, that stability correlations never approach .00 and remain at least moderate in magnitude, even across intervals of several decades (Fraley & Roberts, 2004; see also Clark & Watson, 1999; Costa & McCrae, 1992).

Third, stability correlations for personality increase systematically with age (Caspi et al., 2005; Clark & Watson, 1999; Roberts & DelVecchio, 2000). Older models of trait development assumed that most personality change occurred prior to the age of 30, after which stability correlations should be uniformly high (for discussions, see McCrae & Costa, 1990; Roberts & DelVecchio, 2000; Srivastava et al., 2003). However, the meta-analytic findings of Roberts and DelVecchio (2000) revealed that stability coefficients for personality continue to increase well into middle age.

Fourth, stability estimates do not vary significantly by gender (Costa & McCrae, 1988; Roberts & DelVecchio, 2000; Schuerger, Zarrella, & Hotz, 1989). Indeed, in their meta-analytic review, Roberts and DelVecchio (2000) obtained identical population es-

David Watson and John Humrichouse, Department of Psychology, University of Iowa.

This research was supported by National Institute of Mental Health (NIMH) Grant 1-R01-MH61804-01 to Diane Berry, by NIMH Grant 1-R01-MH068472-1 to David Watson, and by NIMH Grant 1-R03-MH068395-01 to Eva C. Klohnen. We thank Alex Casillas, Elizabeth Gray, Malik Haig, Daniel Heller, Eva C. Klohnen, Shanhong Luo, Ericka Nus Simms, and all of the Iowa Marital Assessment Project staff for their help in the preparation of this article.

Correspondence concerning this article should be addressed to David Watson, Department of Psychology, E11 Seashore Hall, University of Iowa, Iowa City, IA 52242-1407. E-mail: david-watson@uiowa.edu

timates of overall trait stability in men and women (see their Table 4).

Finally, some trait measures consistently show greater stability than others (Watson, 2004). In this regard, it is important to emphasize that the traits comprising the influential five-factor model of personality—neuroticism, extraversion, openness, agreeableness, and conscientiousness—all show very similar levels of stability (see Roberts & DelVecchio, 2000, Table 5). In contrast, however, measures of trait affectivity consistently yield lower stability correlations (Watson, 2004). For instance, Vaidya, Gray, Haig, and Watson (2002) followed a large sample of young adults across an average retest interval of slightly more than 2.5 years. They obtained stability correlations ranging from .59 to .72 (mean $r = .64$) on the Big Five traits, which were assessed using the Big Five Inventory (BFI; John & Srivastava, 1999); in contrast, trait affectivity scales from the Positive and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988) yielded significantly lower stability correlations of .51 (Positive Affect scale) and .49 (Negative Affect scale). This pattern of differential stability was replicated in another sample that completed these same measures twice across a 2-month interval (Watson, 2004). It is particularly striking that BFI Neuroticism showed significantly greater stability than the PANAS Negative Affect scale in both samples, despite the fact that the two scales assess very similar content and were strongly interrelated. We return to this issue of differential stability later.

Mean-Level Change

The second line of evidence concerns mean-level change, that is, whether or not the average levels of a trait change systematically with age. These data clarify how personality evolves over time as a function of changing biological and/or environmental conditions. Once again, recent reviews of this evidence have yielded several clear conclusions (see Caspi et al., 2005; Clark & Watson, 1999; Roberts, Robins, et al., 2003; Roberts et al., 2006). The evidence is especially consistent for the three traits—neuroticism, agreeableness, and conscientiousness—that define the “Alpha” superfactor identified by Digman (1997; see also Markon, Krueger, & Watson, 2005). First, trait levels of neuroticism and negative affectivity show a significant decline with age. The bulk of this decline appears to occur in adolescence and early adulthood (Caspi et al., 2005; Roberts, Robins, et al., 2003); nevertheless, decreases continue to be seen later in life (Clark & Watson, 1999; Roberts et al., 2006; Srivastava et al., 2003). Second, levels of both agreeableness and conscientiousness increase across young adulthood and middle age (Roberts, Robins et al., 2003; Roberts et al., 2006).

In contrast, the two remaining Big Five traits show more complex patterns. The data for extraversion and positive emotionality have been particularly inconsistent (Clark & Watson, 1999; Vaidya et al., 2002). However, Roberts, Robins, et al. (2003) found that the evidence was more consistent at the specific trait level: Measures of dominance tended to increase from adolescence through early middle age, whereas levels of sociability increased during adolescence but then declined starting in young adulthood (see also Roberts et al., 2006). Finally, openness to experience tends to show a curvilinear pattern, exhibiting increases in adolescence and young adulthood (particularly for those individuals who

remain in school), but then declining later in life (see Caspi et al., 2005; Roberts et al., 2006).

Taken together, these data demonstrate that mean-level changes are both meaningful and highly systematic across the life span (Caspi et al., 2005; Roberts, Caspi, & Moffitt, 2001). Indeed, Caspi et al. (2005) concluded that they reflect the influence of a *maturity principle*, arguing that these mean-level shifts “point to increasing psychological maturity over development, from adolescence to middle age” (p. 468).

Finally, it is clear that the pace of change is not constant across the life span. Interestingly, the data indicate that much of this change occurs in young adulthood, a period roughly corresponding to the 20s (Arnett, 2000; Roberts, Caspi, & Moffitt, 2003). This period has been characterized as *demographically dense*, given that it typically involves more changes in identity and major life roles (e.g., the onset of marriage and parenthood; the initiation of one’s career) than any other period in life (Arnett, 2000; Caspi et al., 2005). In fact, the level of change during this period even exceeds that seen during adolescence (Caspi et al., 2005).

Individual-Level Change

The third line of evidence is individual-level change. Several authors recently have emphasized the importance of examining change at the individual level (see Roberts et al., 2001; Roberts, Robins, et al., 2003; Robins, Fraley, Roberts, & Trzesniewski, 2001; Vaidya et al., 2002); these analyses help to clarify the specific processes that are associated with personality development across the life span. Roberts, Caspi, and Moffitt (2003), for example, demonstrated that changes in personality at the individual level are meaningful and can be systematically linked to ongoing life experiences. More specifically, they showed that several aspects of work experience—including occupational attainment, work satisfaction, work involvement, and financial security—were significantly associated with individual-level changes in personality between the ages of 18 and 26 years.

Beyond Self-Report: Evidence From Other Methods

Although the evidence we have considered thus far is impressive, it is limited in a number of ways (see Caspi et al., 2005). A particularly important limitation is that most of the data are based on self-ratings. In this regard, several recent authors have bemoaned the overreliance on self-report in this area and have emphasized the importance of examining stability versus change across multiple methods (e.g., McCrae et al., 2004; McCrae, Terraciano, & Members of the Personality Profiles of Cultures Project, 2005; Roberts, Caspi, & Moffitt, 2003; Roberts, O’Donnell, & Robins, 2004).

Before reaching any fundamental conclusions about the nature of personality development, it obviously is important to examine other types of evidence. What do non-self-report data show? First, the available evidence indicates that rank-order stability coefficients essentially are invariant across methods (Caspi et al., 2005; Costa & McCrae, 1988; Roberts & DelVecchio, 2000). Most notably, in their meta-analytic review of the literature, Roberts and DelVecchio (2000) obtained virtually identical population estimates of overall trait stability across self-report ($\rho = .52$) and observer-rated ($\rho = .48$) data (see their Table 4).

The data regarding mean-level change are both less plentiful and less consistent. Generally speaking, analyses based on observer ratings tend to show patterns that parallel those seen with self-report—that is, age-related declines in neuroticism, extraversion, and openness, and increases in agreeableness and conscientiousness—but that are weaker in magnitude (McCrae et al., 2004, 2005). McCrae et al. (2004), for instance, examined age-related changes in peer ratings collected in Russia and the Czech Republic. They concluded that the observer-based effects were weaker and less consistent but that “whenever significant age differences in observer-rated personality traits appear, they follow the same direction as self-reports” (p. 155).

The bulk of the available observer-rating data comes from cross-sectional designs in which personality scores are correlated with age (e.g., McCrae et al., 2004, 2005). Accordingly, we currently know very little about individual-level change using methods other than self-report. To date, the most comprehensive analysis of this issue was reported by McCrae (1993, Study 3), who examined longitudinal changes in the Big Five across a 6- to 7-year retest interval. McCrae began by establishing that personality changes showed some evidence of within-method consistency: For instance, individuals who indicated that they decreased on one facet of neuroticism (e.g., anxiety) also reported parallel changes on other facets within this domain (e.g., depression, hostility). Thus, these initial results suggest that these self-rated changes are systematic and do not simply reflect random measurement error.

Subsequent analyses, however, revealed that these self-rated changes failed to show substantial convergence with spouse and peer ratings obtained across the same time interval. For instance, McCrae (1993) computed raw change scores for the self-ratings and spouse ratings on each trait. These analyses revealed a significant association between self- versus spouse-rated change on neuroticism, but the correlation was only .16; furthermore, the corresponding coefficients for extraversion and openness were even weaker and nonsignificant. It should be noted, however, that these results are based on relatively small sample sizes (the largest sample size in these analyses is only 135). Consequently, these results need to be replicated in a larger sample before any firm conclusions can be drawn.

The Current Study

The current study adds to this growing literature by examining personality development in a large sample ($N = 460$) of young adults across a time span of approximately 2 years. We included measures of the Big Five and trait affectivity at both assessments, which allowed us to examine the three critical issues of (a) rank-order stability, (b) mean-level change, and (c) individual-level change.

We highlight two unusual features of our study that represent significant extensions of the prior literature. First, most previous longitudinal studies of “emerging adulthood” have focused on the college and immediate postcollege years—that is, the period between the ages of 18 and 26 years (e.g., Roberts et al., 2001; Roberts, Caspi, & Moffitt, 2003; Robins et al., 2001; Vaidya et al., 2002). In contrast, our sample consisted of newlywed couples who were somewhat older than this. Specifically, the mean age of our sample at the Time 1 assessment was approximately 28 years (see

Watson et al., 2004, Table 1). Thus, we are able to examine issues related to personality stability versus change in a sample that is beginning to undergo the transition from emerging adulthood to full adulthood.

