

The Counterpoint of Personality Assessment: Self Reports and Observer Ratings

Robert R. McCrae

Assessment 1994 1: 159

DOI: 10.1177/1073191194001002006

The online version of this article can be found at:

<http://asm.sagepub.com/content/1/2/159>

Published by:



<http://www.sagepublications.com>

Additional services and information for *Assessment* can be found at:

Email Alerts: <http://asm.sagepub.com/cgi/alerts>

Subscriptions: <http://asm.sagepub.com/subscriptions>

Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>

Citations: <http://asm.sagepub.com/content/1/2/159.refs.html>

>> [Version of Record](#) - Jun 1, 1994

[What is This?](#)

THE COUNTERPOINT OF PERSONALITY ASSESSMENT: SELF-REPORTS AND OBSERVER RATINGS

Robert R. McCrae
Gerontology Research Center
National Institute on Aging, NIH
Baltimore, MD

In musical counterpoint, two or more different voices sound together in ways that illuminate the possibilities of the musical theme. Similarly, self-reports and observer ratings together may tell us more about personality than either could separately. This article reviews convergence between these two data sources and illustrates ways in which they can be used jointly in personality research to rule out response biases and estimate stability and heritability of personality traits. A measure of profile agreement is introduced and a rule of thumb is proposed for distinguishing agreement from disagreement on specific traits in individual cases. Averaging may be useful when there is agreement about individuals across observers; when there is disagreement, further information should be sought to resolve the competing hypotheses offered by the discrepant ratings.

In musical counterpoint, two or more different musical voices sound together. Sometimes the voices parallel or imitate each other, sometimes they move apart in contrary motion, sometimes they pit one theme against another. But instead of discord and cacophony, the result is a harmonious unity that explores and illuminates the musical possibilities of the themes. In this article I argue that it might be useful to consider self-reports and observer ratings as different voices that together can tell us more about personality than either could separately.

This may seem to be a fairly obvious approach to personality assessment. Psychologists are accustomed to the idea that different observers may have different opinions about an individual's personality traits and have known for years that

aggregating raters increases the validity of trait scores (Cheek, 1982). But, aggregation, in a sense, is the opposite of what I wish to discuss here. Aggregation emphasizes the common theme across a set of voices and ignores the differences as error variance. An alternative is to take these differences seriously, to assume that different observers may in fact have valid things to say that we as personality assessors need to consider. A clinician who routinely uses both self-reports and spouse ratings (Mutén, 1991) phrases this by saying that both sets of scores are valid, even when they disagree markedly. It is the business of the clinician to figure out what the discrepancy means (E. Mutén, personal communication, June, 1990). Ozer (1989) made a similar point when he urged us to look at the trait-method unit: "Properties of test scores are not distorted by the method through which they are obtained; rather, the test score can only be understood through a consideration of both the trait and the method" (pp. 229-230).

The reason for this is that self-reports and observer ratings are not direct measures of personality traits. Leaving aside the familiar

This article is a version of an invited address presented at the meeting of the American Psychological Association, August 1991, in San Francisco. Lewis R. Goldberg chaired the presentation session.

Correspondence concerning this article should be addressed to Robert R. McCrae, Personality, Stress and Coping Section, Gerontology Research Center, 4940 Eastern Avenue, Baltimore, MD 21224.

problems of measurement validity—problems of self-presentation, random responding, halo, and so on—these assessments are at best accurate representations of how people see themselves and how they are seen by others. What we measure directly are self-concepts and social perceptions, and these are influenced by many factors other than the traits themselves. The self-concept regarding some traits can be affected by other traits of the individual. For example, the self-reports of a narcissist may exaggerate his or her positive attributes (John & Robins, 1994); depressed people accentuate their undesirable traits. Social perceptions are influenced by the relationship between rater and ratee—in what situations they interact, how well they get along, how intimate or superficial their acquaintance is.

We can, and for most purposes do, consider these influences to be error, but ideally a complete science of personality assessment would take them into account. Our theories should explain why individuals with certain traits perceive themselves in certain ways and why they appear to others as they do. Cattell and Digman proposed a systematic approach to this topic as long ago as 1964, but it seems fair to say that little progress has been made in the past 30 years. In this article I hope to revive interest in the topic. After a brief overview of what is now known about self-reports and observer ratings from studies of groups and some comments on how these data sources can be used in research, I turn to the interpretation of individual profiles (which requires some discussion of a measure of agreement or disagreement) and finally suggest a two-stage approach that might improve the assessment of individuals.

Self-Reports and Observer Ratings

Both self-reports and observer ratings have long-established traditions in personality assessment, but until recently (Albright, Kenny, & Malloy, 1988; Funder & Sneed, 1993; Watson & Clark, 1991) they operated more or less independently. Most researchers have used self-report questionnaires, both because they are convenient and because these researchers believe that individuals have more extensive knowledge of their history of behavior and more privileged access to their own inner feelings than do external observers (Osberg & Shrauger, 1990). By contrast, researchers who

prefer observer ratings see them as being more objective and less susceptible to distortions caused by defensiveness or self-presentational strategies (Funder, 1991). Given these differing views, it is not surprising that most researchers choose one or the other method rather than both.¹

Self-Reports as Proxy Ratings

An important exception to the generalization just made is the use of observer ratings as a criterion for the development of self-report scales. Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1943) scales were created by selecting self-report items that discriminated between groups defined by psychiatric diagnosis, and several California Psychological Inventory (CPI; Gough, 1957) scales were similarly based on peer or expert nominations.

These purely empirical scale-construction strategies imply an asymmetry between the two kinds of data: Observer ratings are taken as the gold standard, and related self-report responses serve essentially as their proxies. The self-report items need not have any obvious relevance to the construct being assessed—in fact, so-called subtle items are particularly prized (Megargee, 1972). Such a system was enormously appealing in an era when it was widely assumed that self-reports were inevitably distorted by conscious self-presentation or unconscious defense.

In the case of the MMPI, a further step was taken in an attempt to enhance the validity of self-reports: The responses of individuals who scored low on MMPI scales but were judged to be high in psychopathology were used to select items for the *K* scale, and *K*-scale scores were subsequently used to correct self-reports for presumed defensive bias (Meehl & Hathaway, 1946).

These early approaches to combining self-reports and observer ratings in the development of assessment instruments were not entirely successful. Subtle items are generally inferior to obvious items (Wrobel & Lachar, 1982), and *K* correction of MMPI scales reduces, rather than enhances,

¹A computer search of the PsycLIT database for the period from January 1987 to September 1992 showed 3,921 abstracts with the phrase “self-report(s)” or “self-rating(s)” and 720 articles with “expert,” “peer,” “spouse,” or “observer” and “rating(s).” However, only 184 articles had both, and only 55 of these (including 14 by the present author) appeared to deal with normal adult personality.

their validity, at least in nonclinical samples (McCrae et al., 1989). Observer ratings are fallible, and self-reports are not necessarily biased; both deserve serious scrutiny.

