JOURNAL OF PERSONALITY ASSESSMENT, 1996, *67*(3), 558–578

The Rorschach and MMPI: Toward a More Scientifically Differentiated Understanding of Cross-Method Assessment

Gregory J. Meyer
Department of Psychology
University of Alaska Anchorage

Reasons for Minnesota Multiphasic Personality Inventory (MMPI) and Rorschach disagreement at the nomothetic level are explored. Building on an understanding of measurement distinctions from other sciences, it is proposed that the Rorschach and MMPI procedures are differentially sensitive to unique manifestations of personality. By necessity, each method is then also recognized as having a limited scope of effectiveness, such that neither can provide a complete picture of personality in its full complexity. Drawing on the more extensive self-report literature, the idiosyncracies and limited scope of the self-report method are documented. Finally, an ideographically rooted, cross-method, configural model is proposed for validation research in personality assessment. Several examples consistent with this approach are drawn from the assessment literature and discussed.

Archer and Krishnamurthy (1993a, 1993b) recently completed thorough reviews of the literature addressing Rorschach and Minnesota Multiphasic Personality Inventory (MMPI) relationships. These reviews demonstrate that scales derived from the MMPI and the Rorschach are essentially unrelated across heterogeneous samples. Even though the two procedures may frequently seem to be targeting the same underlying constructs (e.g., depression), there is generally a minimal association between these sources of information (also see Meyer, in press).

The primary question raised by these data is whether one or both tests are invalid. Some current sociopolitical conditions make this question more pressing and also more difficult to address. With third-party payers seeking to trim reimbursements wherever they can, with the movement in psychology to utilize only those procedures that have been "empirically validated," and with some scientists in our field decrying the whole clinical enterprise and holding a particularly vehement condemnation for the Rorschach (Dawes, 1994), it can be a challenge to address the apparent "invalidity" of these two assessment procedures in a calm and scientific manner. However, if personality assessment is to prosper, the issue must be addressed directly.

WHY SHOULD ASSESSMENT METHODS DISAGREE?

A starting premise for this discussion is that all measures of personality are *invalid* if they are expected to consistently measure personality in its full scope and complexity. Stated differently, any state, trait, or process that is operationally defined by a single method of assessment is necessarily invalid if it is assumed that the state, trait, or process being measured is a complete picture of that construct across people.

Framed in such absolute terms, most psychologists would readily agree that no measure is this comprehensive. However, assessment terminology is such that method-specific scores are given undifferentiated names that imply they measure the full scope of a construct. For instance, both the Rorschach and MMPI have scales carrying global and equivalent labels for *depression, schizophrenia,* and *obsessiveness.* As a result, it is easy for an expectation of comprehensiveness to creep into our thinking. It is also easy to take the next step and begin expecting variables drawn from very different methods to correlate—particularly when the scales share a common name. The lack of association between similarly named MMPI and Rorschach measures may then be taken as evidence that something is amiss. Thus, undifferentiated terminology can insidiously lead to the assumption that each method ought to measure the same consensual reality with a sufficient degree of precision to demonstrate some degree of empirical association across methods. Although this assumption may often seem reasonable, it may or may not be accurate depending on the construct being measured, the methods being employed, and the sample of participants being studied (Meyer, in press).

An illustrative digression into another branch of science may help solidify this point and provide a framework for the ensuing discussion. Astrophysicists recognize that different methods of measurement, such as optical, infrared, and ultraviolet telescopes, are incomplete tools that do not uniformly or completely measure a single consensual reality. When excited in the laboratory, all gasses emit a rather similar pale blue glow in the visible light spectrum.[1] Thus, hydrogen and helium

[1] My thanks to Donald Martin, the chair of the physics and astronomy department at the University of Alaska Anchorage, for his advice on these matters. Inaccuracies that may remain reflect my limitations, not his efforts to educate me.

can be indistinguishable in the visible light spectrum using summated optical measurements. Hydrogen, however, may be "seen" quite clearly with a measure that is differentially sensitive to the infrared spectrum because hydrogen has strong, characteristic emissions at this wavelength. At the same time, hydrogen emits no radiation in the ultraviolet region of the spectrum. Thus, if one was using an ultraviolet method of measurement, hydrogen would never be evident. But this does not make ultraviolet readings meaningless. On the contrary, an ultraviolet measure can readily identify a gas like helium, as helium has strong, characteristic emissions at such wavelengths.