Second, as discussed earlier, many recent authors have emphasized the importance of moving beyond self-report measures in studying issues related to personality development across the life span (McCrae et al., 2004; Roberts, Caspi, & Moffitt, 2003; Roberts et al., 2004). Because our sample consisted of newlywed couples, we were able to obtain spouse ratings of personality and trait affectivity for most of our participants ($n = 301$) at both assessments. Consequently, our data allow us to examine issues such as mean-level change and individual-level change in a large sample across two different methods.

On the basis of previous evidence, we had two basic expectations concerning rank-order stability. First, based on the meta-analytic results of Roberts and DelVecchio (2000), we predicted that self-ratings and spouse ratings would produce very similar stability correlations across this 2-year interval. Second, because stability correlations tend to increase with age (Caspi et al., 2005; Clark & Watson, 1999; Roberts & DelVecchio, 2000), we expected these coefficients to be somewhat higher than those reported in previous longitudinal studies of young adulthood (e.g., Roberts et al., 2001; Robins et al., 2001; Vaidya et al., 2002).

We also made two basic predictions related to our analyses of mean-level change. First, we expected our young adult sample to display significant (a) increases in conscientiousness and agreeableness and (b) decreases in neuroticism and negative affectivity across this 2-year interval (Caspi et al., 2005; Clark & Watson, 1999; Roberts, Robins, et al., 2003). In contrast, the findings related to extraversion and openness are more complex and inconsistent (see Roberts et al., 2006). Accordingly, we made no specific predictions regarding these traits. Second, on the basis of earlier results (McCrae et al., 2004, 2005), we expected the self-ratings and spouse ratings to show the same basic patterns over time; however, we expected the magnitude of these age-related changes to be greater in the former than in the latter (i.e., we predicted weaker increases in agreeableness and conscientiousness and smaller decreases in neuroticism and negative affectivity in the spouse ratings).

Finally, as discussed earlier, McCrae (1993) found weak convergence between self-rated versus spouse-rated change at the individual level. On the basis of these results, we expected to obtain significant (because of our larger sample size) but low correlations between self- and spouse-rated change scores on our measures of personality and trait affectivity.

Method

Participants and Procedure

Time 1 assessment. The original sample consisted of 291 married couples who participated in the Iowa Marital Assessment Project (IMAP; for more details regarding IMAP, see Watson et al., 2004). IMAP staff members identified recently married couples from the records of Johnson County and Linn County in eastern Iowa. Couples who met the inclusion criteria for the study (which required that they had been married less than a year at the time of initial contact and that both members of the couple were age 50 or younger) then were sent a letter inviting them to participate. At the time of assessment, the couples had been married an average of

153.9 days (range = 25 to 452 days), that is, approximately 5 months. They indicated that they had known each other an average of 4.69 years (range = less than a year to 30 years) and had begun dating approximately 3.5 years earlier ($M = 3.54$ years; range = less than a year to 15 years).

All participants were assessed in small-group sessions involving from one to three couples. These sessions were conducted from June 2001 through December 2001. The sessions typically lasted from 2 to 2.5 hr and included a battery of self-report measures, spouse ratings, and intelligence testing (see Watson et al., 2004). The couples were compensated \$120 for their participation. To ensure honest and independent responding, each participant sat quietly at a separate desk when completing the self-ratings and spouse ratings. Because of missing data, complete responses were available from a total of 574 participants (98.6%) at this initial assessment.

Time 2 assessment. The Time 2 assessment was conducted in two separate phases. In the first phase, we attempted to contact all of the original IMAP participants, who were invited to return to our laboratory for another small-group session; these Time 2 sessions involved a maximum of six couples. These sessions typically lasted approximately 2 hr and again included a battery of self-report measures and spouse ratings. Each individual was compensated \$50 for his or her participation. These small-group laboratory sessions were conducted between July 2003 and November 2003, that is, approximately 2 years after the original Time 1 assessment.

In most cases, the participants again were assessed as couples; in our invitation letter, however, we emphasized that they were welcome to participate as single individuals (a) if they were separated or divorced from their spouse or (b) if their spouse was unwilling to return for this assessment for any reason. A total of 314 individuals participated in these Time 2 laboratory sessions; this included 154 intact couples and 6 single women. Because of missing responses (including missing Time 1 responses in some instances), we had complete data on 301 of these participants (152 men, 149 women).

Many of the original IMAP participants expressed an interest in being involved in the Time 2 assessment but indicated that they were unable to return to our laboratory for various reasons (most often because they had moved out of eastern Iowa). We therefore initiated the second phase of the Time 2 assessment, in which participants were sent a battery of questionnaires by mail. Because we could not ensure that spouses would complete their ratings independently of one another, these questionnaires were restricted to self-report and did not ask the respondents to provide any spouse ratings. Each participant was compensated \$25 for his or her responses, which were returned to us by mail. We received mailed responses from an additional 159 participants (75 men, 84 women). The bulk of these mailed questionnaires were completed between December 2003 and January 2004 ($n = 146$); however, we received 13 additional questionnaires between February and April 2004. Thus, these participants completed the Time 2 assessment roughly 2 to 2.5 years after Time 1.

Overall, between these two phases, we obtained Time 2 responses from a total of 460 participants (227 men, 233 women). This represents 80.1% of the 574 participants with complete Time 1 data. At Time 1, the men were a mean age of 29.1 years ($SD = 6.5$ years, range = 19–49) and the women were a mean age of 27.3 years ($SD = 6.2$ years, range = 18–50).

Measures

BFI. We used the BFI (Benet-Martinez & John, 1998; John & Srivastava, 1999) to assess the traits comprising the five-factor model. The BFI contains 8-item scales assessing Neuroticism and Extraversion, a 10-item Openness scale, and 9-item measures of Agreeableness and Conscientiousness. In the self-rating version, participants were asked to indicate “the extent to which you agree or disagree” with each item on a 5-point scale ranging from *disagree strongly* to *strongly agree*. The format for the spouse ratings was identical, except that the respondents were asked to “consider the feelings, behaviors, and preferences of your spouse” and then to evaluate the extent to which each item characterized the personality of their partner (again using the same 5-point *disagree–agree* scale). In the

Time 1 assessment, the BFI scales had coefficient alphas ranging from .78 (Agreeableness) to .88 (Neuroticism) in the self-ratings and from .83 (Conscientiousness) to .88 (Neuroticism) in the spouse ratings (see Watson et al., 2004).

PANAS. We assessed affectivity using the trait form of the PANAS. The PANAS includes 10-item scales assessing the general dimensions of Negative Affect (e.g., *nervous, upset, irritable, ashamed, scared*) and Positive Affect (e.g., *enthusiastic, active, interested, proud, determined*). Self-raters were asked to indicate on a 5-point scale (ranging from *very slightly or not at all* to *extremely*) “to what extent you generally feel this way, that is, how you feel on average.” The format and instructions for the spouse ratings were identical, except that respondents were told to rate “to what extent your spouse generally feels or acts this way, that is, how your spouse feels or acts on the average.” In the Time 1 assessment, the Negative Affect scale had coefficient alphas of .89 and .88 in the self-ratings and spouse ratings, respectively; parallel values for the Positive Affect scale were .85 and .87, respectively (see Watson et al., 2004).

Results

Preliminary Analyses

Attrition analyses. Before turning to our main results, we report several other analyses to explicate basic aspects of our data. First, we conducted two series of analyses to evaluate the representativeness of our Time 2 subsamples. In the first series, we compared the Time 1 BFI and PANAS scores of our Time 2 retest participants (RP; $n = 460$) with those of the nonparticipants (NP; $n = 114$). Across the 14 individual analyses (i.e., seven traits assessed using both self-ratings and spouse ratings), we obtained four significant differences. Specifically, the RP respondents rated themselves as more conscientious, RP $M = 34.45$, $SD = 5.72$; NP $M = 32.93$, $SD = 6.51$; $t(572) = 2.46$, $p < .05$, more agreeable, RP $M = 35.58$, $SD = 5.39$; NP $M = 34.47$, $SD = 4.90$; $t(572) = 2.02$, $p < .05$, and less open, RP $M = 38.58$, $SD = 6.11$; NP $M = 39.90$, $SD = 5.55$; $t(572) = -2.11$, $p < .05$, than the NP group. Interestingly, only one of these differences was replicated in the spouse ratings; specifically, the spouses rated the RP group as more conscientious than the NP group, RP $M = 35.17$, $SD = 6.17$; NP $M = 33.52$, $SD = 6.82$; $t(572) = 2.51$, $p < .05$. Thus, replicating a pattern observed in previous longitudinal studies (e.g., Vaidya et al, 2002), we obtained consistent evidence that the Time 2 retest participants were significantly more conscientious than the nonparticipants.

Second, we compared the Time 1 BFI and PANAS scores of the Time 2 laboratory (LAB; $n = 301$) and mail-out (MAIL; $n = 159$) subsamples. Across the 14 individual analyses, we obtained only one significant difference: Spouses in the laboratory subsample rated their partners as higher in trait Negative Affect than did those in the mail-out group, LAB $M = 19.48$, $SD = 6.26$, MAIL $M = 18.14$, $SD = 6.31$; $t(458) = 2.18$, $p < .05$. It is noteworthy, however, that this difference did not approach significance in the self-ratings, LAB $M = 19.04$, $SD = 7.11$; MAIL $M = 18.35$, $SD = 6.44$; $t(458) = 1.06$, *ns*. On the basis of this evidence, it appears that our two Time 2 subsamples are very similar in terms of their personality characteristics; this result is not surprising, given that these groups emerged primarily for pragmatic reasons (i.e., most of the participants in the MAIL group had moved out of the area).

Spousal similarity. Spousal similarity is a potentially important consideration in studies of married couples because it produces statistical nonindependence in data analyzed at the individ-

ual level (Kashy & Snyder, 1995; Kenny, 1995). 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. However, studies consistently have found very little evidence of spousal similarity on a wide range of personality traits (see Watson, Hubbard, & Wiese, 2000b; Watson et al., 2004). Consistent with this broader trend, analyses of the Time 1 IMAP data yielded spousal similarity correlations ranging between $-.17$ and $.18$ (self-ratings) and between $-.14$ and $.13$ (spouse ratings) on the BFI and PANAS scales. Because nonindependence is not a substantial problem in these data, we report our main analyses of personality at the individual level. However, to examine possible gender differences in our data, we also report many of our key results separately for men and women.