The Structure of Personality

Self- and observer perceptions of personality may differ not only in accuracy, but also in form. Block (1961), for example, designed the California Q-set as an instrument to be used by expert raters; he employed the CPI when he wanted to assess personality through self-report (Block, 1981). As Lanning and Gough (1991) have demonstrated, there are meaningful links between these two, but the confounding of instrument with observer greatly complicates the interpretation.

If personality as lived were radically different from personality as observed, these complications would be unavoidable. But there is reason to believe that there are important similarities between self-observations and external observations, beginning in a structure which is common to both: the five-factor model (Digman, 1990). This model of personality was first discovered in studies of trait adjectives in the natural language and was later recovered in questionnaire, Q-sort, and act-frequency measures (Botwin & Buss, 1989; McCrae & John, 1992). Initially, the model was seen as the structure of peer ratings (Tupes & Christal, 1961/1992), but classic studies by Borgatta (1964), Fiske (1949), and Norman and Goldberg (1966) also found these same factors in self-reports.

By now there have been dozens of studies using different instruments that show parallel structures for self-reports and ratings (e.g., Goldberg, 1990). Consider one example. The Revised NEO Personality Inventory (NEO PI-R; Costa & McCrae, 1992a) measures the five factors by summing together six specific traits, or facets, for each of the five domains. Form S of the NEO PI-R is used for self-reports; Form R contains the same items rephrased in the third person for observer ratings. In the development of the test, item selection was based solely on self-report data. Table 1 shows the structural parallelism of self-reports and spouse ratings on this instrument. The self-reports are from a sample of 1,539 men and women in an employment study (Costa, McCrae, & Dye, 1991); the ratings are from 91 spouses of the Baltimore Longitudinal Study of Aging

(BLSA; Shock et al., 1984) participants. Despite the small size of the spouse rating sample relative to the number of variables, the expected solution emerges, with both primary and secondary loadings very similar to those found in the large self-report sample. Coefficients of factor congruence range from .91 to .97. The same kind of parallelism is found in studies of adjectives (e.g., McCrae & Costa, 1987), and replications by Ostendorf (1990) in German and Yang and Bond (1990) in Chinese suggest that this finding may be valid cross-culturally.

One implication of this structural parallelism is that it is meaningful to examine parallel measures in which the same items and scales are used for both self-assessments and observer ratings. That strategy has been adopted explicitly in the design of Forms S and R of the NEO PI-R, and it has also been used in adjective studies and occasionally in research on other questionnaires (e.g., Paunonen, cited in Jackson, 1984). It would appear to be appropriate to use both self-report and observer-rating versions of any scale in which responses are direct indicators of the underlying trait—where, for example, endorsement of “is neat and clean” is taken as evidence that the individual is indeed tidy. It would probably not be appropriate for instruments that treat responses as indirect signs of personality variables. Spouse ratings of how a patient would respond to a Rorschach test are not likely to tell a clinician much about the patient’s personality.² In the remainder of this article, attention will be restricted to personality assessments in which parallel forms for self-reports and observer ratings are or could be used.

Agreement across observers

Given the similarity of factor structure in self-reports and observer ratings, the next question is how well ratings on these common dimensions agree. This has been the topic of considerable research in the past decade, and the data are consistent although their interpretation is still debated. In general, when parallel forms with reliable and valid scales are used, correlations

²Some questionnaires, like the CPI, are conceptualized as measuring indirect signs of personality variables (Gough, 1968), although their scales often behave like ordinary trait measures (McCrae, Costa, & Piedmont, 1993). Whether observer rating forms of such instruments would prove to be parallel to the standard self-report versions is an empirical question.

Table 1
Factor Analysis of NEO PI-R Scales in Self-Reports and Spouse Ratings

NEO PI-R facet	Varimax rotated principal component									
	N		E		O		A		C	
	Self	Spouse	Self	Spouse	Self	Spouse	Self	Spouse	Self	Spouse
N1: Anxiety	82	85								
N2: Angry Hostility	68	61					-46	-60		
N3: Depression	80	81								
N4: Self-Consciousness	72	70								
N5: Impulsiveness	55	51								
N6: Vulnerability	70	70							-40	-46
E1: Warmth			74	69				44		
E2: Gregariousness			72	81						
E3: Assertiveness			48	42					40	
E4: Activity			51	45					48	46
E5: Excitement Seeking			57	47						
E6: Positive Emotions			73	69						
O1: Fantasy				45	60	53				
O2: Aesthetics					76	67				
O3: Feelings	41	40		57	52					
O4: Actions					60	51				
O5: Ideas					76	80				
O6: Values					54	66				
A1: Trust							49	73		
A2: Straightforwardness							70	79		
A3: Altruism			48	42			59	64		
A4: Compliance							74	77		
A5: Modesty							59	72		
A6: Tender-Mindedness							61	68		
C1: Competence									62	72
C2: Order									69	68
C3: Dutifulness									69	77
C4: Achievement Striving									76	78
C5: Self-Discipline									74	75
C6: Deliberation									58	56

Note. $N = 1,539$ for self-reports, 91 for spouse ratings. All loadings over .40 in absolute magnitude are shown. Decimal points are omitted. Adapted in part from Costa, McCrae, and Dye, 1991.

between single peer observations and self-reports tend to be in the .3 to .5 range; correlations between .5 and .7 are not uncommon when spouse ratings are used in place of peer ratings, or when three or four peer ratings are aggregated (Costa & McCrae, 1992a; McCrae & Costa, 1989a). Observer ratings are thus particularly useful criteria for the validation of self-report inventories (e.g., Gough, 1965; Jackson, 1984).

Cross-observer correlations routinely meet or exceed the .30 level that was once thought to represent an upper limit to validity coefficients in personality research, and in that respect they are

substantial. Funder (1989) discussed at length the many reasons why *disagreement* between different observers is to be expected: They have different perspectives on the individual, different amounts and types of exposure, different response biases. Given these circumstances, the degree of agreement across observers is quite remarkable and suggests that both self-reports and observer ratings may provide valid and useful ways of assessing personality.

However, as Fiske (1978) would point out, the fact remains that different ratings show far from perfect agreement: Our different voices most

certainly do not sing in unison. For that reason, they are not strictly interchangeable, and they are often most valuable when employed together in the same research design.