Unlike psychologists, astrophysicists are not interested in the direct correlation between readings obtained from ultraviolet and infrared methods. In fact, they would be puzzled by a suggestion to do so. If the "construct" being measured was hydrogen, a nomothetic correlation of readings would consistently indicate a null relationship because the infrared method would regularly detect hydrogen whereas the ultraviolet method would not. Thus, the two methods would share no overlap with respect to consensual reality at a nomothetic level.[2]

Despite the fact that infrared and ultraviolet methods "disagree," astronomers do not consider one method to be less valid than the other. Rather, they recognize their strengths and limitations. Astronomers know each method is biased by particular forms of interference (e.g., object distance, atmospheric water vapor, temperature, etc.). They also know each method is differentially sensitive to particular wavelengths of radiation. Finally, they know the gasses they study have distinct spectral signatures, emitting large quantities of radiation at characteristic wavelengths. As long as astronomers keep the limitations and confounds associated with each type of telescope (i.e., method variance) in mind, they are able to use the "disagreements" between infrared, optical, and ultraviolet readings to gain a more thorough understanding of the cosmological universe.

Although there is not a one-to-one correspondence, I believe personality assessment would benefit from adopting the core elements of the astrophysicist approach to assessment. Recognizing the limitations and biases associated with each personality assessment method and understanding the ways in which each method is sensitive to different external realities would help to understand how two or more of these incomplete and imperfect tools can be used together to gain a more accurate

[2] This example is probably not really at the proper level of abstraction. Astrophysicists have the conceptual framework and terminology to differentiate elements like hydrogen and helium. As already mentioned, psychology has the habit of applying an equivalent global term to the distinct things that are measured by different assessment methods. Thus, we think of a global construct like "depression" rather than recognize the distinct depressive manifestations within this broad category that are differentially measured by a Rorschach or an MMPI. Consequently, the text example may have more application to psychology if it was framed as a nomothetic effort to correlate infrared and ultraviolet readings for an undifferentiated category of "elements" rather than for a specific element like hydrogen.

picture of clinical phenomena. Accomplishing this, however, requires that several specific pieces of knowledge be considered during any observation. If astronomers simply had the vague notion that ultraviolet readings were somehow different from infrared readings, then disparities between these sources of information would not reveal anything about the complex natural conditions they wished to measure.

The field of personality assessment has not yet evolved a sufficiently differentiated understanding of method sensitivities, method biases, or the distinct "spectral signatures" associated with the many clinical conditions we study. Method-specific confounds continue to be ignored or minimized (see Meyer, in press); we struggle with vague and global notions regarding the limitations and differential sensitivities of our measurement tools, and, as any personality textbook will demonstrate, there is no consensus about how the complex phenomena of personality should be cohesively organized into a dynamic and descriptive taxonomy—much less consensus about how different types of conditions should display unique patterns of scores across methods of measurement.

Further confounding progress in this area, there is a sentiment among some influential psychologists that the self-report method of assessment is sufficiently valid and cost effective such that it is unnecessary to research, teach, or use alternative methods of assessment (see Kagan, 1988, for the roots of this sentiment and a general critique). If this supposition were true, then there would be little need to take any additional steps toward defining a complex cross-method taxonomy like that used in astronomy. As such, it is critical to review evidence on the limits and differential sensitivity of the self-report method.

The thesis of this selective review is not that self-report measures are "invalid." Rather, the thesis is that these tools have a particular domain of effectiveness and this domain does not constitute the full spectrum of personality. At the most basic level, self-report data convey what patients know of themselves and what they wish to share with an evaluator. This is rarely (if ever) equivalent to the full spectrum of personality.

THE IDIOSYNCRACY OF THE SELF-REPORT METHOD: GENERAL REVIEWS

I am aware of two reviews addressing the general adequacy of self-report data. The first is a narrative summary comparing the accuracy of self-ratings to the accuracy of ratings generated by "other" sources of information (Shrauger & Osberg, 1981). The authors conclude that self-assessments are at least as accurate as "other" ratings when predicting various criteria. Unfortunately, effect sizes were not calculated to quantify the magnitude of this accuracy. More problematic, however, was the authors' atypical definition of "self-ratings" and "other ratings." In 19 of 46 studies, a self-rating instrument was classified as an "other" source of information. Thus, in more than 40% of the comparisons, one self-report measure was contrasted with another self-report measure, rather than to some truly alternative source of information. Despite this serious methodological confound, the authors reported that 18 of these 19 studies supported their hypothesis that "self-reports" would be more accurate than "other-reports." An additional confound emerged in 11 studies because self-report data were used as the criterion measure. Given the lack of independence between predictor and criterion self-ratings, the only surprising finding was that the self-rated predictors were not unanimously superior.