Subsequently, we examine the relations between personality ratings and marital satisfaction. Not surprisingly, the husband's and wife's marital satisfaction were substantially related at both Time 1 ($r = .27, n = 289, p < .01$) and Time 2 ($r = .47, n = 144, p < .01$). Accordingly, for all analyses involving marital satisfaction, we do not present any findings on our overall sample but report separate results only for the wives and husbands.

Self-spouse agreement. Finally, to examine the convergent validity of our trait measures, we computed self-spouse agreement correlations at both Time 1 and Time 2; these analyses were based on the 301 participants with complete data at both assessments. Consistent with previous analyses of married couples (Watson et al., 2000b), the Big Five traits showed strong self-spouse convergence at both assessments: Specifically, these agreement correlations ranged from $.43$ (Agreeableness) to $.62$ (Neuroticism; mean $r = .53$) at Time 1, and from $.40$ (Openness) to $.64$ (Extraversion; mean $r = .52$) at Time 2.

Replicating previous research in this area (see Watson et al., 2000b), the PANAS scales showed more moderate—but still significant—agreement at both assessments. As would be expected with relatively low visibility traits (which are particularly susceptible to acquaintanceship effects; see Watson et al., 2000b), these affective scales tended to show better agreement at Time 2 ($r_s =$

$.39$ and $.38$ for Negative Affect and Positive Affect, respectively) than at Time 1 ($r_s = .32$ and $.27$, respectively). These agreement correlations are reported in greater detail in Humrichouse and Watson (2006).

Rank-Order Stability

Basic findings. Table 1 presents rank-order stability correlations for both the self-ratings and the spouse ratings across the 2-year study interval. The table displays stability coefficients computed in the overall sample, as well as separately for women and men; these latter results are organized by the target of assessment (i.e., spouse ratings for women represent the husbands' ratings of their wives, whereas the spouse ratings for men reflect the wives' ratings of their husbands).

Several aspects of these data are noteworthy. First, these correlations are consistently high, ranging from $.67$ (PANAS Positive Affect in the spouse ratings) to $.83$ (BFI Extraversion in the spouse ratings) in the overall sample. Furthermore, consistent with previous investigations of this issue (Costa & McCrae, 1988; Roberts & DelVecchio, 2000; Schuerger et al., 1989), the stability coefficients did not differ substantially by gender. In fact, only 1 of the 14 individual comparisons yielded a statistically significant sex difference: Self-rated Openness was more stable in women ($r = .83$) than in men ($r = .68; z = 3.84, p < .01$). We therefore restrict further discussion to results based on the overall sample.

Second, as predicted, the self-ratings and spouse ratings yielded very similar stability correlations; indeed, in every case the stability correlations for a given scale differed by no more than $|.04|$ from each other. Overall, the mean stability correlations for the two sets of ratings were virtually identical for both the BFI (mean $r_s = .78$ and $.77$ in the self-ratings and spouse ratings, respectively) and the PANAS (mean $r_s = .69$ in both the self-ratings and spouse ratings).

Third, we predicted that the 2-year stability correlations in this study would be significantly higher than those reported in younger adults. We tested this prediction by comparing these stability correlations to those reported by Vaidya et al. (2002) in a some-

Table 1
Rank-Order Stability Correlations Across the 2-Year Interval

Scale	Self-rating			Spouse rating		
	Overall	Women	Men	Overall	Women	Men
BFI						
Neuroticism	.79	.78	.75	.77	.76	.69
Extraversion	.82	.83	.81	.83	.82	.85
Openness	.76	.83	.68	.78	.75	.80
Agreeableness	.75	.72	.77	.71	.65	.76
Conscientiousness	.76	.77	.74	.76	.80	.71
<i>M</i>	.78	.79	.75	.77	.76	.77
PANAS						
Negative Affect	.68	.67	.70	.72	.68	.66
Positive Affect	.69	.71	.66	.67	.68	.66
<i>M</i>	.69	.69	.68	.69	.68	.66

Note. Data are organized by target (e.g., spouse-rating correlations for women represent the husbands' ratings of their wives). For self-ratings, $N = 460, n = 233$, women; $n = 227$, men. For spouse ratings, $N = 301, n = 149$, women; $n = 152$, men. All correlations are significant at $p < .01$. BFI = Big Five Inventory; PANAS = Positive and Negative Affect Schedule.

what younger sample (mean age = 21 years at Time 2). These comparisons strongly confirmed our prediction. Vaidya et al. (see their Table 6) reported 2.5-year stability correlations ranging from .59 to .72 on the BFI (mean $r = .64$), and correlations of .49 (Negative Affect) and .51 (Positive Affect) on the PANAS. Follow-up tests indicated that all seven scales produced significantly higher stability correlations in the current study; this was true for both the self-ratings (z s ranged from 3.21 to 5.26; all p s < .01) and the spouse ratings (z s ranged from 2.73 to 4.84; all p s < .01). Thus, our data again show that stability correlations for personality increase systematically with age (see also Caspi et al., 2005; Roberts & DelVecchio, 2000).

Fourth, our results again indicate that the BFI scales tend to show higher stability correlations than the PANAS (see Vaidya et al., 2002; Watson, 2004). Replicating the results of earlier studies, BFI Extraversion was significantly more stable than PANAS Positive Affect in both the self-ratings (.82 vs. .69; $z = 4.70, p < .01$) and the spouse ratings (.83 vs. .67; $z = 4.81, p < .01$), despite the fact that these scales consistently were strongly correlated (for the self-ratings, $r = .52$ and $.57$ at Time 1 and Time 2, respectively; for the spouse ratings, $r = .55$ and $.53$, respectively). Similarly, BFI Neuroticism had higher stability correlations than the PANAS Negative Affect scale in both sets of ratings; this difference was significant in the self-ratings (.79 vs. .68; $z = 3.70, p < .01$), but not in the spouse ratings (.77 vs. .72; $z = 1.35, ns$). This difference is particularly striking given the very high correlations between these scales in both the self-ratings (r s = .64 and .67 at Time 1 and Time 2, respectively) and the spouse ratings (r s = .72 and .79 at Time 1 and Time 2, respectively). It is important to note, moreover, that the BFI Neuroticism and PANAS Negative Affect scales both are overwhelmingly affective in nature and contain very similar item content (see Watson, 2004, Table 5); taken together with previous findings, these results illustrate the importance of wording, format, and instructional effects on temporal stability (see Watson, 2004).

Moderator analyses of age. As noted earlier, our data are consistent with previous evidence indicating that stability correlations increase with age (Roberts & DelVecchio, 2000). Given that our participants varied widely in age (range = 18–50 years at

Time 1), this raises the further possibility that the level of rank-order stability was substantially higher in our older respondents. We examined this issue in a series of moderated multiple regression analyses, using the Time 2 scores as criteria. We entered the two main effects (i.e., age and the corresponding Time 1 trait score) as predictors in Step 1, followed by the centered interaction term in Step 2.

Across the 14 individual analyses, we found four significant moderators. Only Agreeableness yielded a replicable interaction effect across both the self-ratings ($\beta = .070, \Delta R^2 = .005, p < .05$) and the spouse ratings ($\beta = .109, \Delta R^2 = .012, p < .01$). In addition, Negative Affect ($\beta = .070, \Delta R^2 = .006, p < .05$) and Openness ($\beta = .093, \Delta R^2 = .008, p < .05$) displayed significant moderator effects in the self-ratings and spouse ratings, respectively. Thus, we obtained some scattered evidence indicating that stability correlations were somewhat higher in our older participants.

Mean-Level Change

Analyses of the self-ratings. We turn now to the issue of mean-level change. Table 2 reports the mean Time 1 and Time 2 scores for each scale in the self-ratings, together with an index (Cohen's d) that quantifies the magnitude of the difference between them. We again report results based on the overall sample, as well as separately for men and women.

These results strongly support our predictions. As expected, paired t tests revealed significant increases in both Conscientiousness ($d = .33$) and Agreeableness ($d = .12$), as well as significant declines in Neuroticism ($d = -.19$) and Negative Affect ($d = -.11$) in the overall sample. Conscientiousness showed the largest amount of change; in fact, it displayed significant increases in both women ($d = .31$) and men ($d = .34$). In contrast, our data did not reveal any significant mean-level changes in Extraversion, Openness, and Positive Affect across the 2-year study interval. Thus, consistent with previous work in this area, our findings demonstrate systematic temporal changes in personality that essentially are confined to the three trait domains comprising the Alpha superfactor (Markon et al., 2005). Moreover, these results again

Table 2
Mean-Level Changes in Self-Rated Personality Over the 2-Year Interval

Scale	Overall					Women					Men				
	Time 1		Time 2		d	Time 1		Time 2		d	Time 1		Time 2		d
	M	SD	M	SD		M	SD	M	SD		M	SD	M	SD	
BFI															
Neuroticism	22.2	7.1	21.3	6.8	-.19**	24.8	6.9	23.4	6.7	-.31**	19.5	6.2	19.2	6.2	-.06
Extraversion	27.9	6.6	27.7	6.7	-.06	28.1	6.7	27.9	6.7	-.05	27.7	6.5	27.4	6.6	-.07
Openness	38.6	6.1	38.2	6.4	-.09	38.3	6.3	37.9	6.5	-.11	38.9	5.9	38.5	6.2	-.08
Agreeableness	35.6	5.4	36.0	5.4	.12*	36.2	5.1	36.7	5.2	.15*	35.0	5.6	35.3	5.4	.09
Conscientiousness	34.4	5.7	35.7	5.3	.33**	35.4	5.5	36.6	5.4	.31**	33.5	5.8	34.8	5.1	.34**
PANAS															
Negative Affect	18.6	6.7	18.0	6.3	-.11*	18.7	7.0	18.1	6.0	-.11	18.5	6.3	17.9	6.6	-.12
Positive Affect	38.1	5.7	37.8	6.2	-.07	37.7	5.8	37.3	6.6	-.10	38.6	5.5	38.4	5.8	-.05

Note. $N = 460, n = 233$, women; $n = 227$, men. BFI = Big Five Inventory; PANAS = Positive and Negative Affect Schedule.
* $p < .05$. ** $p < .01$.

show that these changes are positive in nature, such that adult personality development is characterized by increasing psychological maturity with age (see Caspi et al., 2005).