Designs for Joint Use of Multiple Observers

To the extent that differences are due to random error of measurement, greater precision can be obtained by averaging multiple ratings or self-reports and observer ratings. This is the familiar benefit of aggregation. To the extent that differences are due to systematic biases, the use of different observers can often provide conceptual leverage. When scores from two self-report measures are correlated, substance and method are confounded. But when a self-report score is correlated with an observer rating score, shared method artifacts are unlikely, and any relations uncovered are unlikely to be spurious. Observer ratings thus provide powerful replications of self-report findings (e.g., McCrae & Costa, 1989b, 1991).

Estimating stability of personality

There are other, less obvious uses of data from multiple observers. In a 7-year longitudinal study of peer ratings (Costa & McCrae, 1992b), retest correlations for single peers ranged from .63 to .84 for the five NEO-PI domains; a 6-year longitudinal study of self-reports showed retest correlations ranging from .63 to .83 (Costa & McCrae, 1988). Both studies suggest substantial stability of personality, yet even these values systematically underestimate stability because they confound true change with error of measurement.

One solution to that problem would be to disattenuate the observed stability coefficients using estimates of the retest reliability—a strategy that requires only a single data source, but at least three administrations of the instrument (Costa & McCrae, 1988; Costa, McCrae, & Arenberg, 1980). A potential problem with such a design is that it confounds stable method variance with stable trait variance. If an individual's self-concept were fully crystallized despite continuing changes in actual standing on personality traits, self-reports would probably suggest spuriously high retest stability (McCrae & Costa, 1982). Disattenuating for unreliability would only compound the problem. If we wish to minimize the effects of method variance and estimate the stability of the true score, we can

use path analysis and multiple methods: that is, ratings from different observers.

In Figure 1, “True N_1 ” represents true Neuroticism (N) scores for a group of target individuals at Time 1, and “True N_2 ” represents true scores on N some years later. The coefficient s represents the stability of true scores, which is the quantity we are trying to estimate. In this hypothetical design we have spouse ratings at both times and peer ratings only at Time 2. The coefficient a represents the correlation between spouse ratings of N and true N scores on both occasions.³ This is the validity of the spouse ratings, but of course it is an unobserved variable. The coefficient b represents the correlation at Time 2 of true N with peer ratings of N, and it too is unknown.

We do know the concurrent correlation between the spouse ratings and the peer ratings at Time 2, represented here as y , and the cross-lagged correlation between spouse ratings at Time 1 and peer ratings at Time 2, represented as x . This model makes a number of simplifying assumptions—for example, that true scores on N are the only determinant of ratings of N shared by spouses and peers. But given those assumptions, the principles of path analysis tell us that $x = asb$ and $y = ab$. If we divide x by y , we estimate s . In other words, the concurrent correlation between observers sets an upper limit to the agreement we can expect to see, and the closer the cross-time correlation across observers is to this concurrent correlation, the more stable the trait.

This kind of analysis can be used with self-reports and ratings, or with different peer raters. In our longitudinal peer rating study (Costa & McCrae, 1992b) we used concurrent and cross-lagged intra-class correlation coefficients and found estimated true-score stabilities ranging from .86 to 1.00+ for the five NEO-PI domains. Peers tended to agree with each other almost as much when their ratings were separated by 7 years as when they were gathered at the same time, suggesting that personality traits themselves must have changed very little over those 7 years.

³Empirical data on self/spouse agreement on two occasions 6 years apart support the assumption that the validity of spouse ratings is essentially constant, at least over this interval (Costa & McCrae, 1988).

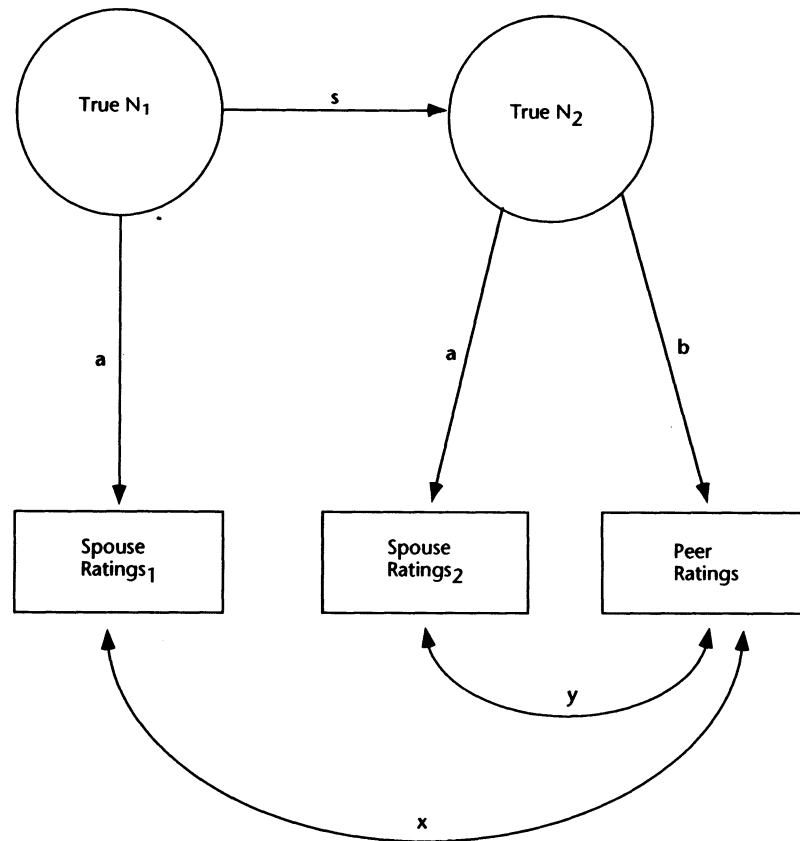


Figure 1. Path diagram for estimating the stability of true Neuroticism scores.

Estimating heritability

A variation on this design may be particularly valuable in studies of behavior genetics. In the usual twin study, some variance can be attributed to genetic influences and some to common environmental influences; the rest, by default, is considered *non-shared environmental influences* (e.g., Plomin & Daniels, 1987). It is tempting to interpret this residual variance substantively as the effect of unique formative influences on the individual, but any such influences are confounded with error of measurement. Correcting for the unreliability of measures reduces this confound, but not completely, again because some errors of measurement are systematic, not random. For example, a stable acquiescent response style could lead to highly reliable scores on a personality test, but at the expense of validity. If one twin had an acquiescent style whereas the other did not, the correlation between their scores and the resulting estimate of heritability would be attenuated.