The second review was more methodologically sophisticated. Mabe and West (1982) conducted a meta-analysis examining the effectiveness of self-rated ability to predict criterion measures of these abilities. The 55 studies they reviewed provided 267 validity coefficients for aggregation. On average, self-rated abilities had only mild associations with external criteria of those abilities, with a mean r = .29 across various skill domains. Nonetheless, effect sizes varied widely depending on the skill being rated. As Table 1 indicates, mentally healthy individuals have some capacity to rate themselves on broad constructs for which they regularly receive environmental feedback, such as general intelligence (mean r = .34), scholastic achievement (mean r = .42), and athletic ability (mean r = .47). However, mentally healthy people are much less adept when it comes to rating themselves on qualities like interpersonal skills (mean r = .17) or managerial ability (mean r = .04).

TABLE 1
Mabe and West's (1982) Meta-Analytic Results on the Association Between Self-Ratings and Performance Criteria

Criterion	*Average Correlation*	*Number of Correlations*
Athletic skill	.47	3
Scholastic ability	.42	71
General intelligence	.34	12
Clerical skill	.33	33
Skilled technical abilities	.33	36
Rehabilitation job skills	.29	7
Job interview skills	.28	3
General mechanical skills	.20	5
Medical skills	.17	42
Interpersonal skills	.17	12
Managerial Ability	.04	41

Note. Number of correlations refers to the number of correlations used to compute the average. The original article did not report the number of subjects that contributed to each correlation. From "Validity of Self-Evaluation of Ability: A Review and Meta-Analysis," by P. A. Mabe, III and S. G. West, 1992, *Journal of Applied Psychology, 67,* pp. 280–296. Copyright 1992 by the American Psychological Association, Inc. Adapted with permission.

THE IDIOSYNCRACY OF THE SELF-REPORT METHOD: SPECIFIC COGNITIVE ABILITIES

Using the Intellectual Efficiency scale from the California Psychological Inventory, Gough (1987, p. 106) presented data from relatively large samples (*N* = 100 to 2,200) that parallel the Mabe and West findings. He reported correlations with achievement tests that are in the range of .35 to .45, and correlations with measures of general intelligence or aptitude that are in the range of .20 to .40. However, the relationship between self-ratings and more focused cognitive abilities are lower and generally in the range between .10 to .20. This is not surprising when it is recognized that people typically receive little direct feedback about their standing on these skills.

Christopher Pierce, Jeff Tysinger, and I recently examined how well 99 college students could judge their performance on a number of specific cognitive and attentional tasks. The tasks included the Digit Span-Forward, Digit Span-Backward, and Digit Symbol subtests from the Wechsler Adult Intelligence Scale–Revised (WAIS–R; Wechsler, 1981), Parts A and B of the Trail Making Test (Reitan & Wolfson, 1985), the Judgement of Line Orientation Test (Benton, Sivan, Hamsher, Varney, & Spreen, 1994), the d2 letter cancellation task (Brickenkamp, 1981), and the Paced Auditory Serial Addition Test (Gronwall, 1977). Participants were asked to rate their abilities as accurately as possible relative to other people of the same age on a 5-point scale ranging from 0 (*very below average*) to 4 (*very above average*).

Two different strategies were used for generating the ratings. Initial ratings were made prior to completing the tasks on slightly abstracted descriptions of the test constructs. For instance, Digit Symbol ability was assessed with item: "Compared to other people my age, my ability to sustain attention, learn new visual symbols, and quickly reproduce these symbols by hand is ..." The second series of ratings were made after the participants had actually completed each of the eight tasks. The questions were also now less abstract and consisted of fairly precise descriptions of the tasks they had just completed. For instance, the question for Digit Symbol was:

> You are given nine unique symbols, each of which corresponds to a single number between 1 and 9. You are then given a long line of the numbers from 1 to 9 mixed up in a random order. Below each of these numbers is an empty box. As quickly as you can, you have to fill in the empty boxes by writing in the unique symbol that corresponds to each number. Compared to other people my age, I believe my ability would be ...

The average correlation between self-rated ability and actual performance was .099 using the first set of ratings and .115 using the second set of ratings. Thus, these college students had essentially no capacity to judge themselves on these kinds of skills despite being emotionally healthy, having average intelligence, and knowing their ratings would be "verified" by actual performance. Similar null findings have been reported by others using summated scales of self-rated attention compared to actual performance (e.g., Turner & Gilliland, 1977).