Analyses of the spouse ratings. Table 3 presents parallel results for the spouse ratings, again computed in the overall sample, as well as separately for women and men; as in Table 1, the latter results are organized by target (i.e., the ratings for women represent the husbands' ratings of their wives, whereas the ratings for men reflect the wives' ratings of their husbands).

We predicted that these data would show the same basic pattern as the self-ratings, albeit in attenuated form; that is, we expected smaller increases in agreeableness and conscientiousness, and weaker declines in neuroticism and negative affectivity. These predictions obviously were not supported. Indeed, the spouse ratings show an entirely different pattern from the self-reports. Looking at the results from the overall sample, the most striking aspect of these data is that they showed a significant *decline* in both Agreeableness ($d = -.28$) and Conscientiousness ($d = -.22$), which is completely opposite to the trend exhibited in the self-ratings. Agreeableness displayed the largest overall level of change and exhibited significant decreases in both women ($d = -.24$) and men ($d = -.32$). Moreover, BFI Neuroticism ($d = .07$) and PANAS Negative Affect ($d = .09$) both showed small, non-significant *increases* in these data. Furthermore, Extraversion ($d = -.20$) and Openness ($d = -.18$)—which did not exhibit significant changes in the self-reports—both showed significant declines in the overall sample. In fact, across the seven trait scales, the only consistent finding was that PANAS Positive Affect displayed small, nonsignificant decreases in both the self-ratings ($d = -.07$) and the spouse ratings ($d = -.07$).

Comparisons of self-ratings and spouse ratings. Clearly, the spouse ratings paint a very different picture of adult personality development, and—unlike the self-reports—they certainly do not suggest that our participants were exhibiting greater psychological maturity over time. These discrepant findings are troubling, and they raise a basic conceptual/interpretative issue: How can these results be reconciled into a coherent model of adult personality development?

One way to begin to address this question is to compare the mean self-rating and spouse rating scores at each of the two

assessments. We therefore computed paired *t* tests to examine whether the average self-ratings and spouse ratings differed significantly from each other at Times 1 and 2; these analyses were based on the 301 participants with complete data at both assessments. Analyses of the Time 1 data revealed evidence of a significant honeymoon effect in these newlywed couples; that is, the mean spouse ratings were significantly greater than the average self-ratings for Agreeableness ($d = .33, p < .01$), Extraversion ($d = .26, p < .01$), and Conscientiousness ($d = .13, p < .05$). Thus, at Time 1, spouses were rating their partners more positively than the partners were rating themselves.

It is noteworthy that only one of these effects persisted at Time 2: The average spouse rating on Extraversion still was significantly higher than the mean self-rating ($d = .17, p < .01$). In contrast, all of the other significant findings suggested that the spouse judgments now were more negative than the self-ratings at Time 2. Specifically, the spouse ratings now were significantly higher than the self-ratings on both Neuroticism ($d = .15, p < .01$) and Negative Affect ($d = .13, p < .05$) and significantly lower on Conscientiousness ($d = -.25, p < .01$) and Openness ($d = -.23, p < .01$). These results clearly suggest that the honeymoon was over at Time 2, so that the spouses now were rating their partners more harshly at this second assessment. We revisit this issue later.

Individual-Level Change

Change score analyses. The self-ratings and spouse ratings obviously displayed very different patterns of mean-level change. Nevertheless, it still is possible that our participants agreed about which specific individuals showed the largest relative increases and decreases on each trait. For example, if a wife reported a very large increase in her level of conscientiousness, it is plausible to suggest that her husband noticed this marked change and also judged her level of the trait to be higher at Time 2. Consequently, it also is important to examine change at the individual level.

We began by computing separate change scores (subtracting the Time 1 score from the Time 2 score) for each scale in the self-ratings and spouse ratings. Before assessing self-spouse convergence, it first is important to establish that these measures assess systematic and meaningful variance, given understandable

Table 3
Mean-Level Changes in Spouse-Rated Personality Over the 2-Year Interval

Scale	Overall					Women					Men				
	Time 1		Time 2		<i>d</i>	Time 1		Time 2		<i>d</i>	Time 1		Time 2		<i>d</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
BFI															
Neuroticism	21.5	7.0	21.8	7.6	.07	24.6	6.3	24.7	7.4	.03	18.5	6.4	19.0	6.8	.11
Extraversion	29.1	6.3	28.4	6.6	-.20**	29.5	5.9	28.6	6.2	-.25**	28.8	6.7	28.2	6.9	-.16
Openness	37.8	6.8	37.0	7.0	-.18**	37.7	6.2	36.4	6.7	-.28**	37.9	7.3	37.5	7.3	-.09
Agreeableness	37.7	5.6	36.4	6.4	-.28**	37.6	5.2	36.5	5.9	-.24**	37.7	5.9	36.3	6.8	-.32**
Conscientiousness	35.1	6.3	34.2	6.6	-.22**	36.0	6.5	35.5	6.5	-.11	34.3	6.0	32.8	6.4	-.31**
PANAS															
Negative Affect	18.1	6.3	18.6	7.0	.09	20.8	6.7	21.2	7.5	.07	15.5	4.6	16.1	5.6	.13
Positive Affect	38.4	5.7	38.1	6.1	-.07	38.5	5.4	38.0	5.9	-.11	38.4	5.9	38.2	6.4	-.03

Note. *N* = 301, *n* = 149, women; *n* = 152, men. BFI = Big Five Inventory; PANAS = Positive and Negative Affect Schedule.
** *p* < .01.

concerns about the unreliability of change scores (see Asendorpf, 1992; McCrae, 1993). As discussed by McCrae (1993), one interesting way to investigate this issue is to examine relations among change scores computed within the same method. If these variables are psychologically meaningful and tap systematic variance, then we should observe significant associations between change scores computed on closely related traits. For instance, we should see significant correlations between self-rated changes on the BFI Neuroticism and PANAS Negative Affect scales.

Accordingly, Table 4 reports the within-method correlations between the change scores in both the self-ratings (below the diagonal) and the spouse ratings (above the diagonal). The most noteworthy aspect of these data is that change scores on related traits were, in fact, significantly correlated with one another. For example, the change scores for BFI Neuroticism and PANAS Negative Affect were moderately correlated in both the self-ratings ($r = .42$) and the spouse ratings ($r = .44$). Similarly, these analyses revealed significant correlations between changes on Extraversion and Positive Affect ($r = .37$ and $.32$ in the self-ratings and spouse ratings, respectively) and between changes on Agreeableness and Conscientiousness ($r = .26$ and $.33$, respectively). These results are reassuring, as they strongly suggest that these change scores tap meaningful psychological variance and do not simply reflect random measurement error.

With that in mind, we now consider the convergence between self- and spouse-rated changes in personality. Based on the results of McCrae (1993), we predicted low, but significant, correlations between these two sets of scores. Table 5 reports these coefficients, both in the overall sample and separately by gender. The Table 5 data offer mixed support for our prediction. As expected, the correlations consistently were low. However, only two of them were significant in the overall sample—those for Conscientiousness ($r = .20$) and Neuroticism ($r = .19$); it is noteworthy, moreover, that neither of these effects replicated across both men and women. In contrast, the coefficients for the five remaining traits all were quite weak, ranging from only $-.01$ (Openness) to $.08$ (Extraversion) in the overall sample. Thus, consistent with our analyses of mean-level change, these data reveal little convergence across our two types of ratings.

Analyses of reliable change index (RCI) scores. Individual scores may randomly fluctuate from one testing to the next simply as a function of measurement error. Accordingly, analyses of raw change scores may be largely tapping random fluctuations over

Table 5
Convergent Correlations Between Self- Versus Spouse-Rated Change Scores

Scale	Overall	Women	Men
BFI			
Neuroticism	.19*	.07	.29**
Extraversion	.08	.11	.06
Openness	-.01	.07	-.07
Agreeableness	.04	.04	.04
Conscientiousness	.20**	.35**	.08
<i>M</i>	.10	.13	.08
PANAS			
Negative Affect	.06	-.01	.16
Positive Affect	.07	.10	.04
<i>M</i>	.06	.04	.10

Note. $N = 301$ (overall), 149 (women), 152 (men). BFI = Big Five Inventory; PANAS = Positive and Negative Affect Schedule.

* $p < .05$. ** $p < .01$.

time rather than true, meaningful change (Roberts et al., 2001; Robins et al., 2001; Vaidya et al., 2002). Consequently, several recent studies have reported analyses based on RCI scores, which allow one to determine how many individuals showed a statistically significant amount of change versus how many essentially stayed the same. We computed RCI scores on our 301 participants with complete data at both assessments (see Roberts et al., 2001, for details regarding the calculation of RCI scores). To calculate the standard error of measurement for each scale, we used (a) its average standard deviation across the two assessments and (b) retest reliability coefficients computed across a 2-week interval in a sample of 446 University of Iowa students (Chmielewski & Watson, 2006). As in previous studies (e.g., Roberts et al., 2001; Robins et al., 2001; Vaidya et al., 2002), we classified individuals as having changed significantly from one assessment to the next if the probability associated with that person's RCI score was less than 5% (i.e., an RCI score of $|1.96|$ or greater).

Table 6 summarizes our RCI results. The first four columns show the number of individuals who showed significant increases or decreases on each scale in the self-ratings and spouse ratings. These results again highlight the marked discrepancy in the patterns of mean-level change across the two sets of ratings. Consider, for instance, the results for Conscientiousness. In the self-ratings,

Table 4
Within-Method (Mono-Rater) Correlations Between Change Scores Across Traits

Scale	1	2	3	4	5	6	7
1. Neuroticism	—	-.21**	-.06	-.36**	-.29**	.44**	-.31**
2. Extraversion	-.15**	—	.17**	.18**	.18**	-.23**	.32**
3. Openness	-.08	.26**	—	.26**	.28**	-.09	.22**
4. Agreeableness	-.20**	.20**	.22**	—	.33**	-.35**	.28**
5. Conscientiousness	-.15*	.14*	.08	.26**	—	-.21**	.31**
6. Negative Affect	.42**	-.02	-.03	-.25**	-.18**	—	-.21**
7. Positive Affect	-.23**	.37**	.19**	.28**	.30**	-.06	—

Note. $N = 301$. Self-rating correlations are shown below the diagonal; spouse-rating correlations are presented above the diagonal.