The joint use of self-reports and observer ratings provides a way to control both random and systematic error, and thus to improve estimates of heritability (Costa & McCrae, 1987). If we substitute two groups of monozygotic twins reared apart for the two test administrations in Figure 1 and substitute self-reports for the spouse ratings, the design would allow us to estimate heritability. Correlations between observer ratings and self-reports of a twin would probably be about .50; correlations between peer ratings of that twin and the self-reports of the co-twin would be greater than zero only to the extent that the twins resembled each other. The ratio of the two correlations would estimate the heritability of the true score.⁴

A recent study (Heath, Neale, Kessler, Eaves, & Kendler, 1992) obtained self-reports from twins in addition to ratings of each twin by her co-twin. Those data could not be used in the design

⁴John C. Loehlin (personal communication, March 21, 1988) first pointed out to me that this concept could be expressed in terms of path analysis.

suggested above because the same individuals provided both self-reports and ratings, and response biases cannot be controlled. Indeed, the authors of the study interpreted some of the data as evidence of rater bias. But this study does demonstrate that it is feasible and informative to include observer ratings in behavior genetics studies.

Moderators of Self/Observer Agreement

Since Bem and Allen's (1974) famous article on predicting some of the people some of the time, there have been many attempts to increase agreement between raters by taking into account features of the raters, the target, or the trait. We know that higher agreement is seen when reliable measures are used, when raters are familiar with the target, and when multiple ratings are aggregated (Funder, 1989; McCrae & Costa, 1989a). But the more intriguing question is whether, with a given instrument and a given set of raters, the strength of the prediction can be increased by taking into account other features, such as the intra-item consistency of responses or the psychological-mindedness of the raters or the nature of the relationship. In some respects, this amounts to a study of Ozer's (1989) trait-method unit; if we knew, for example, that male raters systematically underestimated Openness to Experience in the individuals they rated, whereas female raters systematically overestimated it, we could introduce adjustments for the sex of the rater that should increase accuracy. Unfortunately, this goal has proven illusive.

Unpublished analyses illustrate the typical null findings. In the baseline study (McCrae & Costa, 1987) for our longitudinal peer ratings study, we measured a variety of characteristics of the raters and their relationship to the target, including age, education, and sex of rater, length of acquaintance, closeness of relationship, liking for target, frequency of social interaction, range of situations in which interactions occurred, perceived similarity to target, and self-reported accuracy in understanding others—a total of 32 variables. A matching strategy was used to examine possible moderator effects. For each target and each of five adjective factors, we selected the peer whose rating was most discrepant from the self-report (Low Accuracy) and the peer whose rating was least discrepant (High Accuracy). Paired *t*-tests were used to compare the two accuracy groups on each

of the 32 moderator variables. This is a relatively powerful design, because characteristics of the target, including validity of the self-report, are controlled. Subsamples of targets with 2, 3, or 4 raters (*N*s = 70, 86, and 58 targets, respectively) were examined separately to assess replicability.

Of the 480 *t*-tests, only 30 (6.25%) were significant at $p < .05$, and only two effects were replicated in two of the three subsamples: Raters who were more accurate with respect to Agreeableness were more likely to say that they were a friend of the family of the target; those who were more accurate with respect to Neuroticism were more likely to comment that they had interacted with the target more frequently in the past than they currently did. Neither of these effects is easily explained, and neither was replicated in the third subsample. Such findings discourage the search for moderators based on characteristics of the rater and relationship of rater to target. Within a pool of reasonably close acquaintances—the kinds of individuals likely to serve as personality raters in practical applications—differences in the type and length of acquaintance do not seem to affect accuracy.

Chaplin (1991) reviewed research on moderators based on characteristics of the self-report, including scalability and ipsatized variance, a moderator proposed by Bem and Allen (1974). He concluded that “moderator effect sizes in personality can be expected to correspond to a correlation of about .10” (p. 143). One interpretation of this finding would be that personality assessments from different judges consist partially of true score and otherwise merely of error that no moderator can explain. Yet 7-year retest correlations for peer ratings are about .70 for all five factors; different raters thus have very reliable views of the target's personality. Perhaps different observers have different views of targets for meaningful and potentially understandable reasons, but for reasons unique to each rater/target pair. If so, psychologists will not make much progress looking for general principles that explain discrepancies; instead, they must make assessments one case at a time. The real trait-method unit may not be “peer ratings of Conscientiousness” but rather “John Smith's rating of Mary Jones' Conscientiousness.” What is then needed is not a new moderator variable, but a new approach to the process of assessment itself.

Interpreting Individual Profiles

For decades, psychologists have been in the habit of plotting personality scores as a profile that graphically summarizes information from inventory responses. How should we interpret such a profile? Probably anyone who has gone to the trouble of developing a personality inventory, conducting validation studies, and writing a manual to help users interpret scores comes eventually to believe that the test profile presents true scores, what the person is really like. But anyone who has gone to the additional trouble of constructing and validating a parallel form of the test soon learns that profiles are never really parallel and that both cannot be literally true. This is illustrated in Figure 2, which shows the NEO PI-R profile of a BLSA participant born in 1917. The solid lines in Figure 2 represent her self-reports; the broken lines show a friend's rating of her. The five domain scores are given at the left of the profile; the 30 facets, grouped by

domain, continue to the right. There is substantial agreement for Neuroticism, Extraversion, and Openness domains, but there appear to be differences for Agreeableness and Conscientiousness.

The existence of differences does not mean that either the self-report or the rating form of the NEO PI-R is invalid; the same kinds of discrepancies would be seen with any good personality inventory. They result from properties of the observers, not the instruments. But these discrepancies do call attention to the need to treat *all* personality assessments with due caution.

The clinical importance of this truism is best illustrated by studies of behavioral confirmation (Snyder, 1992). Ideally, the results of psychological testing are used as a starting point in psychotherapy, an initial estimate of what the client is like that can be modified by new information. Unfortunately, clinicians are not immune to self-fulfilling prophecies; they may elicit only information that is consistent with their prior expectations