* $p < .05$. ** $p < .01$.

Table 6
Summary of Results Showing the Number of Participants With Significant Reliable Change Index Scores

Scale	Self-rating		Spouse rating		Cross-method comparison			
	Decreased	Increased	Decreased	Increased	Consistent change	Paradoxical change	Inconsistent change	No change
BFI								
Neuroticism	19	6	12	16	3	0	47	251
Extraversion	10	7	12	4	3	0	27	271
Openness	17	16	16	9	2	3	48	248
Agreeableness	11	10	25	5	0	2	47	252
Conscientiousness	6	22	22	6	3	1	48	249
PANAS								
Negative Affect	13	5	7	13	0	0	38	263
Positive Affect	8	10	9	11	0	0	38	263

Note. $N = 301$. BFI = Big Five Inventory; PANAS = Positive and Negative Affect Schedule.

22 of the 28 individuals (78.6%) who showed significant change reported increased levels of the trait. The spouse ratings showed exactly the opposite pattern: Here, the large majority of reliable change (22 of 28 individuals, or 78.6%) reflected significant declines in Conscientiousness. Similarly, whereas the bulk of the reliable change in self-rated Neuroticism (76%) and Negative Affect (72.2%) indicated decreasing negative emotionality, most of the significant change in the spouse ratings was in the opposite direction (i.e., 57.1% and 65% of the significant RCI scores for Neuroticism and Negative Affect, respectively, were associated with higher scores at Time 2).

Still, these results do not directly address the question of whether the self-raters and spouse raters agreed about which specific individuals showed significant change. This issue is addressed in the last four columns of Table 6, which summarize the cross-method comparisons into four categories: consistent change (i.e., the self-ratings and spouse ratings both yielded significant RCI scores in the same direction), paradoxical change (i.e., the two sets of ratings both yielded significant RCI scores, but one showed an increase and the other a decrease), inconsistent change (i.e., one rating yielded a significant RCI score, but the other did not), and no change (i.e., both ratings produced nonsignificant RCI scores). These results yield very little evidence of consistency across the two methods. Across the seven traits, there were a total of only 11 instances of consistent change (all on the BFI scales). It is noteworthy, moreover, that there actually were 6 cases of paradoxical change in our data. Taken together with our earlier analyses of raw change scores, these results indicate that individual-level change shows little consistency across our two rating methods.

Exploring the Self-Spouse Discrepancy: The Influence of Marital Satisfaction

Personality ratings and marital satisfaction. How can we explain this marked divergence between self-rated versus spouse-rated changes in personality? Earlier, we suggested that the spouse ratings reflected a honeymoon effect at Time 1 that had dissipated by Time 2. This, in turn, suggests that marital satisfaction may play a key role in explaining our findings. In this regard, previous research has established strong links between current relationship satisfaction and personality ratings of romantic partners. For ex-

ample, Watson, Hubbard, and Wiese (2000a) analyzed ratings of the Big Five and trait affectivity in dating and married couples. They found that the judge's current level of relationship satisfaction was only weakly related to the *self-rated* personality characteristics of their romantic partners. For instance, relationship satisfaction in the dating women correlated only .18 and .17 with the self-reported agreeableness and conscientiousness of their male partners; conversely, satisfaction among the dating men correlated only .11 and .21 with the self-rated agreeableness and conscientiousness, respectively, of their female partners. These data therefore suggest that a person's self-perceived standing on these traits had relatively little effect on the satisfaction of their romantic partner.

In sharp contrast, however, satisfaction was much better predictor of the *judges' ratings* of their partners' traits. Thus, a dating woman's relationship satisfaction correlated .41 and .50, respectively, with her ratings of her male partner's agreeableness and conscientiousness; similarly, satisfaction among dating men correlated .30 and .55, respectively, with their ratings of their female partners' agreeableness and conscientiousness.

The IMAP participants rated their current level of marital satisfaction at both assessments. Satisfaction was assessed using a single global rating derived from the Locke-Wallace Marital Adjustment Test (Locke & Wallace, 1959). Participants chose "the number which best describes the degree of happiness, everything considered, that you feel in your present marriage"; these ratings were made on a 7-point scale ranging from *very unhappy* to *perfectly happy*. Not surprisingly, given that the participants were newlyweds at the initial assessment, their mean level of satisfaction declined significantly from Time 1 to Time 2. Specifically, the mean level of satisfaction among the wives dropped from 5.68, $SD = 1.02$, to 5.39, $SD = 1.21$; $t(150) = 2.98$, $p < .01$; Cohen's $d = .24$, whereas the husbands' satisfaction decreased from 5.77, $SD = 0.95$, to 5.54, $SD = 1.12$; $t(145) = 2.76$, $p < .01$; Cohen's $d = .23$. These data raise the possibility that spouse-rated changes in personality—which, in marked contrast to the self-reports, tended to indicate a negative developmental trajectory (e.g., lower agreeableness and conscientiousness)—may reflect, in part, the significant decline in satisfaction that occurred over the course of our study.

Correlational analyses. We conducted several series of analyses to examine this possibility. In the first set, we sought to replicate the results of Watson et al. (2000a), demonstrating an especially strong link between marital satisfaction and personality ratings of the romantic partner. Table 7 presents hetero-rater correlations between a participant's current marital satisfaction and the self-rated personality characteristics of his or her spouse. For instance, the coefficient in the first row and column of Table 7 represents the correlation between a wife's marital satisfaction and her husband's self-rated Neuroticism at Time 1. These correlations are reported separately for the husbands and wives at each assessment.

These results reveal several significant associations between marital satisfaction and the spouse's self-rated characteristics. Most notably, marital satisfaction tended to be positively associated with Agreeableness and negatively related to Neuroticism/Negative Affectivity. Thus, participants tended to report greater satisfaction if they were married to spouses who were agreeable and emotionally stable. Overall, 14 of the 28 coefficients (50%) were significant at $p < .05$. At the same time, however, it also should be noted that these correlations tend to be relatively low in magnitude. Specifically, only five coefficients (17.9%) are as high as $|.20|$, and only two (7.1%) exceed $|.30|$.

These results provide an interesting context for interpreting the mono-rater correlations between a participant's marital satisfaction and his or her ratings of the spouse's trait characteristics; these coefficients are reported in Table 8. For example, the coefficient in the first row and column of Table 8 reflects the correlation between a wife's satisfaction and her ratings of her husband's Neuroticism at Time 1. These correlations clearly are systematically stronger than those reported in Table 7. Of the 28 correlations, 27 (96.4%) are significant at $p < .05$. Furthermore, 25 coefficients (89.3%) exceed $|.20|$, and 19 (67.9%) exceed $|.30|$; indeed, 10 correlations (35.7%) are $|.40|$ and greater. Follow-up tests indicated that 27 of the 28 correlations in Table 8 differed significantly from the corresponding values reported in Table 7 (z s ranged from $|2.00|$ to $|5.47|$; all $ps < .05$); the single exception was that the

Table 7
Hetero-Rater Correlations Between Marital Satisfaction and the Spouse's Self-Rated Personality

Scale	Wife's satisfaction: Husband's personality		Husband's satisfaction: Wife's personality	
	Time 1	Time 2	Time 1	Time 2
BFI				
Neuroticism	-.12*	-.18*	-.15*	-.33**
Extraversion	.02	-.07	-.01	.19**
Openness	.02	.06	.10	.05
Agreeableness	.16**	.23**	.10	.23**
Conscientiousness	-.07	.11	.03	.17*
PANAS				
Negative Affect	-.16**	-.26**	-.09	-.31**
Positive Affect	-.02	.06	.15*	.18*

Note. $N = 283$ (Time 1); $n = 151$, wife's Time 2 satisfaction; $n = 149$, husband's Time 2 satisfaction. BFI = Big Five Inventory; PANAS = Positive and Negative Affect Schedule.
* $p < .05$. ** $p < .01$.

Table 8
Mono-Rater Correlations Between Marital Satisfaction and the Judge's Ratings of the Spouse's Personality

Scale	Wife's satisfaction: Husband's personality traits		Husband's satisfaction: Wife's personality traits	
	Time 1	Time 2	Time 1	Time 2
BFI				
Neuroticism	-.28**	-.39**	-.36**	-.49**
Extraversion	.21**	.11	.26**	.41**
Openness	.24**	.22**	.33**	.42**
Agreeableness	.45**	.53**	.37**	.48**
Conscientiousness	.19**	.31**	.19**	.37**
PANAS				
Negative Affect	-.35**	-.40**	-.29**	-.55**
Positive Affect	.34**	.41**	.30**	.55**

Note. $N = 283$ (Time 1); $n = 151$, wife's Time 2 satisfaction, $n = 149$, husband's Time 2 satisfaction. BFI = Big Five Inventory; PANAS = Positive and Negative Affect Schedule.
** $p < .01$.

correlations between a wife's satisfaction and her husband's negative affectivity at Time 2 did not differ from one another ($-.40$ vs. $-.26$; $z = -1.66$, $p < .10$).

Thus, our findings replicate those reported by Watson et al. (2000a). These results again indicate that spouse ratings are closely linked to current levels of marital satisfaction.

Predicting changes in spouse-rated personality. Next, we conducted hierarchical regression analyses to test whether spouse-rated changes in personality could be predicted from changes in marital satisfaction between Time 1 and Time 2. The results of these analyses are presented in Tables 9 (husbands' ratings) and 10 (wives' ratings). The participant's Time 2 ratings of his or her spouse's trait characteristics served as the criteria in these regressions. In each case, the rater's Time 1 trait rating was entered as a predictor in Step 1, followed by Time 1 marital satisfaction in Step

Table 9
Hierarchical Multiple Regressions: Predicting Time 2 Spouse-Rated Traits From Time 2 Marital Satisfaction (Husbands' Ratings)

Scale	Step 1	Step 2	Step 3	
	T1 trait	T1 satisfaction	T2 marital satisfaction	β
	ΔR^2	ΔR^2	ΔR^2	
BFI				
Neuroticism	.554**	.003	.035**	-.234**
Extraversion	.648**	.000	.017**	.161**
Openness	.552**	.030**	.029**	.204**
Agreeableness	.421**	.015	.097**	.376**
Conscientiousness	.628**	.001	.041**	.245**
PANAS				
Negative Affect	.452**	.001	.091**	-.375**
Positive Affect	.389**	.047**	.087**	.352**

Note. $N = 146$. T1 = Time 1; T2 = Time 2; BFI = Big Five Inventory; PANAS = Positive and Negative Affect Schedule.
** $p < .01$.

Table 10
Hierarchical Multiple Regressions: Predicting Time 2 Spouse-Rated Traits From Time 2 Marital Satisfaction (Wives' Ratings)

Scale	Step 1	Step 2	Step 3	
	T1 trait	T1 satisfaction	T2 marital satisfaction	β
	ΔR^2	ΔR^2	ΔR^2	
BFI				
Neuroticism	.479**	.057**	.012	-.127
Extraversion	.723**	.000	.000	.008
Openness	.629**	.002	.011*	.123*
Agreeableness	.569**	.010	.036**	.225**
Conscientiousness	.509**	.011	.012	.128
PANAS				
Negative Affect	.461**	.064**	.018*	-.153*
Positive Affect	.477**	.014*	.029**	.198**

Note. $N = 146$. T1 = Time 1; T2 = Time 2; BFI = Big Five Inventory; PANAS = Positive and Negative Affect Schedule.

* $p < .05$. ** $p < .01$.

2; the rater's Time 2 satisfaction then was added in Step 3 to determine its incremental predictive power. For example, the first row in Table 9 shows an analysis in which the husbands' ratings of their wives' Neuroticism at Time 2 served as the criterion. We entered the husband's Time 1 rating of his wife's Neuroticism in Step 1 and the husband's Time 1 marital satisfaction in Step 2; we then added the husband's Time 2 marital satisfaction in Step 3. Thus, by controlling for both Time 1 satisfaction and the corresponding Time 1 personality rating, these analyses allow us to determine whether changes in marital satisfaction predict changes in the spouse ratings.

We consider first the analyses of the husbands' ratings (see Table 9). These results clearly establish that changes in spouse ratings can be predicted from changes in marital satisfaction. In every case, the inclusion of Time 2 marital satisfaction in Step 3 was associated with a significant increase in predictive power; across the seven analyses, Time 2 marital satisfaction contributed from 1.7% to 9.7% incremental variance ($M = 5.7%$). The largest incremental variance effects were seen for ratings of Agreeableness (9.7%), Negative Affect (9.1%), and Positive Affect (8.7%). These results are particularly impressive given the very strong rank-order stabilities of these scales. Indeed, Table 9 indicates that the Time 1 trait scores already accounted for 42.1% to 64.8% ($M = 52.1%$) of the criterion variance in Step 1.

Table 10 presents parallel results for the wives' ratings. The effects here clearly are weaker overall. Across the seven analyses, Time 2 marital satisfaction contributed from 0% to 3.6% incremental variance, with a mean value of 1.7%. Still, Time 2 satisfaction contributed significantly in four of the seven analyses. It is noteworthy, moreover, that Agreeableness, Negative Affect, and Positive Affect again showed the strongest incremental effects, suggesting that these ratings are especially sensitive to changes in marital satisfaction.

The data presented in Tables 9 and 10 generally indicate that spouse-rated changes in personality reflect, in part, changes in marital satisfaction between Time 1 and Time 2. Given that marital satisfaction declined significantly across the study interval, these analyses help to explain why the spouse ratings—in marked con-

trast to the self-ratings—tended to be somewhat more negative in character at Time 2.

Predicting changes in self-rated personality. Is this link with marital satisfaction specific to spouse ratings, or does it also characterize self-reports? We examined this important issue in a parallel series of hierarchical regression analyses. These regressions were identical in form to those presented in Tables 9 and 10, except that they were based on self-rated—rather than spouse rated—trait scores. In one analysis, for instance, the criterion was the husband's self-rated Neuroticism at Time 2. We added his self-rated Time 1 Neuroticism in Step 1 and his Time 1 marital satisfaction in Step 2; his Time 2 marital satisfaction was then entered in Step 3. These analyses therefore allow us to determine whether changes in marital satisfaction predict changes in self-ratings between Time 1 and Time 2.

These analyses established that changes in marital satisfaction essentially were unrelated to self-rated change in personality. Time 2 marital satisfaction contributed from 0% to 2.2% incremental variance in the husbands' ratings ($M = 0.5%$), and from 0% to 0.6% incremental variance in the wives' ratings ($M = 0.2%$). Across the 14 analyses, only 1 produced a significant effect: Time 2 satisfaction was negatively related to change on the PANAS Negative Affect scale in the husbands' data ($\Delta R^2 = .022$, $\beta = -.182$, $p < .05$); in other words, husbands who reported greater negative affectivity also experienced a decline in marital satisfaction. Overall, however, our results establish that marital satisfaction is a much more powerful predictor of spouse-rated change than of self-rated change.

Predicting changes in marital satisfaction. Earlier, we established that changes in marital satisfaction predict changes in spouse-rated personality traits. We conducted a final series of hierarchical regression analyses to determine whether similar effects could be observed in the opposite direction, that is, whether changes in spouse-rated personality traits also could predict changes in marital satisfaction. The rater's Time 2 marital satisfaction served as the criterion in these regressions. In each case, the rater's Time 1 marital satisfaction was entered as a predictor in Step 1, followed by his or her Time 1 trait rating of the partner in Step 2; the rater's Time 2 trait rating of the spouse then was added in Step 3 to examine its incremental predictive power. For example, the wife's Time 1 marital satisfaction was entered in Step 1, followed by her Time 1 rating of her husband's Agreeableness in Step 2; finally, her Time 2 rating of her husband's Agreeableness was entered in Step 3.

Tables 11 (husbands' ratings) and 12 (wives' ratings) summarize the results from Steps 2 and 3 of these regressions. The pattern of these results is very similar to that reported in Tables 9 and 10, but the overall magnitude of the effects is slightly stronger in these analyses. In the husbands' data, the inclusion of the Time 2 spouse rating in Step 3 was associated with a significant increase in predictive power in every case; across the seven analyses, the Time 2 trait ratings contributed from 3.2% to 11.8% incremental variance ($M = 7.7%$) in predicting Time 2 marital satisfaction. The largest effects again were seen for ratings of Agreeableness (11.8% incremental variance), Positive Affect (10.8%), and Negative Affect (10.7%). As in the earlier analyses, the wives' ratings yielded weaker effects overall. Across the seven traits, the Time 2 spouse rating contributed from 0% to 6.0% incremental variance, with a mean value of 2.8%. Still, the Time 2 trait score contributed

Table 11
Hierarchical Multiple Regressions: Predicting Time 2 Marital Satisfaction From Time 2 Spouse-Rated Traits (Husbands' Ratings)

Scale	Step 2	Step 3		Final <i>R</i>
	T1 trait	T2 trait	T2 trait	
	ΔR^2	ΔR^2	β	
BFI				
Neuroticism	.069**	.051**	-.338**	.64
Extraversion	.053**	.032**	.301**	.61
Openness	.005	.049**	.344**	.59
Agreeableness	.022*	.118**	.458**	.66
Conscientiousness	.024*	.076**	.452**	.63
PANAS				
Negative Affect	.073**	.107**	-.443**	.68
Positive Affect	.016	.108**	.437**	.64

Note. $N = 146$. T1 marital satisfaction was entered in Step 1 of the regression. T1 = Time 1; T2 = Time 2; BFI = Big Five Inventory; PANAS = Positive and Negative Affect Schedule.

* $p < .05$. ** $p < .01$.

significantly in four of the seven analyses, with Agreeableness, Negative Affect, and Positive Affect again showing significant incremental effects. Overall, therefore, these results further demonstrate the significant link between spouse ratings and changes in marital satisfaction.

Discussion

Rank-Order Stability

Our findings regarding rank-order stability supported our predictions and were broadly consistent with previous work in this area. As expected, the stability correlations for the self-ratings and spouse ratings were very similar; in the overall sample, in fact, the average stability correlations were virtually identical for the self-ratings (mean $r_s = .78$ and $.69$ for the BFI and PANAS, respectively) and the spouse ratings (mean $r_s = .77$ and $.69$, respectively). Moreover, the correlations in our study generally were quite high, ranging from $.67$ to $.83$. As we predicted, all of our stability correlations exceeded those previously reported by Vaidya et al. (2002) for these same scales in a somewhat younger sample. Thus, our results again demonstrate that the stability of personality increases systematically with age (Roberts & DelVecchio, 2000).

Our data also replicated previous research indicating that the BFI scales show higher stability correlations than the PANAS (see Vaidya et al., 2002; Watson, 2004). The findings for Extraversion and Neuroticism are particularly striking. BFI Extraversion was significantly more stable than the PANAS Positive Affect scale in both the self-ratings and the spouse ratings; BFI Neuroticism also had higher stability correlations than PANAS Negative Affect in both sets of ratings, although this difference was not significant in the spouse ratings. These differences in stability are noteworthy in light of the consistently strong correlations (which ranged from $.52$ to $.79$ in our study) between these BFI and PANAS scales.

The stability gap between Extraversion and Positive Affect most likely reflects systematic differences in content between the two scales. The BFI Extraversion scale asks respondents to indicate

whether they are talkative, outgoing, and reserved (reverse-keyed); only two items ask directly about affect-related content ("is full of energy," "generates a lot of enthusiasm"). More generally, Pytlik Zillig, Hemenover, and Dienstbier (2002) found that only 22.7% of the content in the BFI Extraversion scale was affective in nature. Thus, this gap likely reflects the fact that individual differences in sociability are more stable over time than positive emotionality.