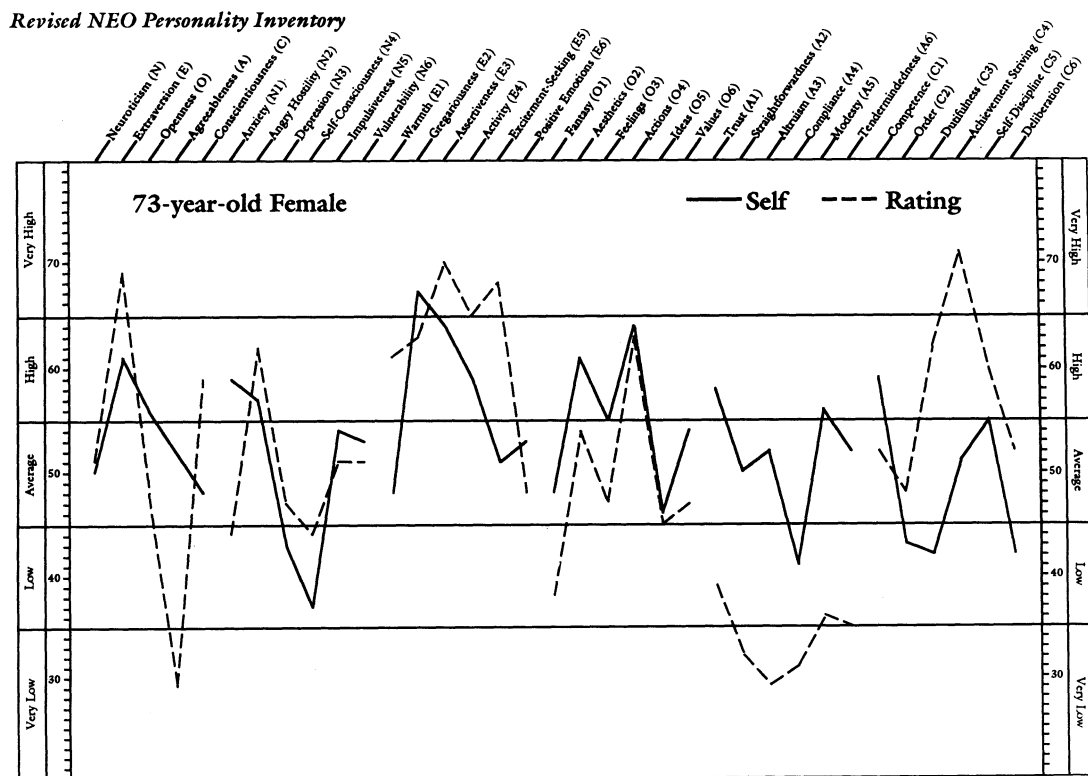


Figure 2. Revised NEO Personality Inventory profile for a 73-year-old female. Solid lines represent self-reports; broken lines represent a peer rating. Profile form reproduced by special permission of the publisher from the Revised NEO Personality Inventory. Copyright 1978, 1985, 1989, 1992 by Psychological Assessment Resources (PAR). Further reproduction is prohibited without permission of PAR.

(Wilson-Dallas & Baron, 1985), perpetuating misinformation that is likely to interfere with effective treatment. However, when both self-reports and observer ratings are employed, their points of agreement are probably accurate descriptions of the client's personality, and their points of disagreement are useful reminders to the clinician that personality profiles are tentative. Behavioral confirmation is unlikely to occur in such situations (Snyder, 1992).

A Metric for Agreement

Most clinicians have little experience interpreting profiles from multiple sources, although they frequently look for changes in a client's profile across time. In both cases, a similar interpretive problem arises: How does one know when a difference (or change) really is a difference? How large a discrepancy between two scores should be considered evidence of real change or of real disagreement between two sources? I would like to describe a general approach to assessing the degree of agreement of two sets of ratings (McCrae, 1993). This treatment omits the mathematical details but may convey the logic of the approach and illustrate how we might begin to think about interpreting self-reports and ratings.

In comparing profiles, the musical metaphor I have been using so far can be misleading. A musical theme is identifiable by its profile of intervals—rising and falling pitch—and if two voices have a similar progression of intervals, they will both produce the same theme, even if they start on different notes. This is why we recognize a tune regardless of the key in which it is played. By analogy, we might think that personality profiles are similar if they show the same pattern of relative highs and lows. This is the approach used in interpreting two-point codes for the MMPI and is closely related to the logic of Stephenson's (1953) Q-sorts. A Pearson correlation of self-report scores with ratings across the elements in the profile would be a simple measure of this kind of profile agreement.

But, as Cattell (1949) recognized long ago, this measure has many problems. Two profiles may show perfectly parallel patterns of ups and downs, yet describe dramatically different personality traits because the profiles have different elevations. Alternation between *high* and *very high* scores parallels alternation between

very low and *low* scores, but two such profiles could hardly be considered to be in agreement.

Cattell (1949) suggested an alternative measure, the pattern similarity coefficient r_p . This coefficient is based on the sum of the squared differences between the corresponding elements of two profiles and appears to be better than most other measures of profile similarity (Carroll & Field, 1974). There are, however, some limitations to this (and most other) measures of pattern similarity. In essence, because they are based on the *difference* between elements, they are really measures of *dissimilarity*. In clinical judgments of the similarity of self-reports and ratings, we are more often struck by the evidence of agreement. In Figure 2, for example, both sources agree that the individual is high in the Angry Hostility facet of Neuroticism and low in the Self-Consciousness facet. Instances of agreement are especially telling when the trait is extreme, because it is unlikely that an individual would receive two extreme ratings on the same trait by chance.

It is possible to take such agreement into account by considering the mean elevation of the two ratings on the trait, and a measure that incorporates both difference and extremeness of the mean was superior to Cattell's r_p in distinguishing matched from mismatched pairs of self-reports and ratings (McCrae, 1993).⁵ An overall *coefficient of profile agreement* across a set of traits, r_{pa} , might be interpreted as a measure of how well the observers know the target or how skilled they are in person-perception, and it has several potential uses in research on groups. For example, we might compare trained psychologists with laypersons as personality raters.

⁵The index of profile agreement is defined as:

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

where k is the number of traits in the profile, M is the mean of the two ratings for a trait, and d is the difference between the two ratings. The coefficient of profile agreement is a transformation of this index. The equation can also be expressed in terms of z scores for the two sets of ratings, and solving that equation for one rating in terms of the other produces the boundary of the nomograph given in Figure 3.

In considering individual cases, however, interest usually shifts from an estimate of overall agreement to agreement on specific traits. The value of r_{pa} for the domain scores in Figure 2 is .30, suggesting low-average agreement overall, but there is clearly more agreement on Extraversion than on Agreeableness. A related *index of profile agreement*, I_{pa} , can be used to estimate agreement on each trait in a profile separately and is particularly useful in understanding individual profiles.

Figure 3 depicts a nomograph that can be used to judge whether two estimates of standing on a single trait are in agreement. When the intersection of the self-report and observer rating falls within the shaded area, the value of I_{pa} is positive and the two scores can be considered to be in agreement. When the intersection is outside this shaded area, I_{pa} is negative and we can interpret the self-report and observer rating as being in

disagreement. Note that there is a wider latitude for difference when traits are extremely high or low than when traits are near average. When peer ratings and self-reports in BLSA participants are evaluated by this criterion, about 82% of the factor-score pairs show agreement (McCrae, 1993). This means that, on average, about one factor per case would be in disagreement.