However, simple content-based considerations cannot explain the fact that BFI Neuroticism tends to be more stable than the PANAS Negative Affect scale. The content of the BFI Neuroticism scale is predominantly affective in nature; indeed, Pytlik Zillig et al. (2002) classified 75.3% of its item content as affective (see also Watson, 2004, for an examination of the stability of individual BFI Neuroticism and PANAS Negative Affect items). It therefore seems likely that noncontent considerations—such as instructional or format effects—are at least partly responsible for this evidence of differential stability.

Watson (2004) tested this possibility by constructing a new instrument, the Temperament and Emotion Questionnaire (TEQ). The 60-item TEQ was created by taking individual mood descriptors from the Expanded Form of the PANAS (PANAS-X; Watson & Clark, 1999) and embedding them in complete sentences (which are rated using a 5-point *agreedisagree* format). For instance, the PANAS/PANAS-X term *irritable* became the TEQ item "I have days on which I can be rather irritable." Watson (2004) then compared retest correlations for the PANAS-X and TEQ negative affectivity scales across a 2-month interval. In three of five comparisons, the stability of the TEQ scale was significantly higher than that of its PANAS-X counterpart. These data demonstrate that stability correlations can be significantly influenced by changes in wording and format, even while maintaining the same basic item content. More generally, they suggest that stability researchers should carefully attend to issues such as wording and format when choosing measures to be used in studies of adult personality development.

Mean-Level Change

Summary of findings. Our analyses of mean-level change revealed a striking discrepancy between the self-ratings and spouse

Table 12
Hierarchical Multiple Regressions: Predicting Time 2 Marital Satisfaction From Time 2 Spouse-Rated Traits (Wives' Ratings)

Scale	Step 2	Step 3		Final <i>R</i>
	T1 trait	T2 trait	T2 trait	
	ΔR^2	ΔR^2	β	
BFI				
Neuroticism	.010	.020	-.207	.52
Extraversion	.000	.000	.022	.49
Openness	.005	.028*	.251*	.52
Agreeableness	.060**	.060**	.376**	.60
Conscientiousness	.026*	.018	.196	.53
PANAS				
Negative Affect	.014	.028*	-.241*	.53
Positive Affect	.016	.043**	.290**	.54

Note. $N = 146$. T1 marital satisfaction was entered in Step 1 of the regression. T1 = Time 1; T2 = Time 2; BFI = Big Five Inventory; PANAS = Positive and Negative Affect Schedule.

* $p < .05$. ** $p < .01$.

ratings. Our self-report data were consistent with prediction and largely replicated previous findings in this area. Specifically, our participants reported significant increases in both conscientiousness and agreeableness, and significant declines in neuroticism/negative affectivity. We did not observe any systematic change in extraversion, openness, and positive affectivity across this 2-year interval. Overall, therefore, our findings are consistent with the broader literature indicating that adult personality development is associated with increasing psychological maturity with age (see Caspi et al., 2005).

However, the spouse ratings yielded a markedly different pattern and suggested very different conclusions about the nature of adult development. Specifically, these scores showed significant declines in agreeableness, conscientiousness, extraversion, and openness, a pattern that certainly would not be characterized as reflecting greater psychological maturity. Moreover, across the seven assessed scales, we obtained only one consistent finding: PANAS Positive Affect showed a small, nonsignificant decline in both sets of ratings.

For reasons discussed earlier, we suspected that changes in marital satisfaction might provide at least a partial explanation for these discrepant results. We conducted correlational and hierarchical regression analyses to examine this possibility. We found that changes in marital satisfaction were significantly associated with changes in spouse-rated personality, with each predicting the other in a separate series of analyses; the effects were particularly strong for agreeableness and trait affectivity. Subsequent analyses revealed that changes in marital satisfaction essentially were unrelated to self-rated change in personality. Thus, our findings suggest that marital satisfaction played a key role in producing the discrepant findings presented in Tables 2 and 3. Given that marital satisfaction declined significantly across the study period, it makes sense that our spouse ratings tended to be more negative overall at Time 2.

Explaining the findings. This, then, leads to a more basic set of questions: How can these findings be reconciled and integrated into an overall model of adult personality development? Is the nature of this development largely positive (as indicated by the self-ratings) or rather negative (as suggested by the spouse ratings)? Put differently, which rating source provides a more accurate picture of mean-level change in emerging adulthood?

One possible explanation of our data is that the spouse raters were able to remain more objective and, thus, ultimately had better insight into the targets' true personalities than the self-raters. In other words, our IMAP participants actually showed negative changes in their trait characteristics—including significant declines in agreeableness and conscientiousness—across the 2-year study interval. Because these changes are negative and socially undesirable, however, the self-raters understandably may have been reluctant to acknowledge them. In contrast, spouse raters were not as motivated to gloss over these negative developmental trends and so were able to provide a more accurate account of the targets' true personalities at Time 2.

This explanation is consistent with our hierarchical regression analyses that established that changes in spouse ratings predicted changes in marital satisfaction (see Tables 11 and 12); it also is supported by evidence establishing the existence of a self-enhancement bias in self-report data (John & Robins, 1993; Kwan, John, Kenny, Bond, & Robins, 2004; Paulhus, Harms, Bruce, &

Lysy, 2003). However, there are two significant problems with a simple self-enhancement explanation of our data. First, as we reviewed earlier, a large body of evidence—based on both self-reports and, to a lesser extent, observer ratings—indicates that personality development is associated with greater maturity (i.e., higher levels of conscientiousness and agreeableness, lower levels of neuroticism/negative affectivity) over time. The mean-level changes in our self-rating data obviously conformed much more closely to this typical pattern, and so they appear to be more credible than the results that emerged in the spouse ratings. Second, a self-enhancement explanation clearly cannot account for the “honeymoon effect” we observed at Time 1, such that the mean spouse ratings significantly exceeded the average self-ratings for Agreeableness, Extraversion, and Conscientiousness. In other words, the spouse ratings actually were more positive and socially desirable than the self-ratings at Time 1.

In light of these considerations, we believe it is unlikely that our participants actually showed true negative changes in their trait characteristics from Time 1 to Time 2. A more likely explanation is that the spouse ratings were unduly positive during the honeymoon period of Time 1, and then declined to more realistic levels at Time 2. However, this then leads to the further question of what caused these changes in the spouse ratings over the course of our study. One possibility is that the actual day-to-day behavior of our participants changed significantly between Time 1 and Time 2. That is, our IMAP participants may have been on their “best behavior” during the honeymoon period of Time 1 and acted as more agreeable and conscientious—and less neurotic—than they really were; this would explain why the spouse ratings were more positive than the self-ratings at this initial assessment. As the marriage wore on and the honeymoon ended, however, the participants gradually reverted to their true baseline levels of behavior. If so, then this would naturally lead to declines in marital satisfaction and to less positive spouse ratings at Time 2.

We suspect that behavioral changes of this type offer at least a partial explanation of our findings. However, these changes alone cannot explain why the spouse ratings were significantly more negative than the self-ratings (i.e., higher levels of Neuroticism and Negative Affect, lower levels of Conscientiousness and Openness) at Time 2. We believe that the observed decline in marital satisfaction played a key role in producing this Time 2 negativity in the spouse ratings. Although they know each other well, husbands still are likely to have significant gaps in their knowledge about their wives, and vice versa. We now have extensive evidence indicating that judges use various rating strategies or heuristics to fill in these informational gaps. Of particular relevance here, Watson et al. (2000a) reviewed evidence suggesting that relationship satisfaction represents a significant heuristic that can be used when rating romantic partners: That is, raters compensate for informational gaps by using their current level of satisfaction as a basis for making strongly evaluative inferences about the personalities of their partners. Thus, at Time 1, the participants were extremely satisfied with their marriages and so rated their spouses very positively (even more positively than their partners rated themselves). Because the spouses became significantly more dissatisfied with their marriages over time, however, they judged their partners more harshly at Time 2.

This heuristic-based model also offers a parsimonious explanation for the well-established finding that marital satisfaction (a) is

only weakly associated with the partner's self-rated traits but (b) is much more substantially linked to the judge's ratings of the partner's characteristics (see Watson et al., 2000a); we replicated this pattern at both Time 1 and Time 2 (see Tables 7 and 8). It is noteworthy that these mono-rater correlations tend to be strong even when self–other agreement is poor and judges are rating low visibility traits that are difficult to observe in others. For example, Watson et al. (2000b) obtained relatively low self–other agreement correlations for ratings of trait affectivity in a sample of dating couples ($r_s = .22$ and $.33$ on the PANAS Negative Affect and Positive Affect scales, respectively). Nevertheless, relationship satisfaction was strongly correlated with partner ratings of trait affectivity in this same sample, with coefficients ranging from $|.40|$ to $|.56|$ (see Watson et al., 2000a, Table 7). Thus, consistent with a heuristic-based explanation, we tend to see strong correlations between partner-rated personality and satisfaction even under conditions in which trait-related information is limited and the ratings are unlikely to be highly accurate. Consequently, we believe that this heuristic-based account also provides at least a partial explanation for our findings.

We emphasize, however, that our data do not allow us to draw clear, general conclusions regarding the relative merits of these various explanations. We emphasize, moreover, that these models are not mutually exclusive, and they each may offer partial explanations for our findings. This is a crucial issue that needs to be investigated more thoroughly in future research. In order to weigh the relative merits of these models, it would be helpful to include a third rating source (e.g., friends' ratings of both spouses) in future studies. It also would be highly informative to collect converging data using other assessment approaches (e.g., time sampling of trait-related behaviors and feelings). Finally, future research in this area should use longer, more reliable measures of marital satisfaction.

Individual-Level Change

We also investigated change at the individual level. We began by examining the within-method correlations among the raw change scores in both the self-ratings and the spouse ratings (see Table 4). These data demonstrated that change scores on related traits were significantly correlated with one another. These results help to establish that these scores tap some systematic variance and do not simply reflect random measurement error. Having said that, however, we must emphasize that this does not necessarily mean that this systematic variance is valid. Among other things, it may reflect systematic measurement errors, such as transient error (e.g., Becker, 2000; Schmidt, Le, & Ilies, 2003). Transient error reflects the influence of time-limited factors, such as the current mood of the respondent. For instance, if some participants were more distressed and upset at Time 2 than at Time 1, this could have influenced their responses on both BFI Neuroticism and PANAS Negative Affect, thereby producing a positive correlation between the change scores for these two scales.