A Two-Stage Strategy for Interpreting Profiles

Figure 3 also plots the data from the domain scores given in Figure 2. Four of the five domains lie within the shaded area and can be considered to show reasonable agreement. For these four scores, the best estimate of true standing on the trait might be obtained by averaging the two ratings, following the principle of aggregation. The woman whose profile is depicted in Figure 2 might be described at a global level as being average in Neuroticism,

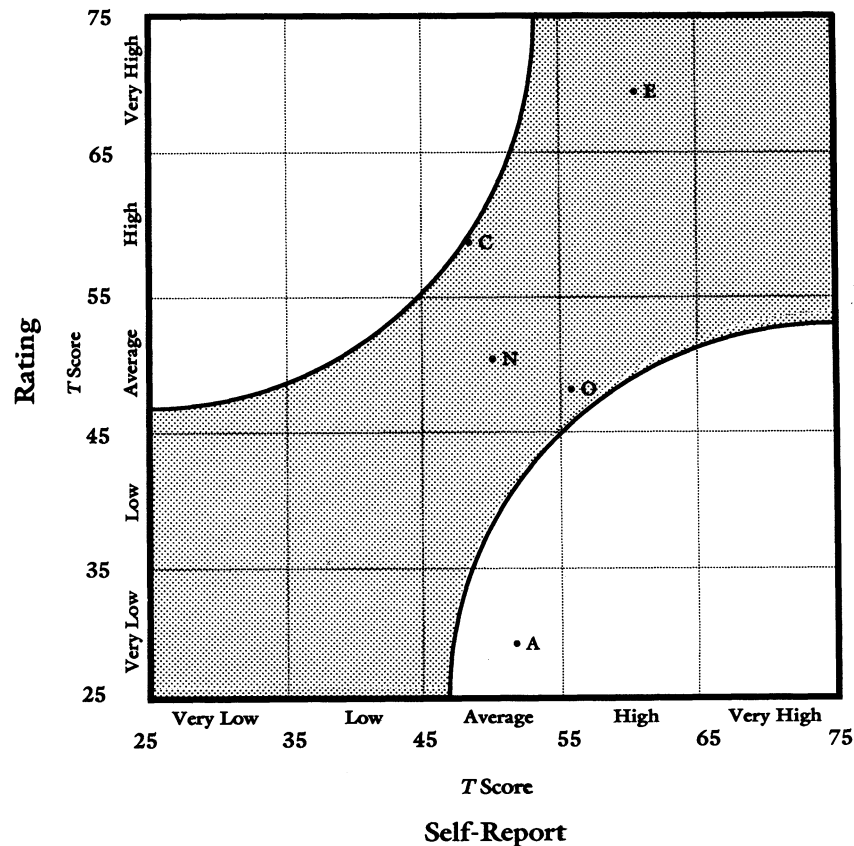


Figure 3. A nomograph for determining agreement or disagreement between self-reports and ratings. Data from the domain scores of the profile shown in Figure 2 are plotted. Adapted from McCrae (1993).

Openness, and Conscientiousness and high in Extraversion.⁶

When self-reports and ratings are clearly in disagreement, however, aggregation does not seem to be justified. Instead, the appropriate step may be to gather more data. There is, after all, no reason why the process of personality assessment has to stop when the individual closes the test booklet. Indeed, clinical assessment is often viewed as a hypothesis-testing process in which tentative interpretations are revised in light of new information. Again, we might consider this two-stage strategy a form of adaptive testing in which assessment continues until reliable scores are obtained.

Researchers who are accustomed to gathering data on a single occasion may find this prescription difficult to follow, although a new interactive computer program for administering, scoring, and reconciling Forms S and R of the NEO PI-R could facilitate the process for those who use that instrument. Because of the potential increase in validity, the extra effort involved may well be cost-effective, especially in longitudinal studies where baseline personality predictors are used repeatedly. Clinicians will normally have ample opportunity to gather new information after the initial assessment.

What other information should be considered in resolving the disagreements seen in Figure 2? There are many options. We could readminister the NEO PI-R to see if careless responding clouded the results. We might administer some alternative measure of the five factors, such as Goldberg's (1990), Wiggins' (Trapnell & Wiggins, 1990), or Hogan's (1986). Or we might ask a second rater to provide a third opinion.

In the case of the profile depicted in Figure 2, we have one other source of information. The individual in question is one of a small number of BLSA volunteers who wrote autobiographical sketches a few years ago. Given very general instructions to write about her adult life, this woman produced 14 handwritten pages that include some telling material.

On the NEO PI-R (Figure 2), this woman described herself as being trusting and modest, though relatively low in compliance; she saw herself as average in straightforwardness, altruism, and tender-mindedness, as well as in total Agreeableness. But that description does not

square well with the material in her autobiography, where the antagonism noted by the peer rater seems evident. As an adolescent she felt unappreciated at home so she "got a job as far away from home as possible and decided to fight for everything I wanted out of life." As a teacher, she didn't get along with her first principal and "felt like quitting...but jobs were scarce and I fought to stay." Years later, when this principal retired, "I was invited to attend and give a contribution toward his gift. I declined to do either. I told the committee that I didn't like him any better now than I did when I worked for him."

With regard to her married life she wrote that in the early years "We had plenty of battles. I can take things for a while; but I am volatile at times.... On occasion I did throw things at my husband."

She wrote that she likes people and has many acquaintances, and the peer rating of Extraversion as well as her election to leadership positions in social groups supports this assertion. But she also added that "One doesn't always like everybody he meets....I do not like people who are users. You can use me until I discover it. Then it ends. I also distrust people who are sweet and insincere. 'Sweet' and 'insincere' are synonymous to me."

Her own words suggest that this is indeed a disagreeable, defiant, and somewhat cynical individual. If so, we should not aggregate her self-report on Agreeableness with the peer rating; for this domain in this case we should accept the peer rating and reject the self-report.

An accurate assessment of personality traits, however, is only half the story. To understand our case fully, we must also know why she gave "average" responses to Agreeableness items. Is there some sense in which they are also valid? In general, what remains to be discovered is why self-reports and ratings are sometimes (apparently) inaccurate. Do respondents use different reference groups? Do they draw inferences based on behav-

⁶A more sophisticated procedure would be to calculate an adjusted mean. Because self-reports and ratings are imperfectly correlated, the standard deviation of means from pairs of raters will be somewhat less than the standard deviations of the original scores; an adjustment based on the normative correlation between self-reports and ratings could restore the initial *T*-score metric. In practice, this adjustment is usually small; however, such an adjustment would place the woman in Figure 2 in the very high, rather than high, range of Extraversion.

ior in different spheres of life, such as school and family (Achenbach, McConaughy, & Howell, 1987)? Are ratings biased by the nature of the rater's relationship to the target or by the rater's own personality? Exploratory research in which pairs of raters are confronted with discrepancies and asked to account for them could suggest answers to these questions.

Zimmerman, Pfohl, Coryell, Stangl, and Corenthal (1988) examined some of these issues in the context of diagnosing personality disorders. A structured interview was administered by one interviewer to a psychiatric patient and by a second interviewer to a family member or other informant. After both interviewers made independent diagnoses, the first interviewer listened to a tape of the informant interview and made a consensus diagnosis based on both interviews. For some disorders (notably schizoid, schizotypal, and avoidant), the consensus diagnosis resembled the diagnosis based on the patient interview; for others (including paranoid, histrionic, and narcissistic) it followed the lead of the informant interview. In this case there appears to be an interaction between the type and source of information, some of which is understandable: A real narcissist is probably not a good source of self-report data.

A Lesson from Musical History

I have argued here that we should make more—and more creative—use of multiple ratings of personality. We know that the same five-factor structure is found across observers, and that agreement between different knowledgeable observers on the standing of individuals on these five factors is sufficient to indicate that both self-reports and ratings include a substantial element of truth. Merely aggregating ratings does improve their validity, but much more can be done by comparing and contrasting different data sources.

A measure of profile agreement proposed here may enhance the assessment of individuals. Better measures can doubtless be developed and much more sophisticated ways of using them can be devised that will help us understand not only how each individual's personality should be characterized, but also how personality traits are represented in the self-concept, how they are viewed by others, how some traits influence the perception of other traits, and how personality affects rela-

tionships and relationships affect the perception of personality.

So complex and difficult are these questions that it is understandable that personality psychologists have made little progress on the problems posed by Cattell and Digman (1964). But the history of music has a lesson here. Polyphony, the use of two different voices in singing, began sometime around the year 1000 with experiments in contrary motion. Crocker (1966), in his *History of Musical Style*, asserted that the result was "a wealth of variety in interval progressions—an embarrassment of riches, in fact; composers had learned how to produce but not yet control a varied stream of intervals." Music theorists of that day had little to say about this new style of singing; as Crocker (1966) noted, "Throughout the 1000s and for most of the 1100s, polyphony was so inferior to contemporary [monophonic] chant in style and technique that it was hardly worth considering from an artistic point of view" (p. 61, 62).

But the musical innovators persisted—and with what results! It was precisely this step of learning to control multiple voices that led to the musical tradition that progressed from medieval canons and motets to the masses of Dufay and Palestrina, to the Baroque masterpieces of Vivaldi and Bach, the symphonies of Haydn and Mozart, Beethoven and Schubert, and the romantic glories of Chopin, Wagner, Brahms, Bizet, and Tchaikovsky. So difficult are the intricacies of music that it took 900 years to reach that point. The intricacies of personality assessment are equally formidable, but the rewards of understanding human nature are surely worth the effort.

References

- Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, *101*, 213-232.
- Albright, L., Kenny, D. A., & Malloy, T. E. (1988). Consensus in personality judgments at zero acquaintance. *Journal of Personality and Social Psychology*, *55*, 387-395.
- Bem, D. J., & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review*, *81*, 88-104.
- Block, J. (1961). *The Q-sort method in personality assessment and psychiatric research*. Springfield, IL: Charles C Thomas.

- Block, J. (1981). Some enduring and consequential structures of personality. In A. I. Rabin, J. Aronoff, A. M. Barclay, & R. A. Zucker (Eds.), *Further explorations in personality* (pp. 27-43). New York: Wiley-Interscience.
- Borgatta, E. F. (1964). The structure of personality characteristics. *Behavioral Science*, *9*, 8-17.
- Botwin, M. D., & Buss, D. M. (1989). The structure of act report data: Is the five-factor model recaptured? *Journal of Personality and Social Psychology*, *56*, 988-1001.
- Carroll, R. M., & Field, J. (1974). A comparison of the classification accuracy of profile similarity measures. *Multivariate Behavioral Research*, *9*, 373-380.
- Cattell, R. B. (1949). r_p and other coefficients of pattern similarity. *Psychometrika*, *14*, 279-298.
- Cattell, R. B., & Digman, J. M. (1964). A theory of the structure of perturbations in observer ratings and questionnaire data in personality research. *Behavioral Science*, *9*, 341-358.
- Chaplin, W. F. (1991). The next generation of moderator research in personality psychology. *Journal of Personality*, *59*, 143-178.
- Cheek, J. M. (1982). Aggregation, moderator variables, and the validity of personality tests: A peer rating study. *Journal of Personality and Social Psychology*, *43*, 1254-1269.
- Costa, P. T., Jr., & McCrae, R. R. (1987). On the need for longitudinal evidence and multiple measures in behavior-genetics studies of adult personality. *Behavioral and Brain Sciences*, *10*, 22-23.
- 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. (1992a). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.
- Costa, P. T., Jr., & McCrae, R. R. (1992b). Trait psychology comes of age. In T. B. Sonderegger (Ed.), *Nebraska Symposium on Motivation: Psychology and Aging* (pp. 169-204). Lincoln, NE: University of Nebraska Press.
- Costa, P. T., Jr., McCrae, R. R., & Arenberg, D. (1980). Enduring dispositions in adult males. *Journal of Personality and Social Psychology*, *38*, 793-800.
- Costa, P. T., Jr., McCrae, R. R., & Dye, D. A. (1991). Facet scales for agreeableness and conscientiousness: A revision of the NEO Personality Inventory. *Personality and Individual Differences*, *12*, 887-898.
- Crocker, R. L. (1966). *A history of musical style*. New York: Dover.
- Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, *41*, 417-440.
- Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *Journal of Abnormal and Social Psychology*, *44*, 329-344.
- Fiske, D. W. (1978). *Strategies for personality research*. San Francisco: Jossey-Bass.
- Funder, D. C. (1989). Accuracy in personality judgment and the dancing bear. In D. M. Buss & N. Cantor (Eds.), *Personality psychology: Recent trends and emerging directions* (pp. 210-223). New York: Springer-Verlag.
- Funder, D. C. (1991). Global traits: A Neo-Allportian approach to personality. *Psychological Science*, *2*, 31-39.
- Funder, D. C., & Sneed, C. D. (1993). Behavioral manifestations of personality: An ecological approach to judgmental accuracy. *Journal of Personality and Social Psychology*, *64*, 479-490.
- Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, *59*, 1216-1229.
- Gough, H. G. (1957). *California Psychological Inventory manual*. Palo Alto, CA: Consulting Psychologists Press.
- Gough, H. G. (1965). Conceptual analysis of psychological test scores and other diagnostic variables. *Journal of Abnormal Psychology*, *70*, 294-302.
- Gough, H. G. (1968). An interpreter's syllabus for the California Psychological Inventory. In P. McReynolds (Ed.), *Advances in psychological assessment* (Vol. 1, pp. 55-79). Palo Alto, CA: Science and Behavior Books.
- Hathaway, S. R., & McKinley, J. C. (1943). *The Minnesota Multiphasic Personality Inventory manual*. New York: The Psychological Corporation.
- Heath, A. C., Neale, M. C., Kessler, R. C., Eaves, L. J., & Kendler, K. S. (1992). Evidence for genetic influences on personality from self-reports and informant ratings. *Journal of Personality and Social Psychology*, *63*, 85-96.
- Hogan, R. (1986). *Hogan Personality Inventory manual*. Minneapolis: National Computer Systems.
- Jackson, D. N. (1984). *Personality Research Form manual* (3rd ed.). Port Huron, MI: Research Psychologists Press.
- John, O. P., & Robins, R. W. (1994). Accuracy and bias in self-perception: Individual differences in self-enhancement and narcissism. *Journal of Personality and Social Psychology*, *66*, 206-219.
- Lanning, K., & Gough, H. G. (1991). Shared variance in the California Psychological Inventory and the California Q-set. *Journal of Personality and Social Psychology*, *60*, 596-606.
- McCrae, R. R. (1993). Agreement of personality profiles across observers. *Multivariate Behavioral Research*, *28*, 13-28.
- McCrae, R. R., & Costa, P. T., Jr. (1982). Self-concept and the stability of personality: Cross-sectional comparisons of self-reports and ratings. *Journal of Personality and Social Psychology*, *43*, 1282-1292.
- McCrae, R. R., & Costa, P. T., Jr. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, *52*, 81-90.
- McCrae, R. R., & Costa, P. T., Jr. (1989a). Different points of view: Self-reports and ratings in the assessment of personality. In J. P. Forgas & M. J. Innes (Eds.), *Recent advances in social psychology: An international perspective* (pp. 429-439). Amsterdam: Elsevier Science Publishers.
- McCrae, R. R., & Costa, P. T., Jr. (1989b). Reinterpreting the Myers-Briggs Type Indicator from the perspective of the five-factor model of personality. *Journal of Personality*, *57*, 17-40.
- McCrae, R. R., & Costa, P. T., Jr. (1991). Adding Liebe und Arbeit: The full five-factor model and well-being. *Personality and Social Psychology Bulletin*, *17*, 227-232.
- McCrae, R. R., Costa, P. T., Jr., Dahlstrom, W. G., Barefoot, J. C., Siegler, I. C., & Williams, R. B., Jr. (1989). A caution on the use of the MMPI K-correction in research on psychosomatic medicine. *Psychosomatic Medicine*, *51*, 58-65.