Indeed, our subsequent analyses raised some significant concerns about the overall validity of change scores. We next examined the convergence between self- and spouse-rated change using these raw change scores. These findings offered mixed support for our predictions. We obtained significant convergent correlations for only two of seven scales in the overall sample: Conscientious-

ness ($r = .20$) and Neuroticism ($r = .19$). Moreover, the coefficients were consistently low in magnitude, ranging from $-.01$ to $.20$, with a median value of only $.07$. Thus, our data revealed little convergence between change assessments across two different rating methods. Our subsequent analyses of RCI scores yielded the same basic conclusion: Again, we found little consistency in individual-level change across our two rating methods.

Our results replicate those of McCrae (1993), who reported very low correlations between self- and spouse-rated changes on neuroticism, extraversion, and openness. It is noteworthy, moreover, that McCrae also found little convergence with change scores that were based on the ratings of two peers. It therefore appears that this problem may not be restricted to spouse ratings, but instead reflects a more general pattern.

This poor convergence across methods is troubling and raises significant concerns about the meaningfulness of change assessed at the individual level. The available evidence remains quite limited, however, so we strongly encourage more extensive investigation of this issue in subsequent research. Paralleling our earlier discussion of mean-level change, we believe it would be particularly informative to assess individual-level change in personality across multiple methods (e.g., self-reports, spouse ratings, peer ratings, time sampling) in a reasonably large sample.

Strengths and Limitations

This study contributes to the literature on personality development in emerging adulthood in several ways. First, we examined personality stability and change in a relatively large sample of young adults across a time span of approximately 2 years. Second, our participants were somewhat older than those typically investigated in studies of young adulthood. We therefore were able to investigate key developmental issues during the transitional period from "emerging adulthood" to full adulthood. Our data revealed that this critical developmental period is characterized by both very strong rank-order stability and significant mean-level change. Third, because our sample consisted of newlywed couples, we were able to examine stability and change across two different rating methods. Although these methods yielded virtually identical results in our examination of rank-order stability, they diverged sharply in our analyses of both mean-level change and individual-level change. These data highlight the importance of multimethod assessment in studies of adult personality development. We therefore join others (e.g., McCrae, Terracciano, et al., 2005; Roberts et al., 2004) who recently have called for increased reliance on multisource data in this area.

At the same time, we also must acknowledge two significant limitations of our study. First, we examined personality stability and change in a single sample across only two assessments. Accordingly, certain aspects of our results may reflect study-specific factors that may not generalize across other samples and occasions. Second, because we collected personality data from only two sources, we were unable to resolve the striking discrepancies that emerged in our analyses of the self- and spouse-rating data. As noted previously, we believe that it will be extremely informative for future studies to obtain relevant longitudinal data from multiple sources. An expanded design of this type would be invaluable in clarifying the true nature and course of adult personality development.

References

- Arnett, J. J. (2000). Emerging adulthood: A theory of development from the late teens through the twenties. *American Psychologist*, *55*, 469–480.
- Asendorpf, J. (1992). Beyond stability: Predicting inter-individual differences in intra-individual change. *European Journal of Personality*, *6*, 103–117.
- Becker, G. (2000). How important is transient error in estimating reliability? Going beyond simulation studies. *Psychological Methods*, *5*, 370–379.
- Benet-Martinez, V., & John, O. P. (1998). *Los Cinco Grandes* across cultures and ethnic groups: Multitrait multimethod analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology*, *75*, 729–750.
- Caspi, A., Roberts, B. W., & Shiner, R. L. (2005). Personality development: Stability and change. *Annual Review of Psychology*, *56*, 453–484.
- Chmielewski, M., & Watson, D. (2006). *Investigations of dependability: Comparisons of two-week versus two-month retest correlations for measures of personality and psychopathology*. Manuscript in preparation.
- Clark, L. A., & Watson, D. (1999). Temperament: A new paradigm for trait psychology. In L. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 399–423). New York: Guilford Press.
- Costa, P. T., Jr., & McCrae, R. R. (1988). Personality in adulthood: A six-year longitudinal study of self-reports and spouse ratings on the NEO Personality Inventory. *Journal of Personality and Social Psychology*, *54*, 853–863.
- Costa, P. T., Jr., & McCrae, R. R. (1992). 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. (1994). Set like plaster: Evidence for the stability of adult personality. In T. F. Heatherton & J. L. Weinberger (Eds.), *Can personality change?* (pp. 21–40). Washington, DC: American Psychological Association.
- Digman, J. M. (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology*, *73*, 1246–1256.
- Fraley, R. C., & Roberts, B. W. (2004). Patterns of continuity: A dynamic model for conceptualizing the stability of individual differences in psychological constructs across the life course. *Psychological Review*, *112*, 60–74.
- Humrichouse, J., & Watson, D. (2006). *Self-spouse ratings, trait visibility, and acquaintanceship: A longitudinal study of newlywed couples*. Manuscript in preparation.
- John, O. P., & Robins, R. W. (1993). Determinants of interjudge agreement on personality traits: The Big Five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality*, *61*, 521–551.
- John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality* (2nd ed., pp. 102–138). New York: Guilford Press.
- Kashy, D. A., & Snyder, D. K. (1995). Measurement and data analytic issues in couples research. *Psychological Assessment*, *7*, 338–348.
- Kenny, D. A. (1995). The effect of nonindependence on significance testing in dyadic research. *Personal Relationships*, *2*, 67–75.
- Kwan, V., John, O. P., Kenny, D. A., Bond, M. H., & Robins, R. W. (2004). Reconceptualizing individual differences in self-enhancement bias: An interpersonal approach. *Psychological Review*, *111*, 94–110.
- Locke, H., & Wallace, K. (1959). Short marital-adjustment and prediction tests: Their reliability and validity. *Marriage and Family Living*, *21*, 251–255.
- Markon, K. E., Krueger, R. F., & Watson, D. (2005). Delineating the structure of normal and abnormal personality: An integrative hierarchical approach. *Journal of Personality and Social Psychology*, *88*, 139–157.
- McCrae, R. R. (1993). Moderated analyses of longitudinal personality stability. *Journal of Personality and Social Psychology*, *65*, 577–585.
- McCrae, R. R., & Costa, P. T., Jr. (1990). *Personality in adulthood*. New York: Guilford Press.
- McCrae, R. R., Costa, P. T., Jr., Hřebíčková, M., Urbánek, T., Martin, T. A., Oryol, V. E., et al. (2004). Age differences in personality traits across cultures: Self-report and observer perspectives. *European Journal of Personality*, *18*, 143–157.
- McCrae, R. R., Terracciano, A., & Members of the Personality Profiles of Cultures Project (2005). Universal features of personality traits from the observer's perspective: Data from 50 cultures. *Journal of Personality and Social Psychology*, *88*, 547–561.
- Paulhus, D. L., Harms, P. D., Bruce, M. N., & Lysy, D. C. (2003). The over-claiming technique: Measuring self-enhancement independent of ability. *Journal of Personality and Social Psychology*, *84*, 890–904.
- Pytlík Zillig, L. M., Hemenover, S. H., & Dienstbier, R. A. (2002). What do we assess when we assess a Big 5 trait? A content analysis of the affective, behavioral, and cognitive processes represented in Big 5 personality inventories. *Personality and Social Psychology Bulletin*, *28*, 847–858.
- Roberts, B. W., Caspi, A., & Moffitt, T. E. (2001). The kids are alright: Growth and stability in personality development from adolescence to adulthood. *Journal of Personality and Social Psychology*, *81*, 670–683.
- Roberts, B. W., Caspi, A., & Moffitt, T. E. (2003). Work experiences and personality development in young adulthood. *Journal of Personality and Social Psychology*, *84*, 582–593.
- Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin*, *126*, 3–25.
- Roberts, B. W., O'Donnell, M., & Robins, R. W. (2004). Goal and personality trait development in emerging adulthood. *Journal of Personality and Social Psychology*, *87*, 541–550.
- Roberts, B. W., Robins, R. W., Trzesniewski, K., & Caspi, A. (2003). Personality trait development in adulthood. In J. Mortimer & M. Shanahan (Eds.), *Handbook of the life course* (pp. 579–598). New York: Kluwer.
- Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, *132*, 1–25.
- Robins, R. W., Fraley, R. C., Roberts, B. W., & Trzesniewski, K. H. (2001). A longitudinal study of personality change in young adulthood. *Journal of Personality*, *69*, 617–640.
- Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods*, *8*, 206–224.
- Schuerger, J. M., Zarrella, K. L., & Hotz, A. S. (1989). Factors that influence the temporal stability of personality by questionnaire. *Journal of Personality and Social Psychology*, *56*, 777–783.
- Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology*, *84*, 1041–1053.
- Vaidya, J. G., Gray, E. K., Haig, J., & Watson, D. (2002). On the temporal stability of personality: Evidence for differential stability and the role of life experiences. *Journal of Personality and Social Psychology*, *83*, 1469–1484.
- Watson, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality*, *38*, 319–350.
- Watson, D., & Clark, L. A. (1999). *The PANAS-X: Manual for the Positive and Negative Affect Schedule—Expanded Form*. Retrieved May 26,

- 2006 from University of Iowa, Department of Psychology Web site: www.psychology.uiowa.edu/faculty/watson/watson.html
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of Positive and Negative Affect: The PANAS Scales. *Journal of Personality and Social Psychology, 54*, 1063–1070.
- Watson, D., Hubbard, B., & Wiese, D. (2000a). General traits of personality and affectivity as predictors of satisfaction in intimate relationships: Evidence from self- and partner-ratings. *Journal of Personality, 68*, 413–449.
- Watson, D., Hubbard, B., & Wiese, D. (2000b). Self–other agreement in personality and affectivity: The role of acquaintanceship, trait visibility, and assumed similarity. *Journal of Personality and Social Psychology, 78*, 546–558.
- Watson, D., Klohnen, E. C., Casillas, A., Simms, E. N., Haig, J., & Berry, D. S. (2004). Match makers and deal breakers: Analyses of assortative mating in newlywed couples. *Journal of Personality, 72*, 1029–1068.

Received September 12, 2005

Revision received May 26, 2006

Accepted June 1, 2006 ■