- McCrae, R. R., Costa, P. T., Jr., & Piedmont, R. L. (1993). Folk concepts, natural language, and psychological constructs: The California Psychological Inventory and the five-factor model. *Journal of Personality, 61*, 1-26.
- McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60*, 175-215.
- Meehl, P. E., & Hathaway, S. R. (1946). The K factor as a suppressor variable in the MMPI. *Journal of Applied Psychology, 30*, 525-564.
- Megargee, E. I. (1972). *The California Psychological Inventory handbook*. San Francisco: Jossey-Bass.
- Mutén, E. (1991). Self-reports, spouse ratings, and psychophysiological assessment in a behavioral medicine program: An application of the five-factor model. *Journal of Personality Assessment, 57*, 449-464.
- Norman, W. T., & Goldberg, L. R. (1966). Raters, ratees, and randomness in personality structure. *Journal of Personality and Social Psychology, 4*, 681-691.
- Osberg, T. M., & Shrauger, J. S. (1990). The role of self-prediction in psychological assessment. In J. N. Butcher & C. D. Spielberger (Eds.), *Advances in personality assessment* (Vol. 8, pp. 97-120). Hillsdale, NJ: Lawrence Erlbaum.
- Ostendorf, F. (1990). *Sprache und Persönlichkeitsstruktur: Zur Validität des Fünf-Faktoren-Modells der Persönlichkeit* [Language and personality structure: Toward the validation of the five-factor model of personality]. Regensburg: S. Roderer Verlag.
- Ozer, D. J. (1989). Construct validity in personality assessment. In D. M. Buss & N. Cantor (Eds.), *Personality psychology: Recent trends and emerging directions* (pp. 224-234). New York: Springer-Verlag.
- Plomin, R., & Daniels, D. (1987). Why are children in the same family so different from one another? *Behavioral and Brain Sciences, 10*, 1-16.
- Shock, N. W., Greulich, R. C., Andres, R., Arenberg, D., Costa, P. T., Jr., Lakatta, E. G., & Tobin, J. D. (1984). *Normal human aging: The Baltimore Longitudinal Study of Aging* (NIH Publication No. 84-2450). Bethesda, MD: National Institutes of Health.
- Snyder, M. (1992). Motivational foundations of behavioral confirmation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 25, pp. 67-114). New York: Academic Press.
- Stephenson, W. (1953). *The study of behavior*. Chicago: University of Chicago.
- Trapnell, P. D., & Wiggins, J. S. (1990). Extension of the Interpersonal Adjective Scales to include the Big Five dimensions of personality. *Journal of Personality and Social Psychology, 59*, 781-790.
- Tupes, E. C., & Christal, R. E. (1992). Recurrent personality factors based on trait ratings. *Journal of Personality, 60*, 225-251. (Original work published 1961)
- Watson, D., & Clark, L. A. (1991). Self- versus peer ratings of specific emotional traits: Evidence of convergent and discriminant validity. *Journal of Personality and Social Psychology, 60*, 927-940.
- Wilson-Dallas, M. E., & Baron, R. S. (1985). Do psychotherapists use a confirmatory strategy during interviewing? *Journal of Social and Clinical Psychology, 3*, 106-122.
- Wrobel, T. A., & Lachar, D. (1982). Validity of the Wiener subtle and obvious scales for the MMPI: Another example of the importance of inventory-item content. *Journal of Consulting and Clinical Psychology, 50*, 469-470.
- Yang, K., & Bond, M. H. (1990). Exploring implicit personality theories with indigenous or imported constructs: The Chinese case. *Journal of Personality and Social Psychology, 58*, 1087-1095.
- Zimmerman, M., Pfohl, B., Coryell, W., Stangl, D., & Corenthal, C. (1988). Diagnosing personality disorder in depressed patients: A comparison of patient and informant interviews. *Archives of General Psychiatry, 45*, 733-737.