In a recent study specific to the MMPI, self-rated memory and attentional complaints were correlated with actual performance on memory and concentration tasks. Gass, Russell, and Hamilton (1990) used a sample of 70 patients with closed head trauma. They had patients complete the MMPI, Digit Span from the WAIS–R, and Russell's revision of the Wechsler Memory Scale, from which they obtained four scores. MMPI scales included: Organic Symptoms (*Org*), Mental Dullness (*D4*), Lack of Ego Mastery-Cognitive (*Sc3*), and a new scale termed the *Cognitive Complaint Index,* comprising the eight MMPI items that specifically ask about memory problems. Exclusion criteria ensured no invalid MMPI protocols were included in the analysis. Despite validity controls and the use of theoretically pertinent scales, there were no statistically significant correlations between the MMPI scales and actual cognitive performance. Collapsing across 20 correlations, the average validity coefficient was –.04 (range –.20 to .13). In a general review of the literature, Herrmann (1982) cautioned against using any self-report measure for the accurate identification of memory problems because of low validity coefficients.

THE IDIOSYNCRACY OF THE SELF-REPORT METHOD: PERSONALITY CHARACTERISTICS

Despite the low to null findings from the ability/cognitive performance domains, one could suppose the results may be different when it comes to the assessment of personality characteristics. However, after excluding those studies that capitalize on shared method variance by examining the relationship between one self-report measure and another self-report measure (Campbell & Fiske, 1959), a representative review of the literature does not provide any reason to suspect self-report instruments provide a comprehensive portrait of personality.

Self-Reports and Other "Projective" Test Scores

Historically, the study of motivation generated contentious debate because self-ratings of motivation were unrelated to TAT-based measures of motivation. In a recent meta-analysis of this issue, Spangler (1992) gathered data from 36 observations drawn from 2,785 participants. He found the average correlation between TAT and self-rated achievement motivation was .088. This degree of association appears typical of all TAT and self-rated motivational constructs and it is probably not very different from the value Archer and Krishnamurthy (1993a, 1993b) would have derived, had they generated an overall effect size for Rorschach and MMPI relationships.

Self-Reports and Observer-Ratings: Adult Psychiatric Samples

During the development of the third edition of the Millon Clinical Multiaxial Inventory (MCMI–III; Millon, 1994), Millon had clinicians rate the personality characteristics of their patients. To ensure the cross-method validity coefficients were as accurate as possible, Millon discarded MCMI–IIIs that may have been invalid and only included ratings from about 275 clinicians who had "moderate" to "total confidence" in the accuracy of their judgments.[3] Despite the quality of these data, the average correlation between 24 self-rated MCMI–III scales and clinician ratings on the same dimensions was only .185 (range = –.11 to .37). On average, slightly less than 3.5% of the variance was shared between the patient and clinician perspectives on symptomatology. Millon did not record the length of contact clinicians had with these patients and speculates the ratings may have been stronger if the data were restricted to clinicians who had treated patients for an extended period of time.

Using a sample of 552 adults being seen in psychotherapy, Hyler et al. (1988) conducted such an analysis. He and his colleagues obtained correlations between self-rated personality disorder characteristics (using the Personality Diagnostic Questionnaire [PDQ]) and clinician ratings on the same dimensions. On average, the clinicians had been treating their patients for 50 weeks. In this study, clinicians were asked to select one patient from their practice who had a "significant personality disturbance" and another patient who had no salient personality disturbance. This procedure should increase variance in both sets of ratings and magnify the resulting validity correlations (Hunter & Schmidt, 1990). The coefficients obtained by Hyler et al. (1989) were slightly higher than those found by Millon (1994). However, they still had an average value of only .33 (range .16 to .51).

Self-Reports and Observer Ratings: Adult Nonpatient Samples

One could argue that the MCMI and PDQ are relative newcomers on the personality assessment scene and that the data may be better for an established measure like the MMPI. Alternatively, one could argue that clinical samples have low cross-

[3] Millon did not provide the exact number of clinicians and patients included in this analysis. Consequently, this number was estimated by determining how many participants would be required to achieve the reported level of statistical significance.

method correlations because distorted self-perceptions can be a defining characteristic of people with psychiatric troubles. To address either argument, the nonpatient data generated during the MMPI restandardization can be considered. The MMPI–2 manual (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) provides validity correlations between MMPI–2 scales and criterion ratings made by spouses or live-in partners on the Katz Adjustment Rating Scale. Even under the relatively optimal conditions when nonpatients provide self-ratings they know will be compared to the ratings of their partner, the results are far from optimal.

The MMPI–2 manual does not indicate which MMPI–2 scales should be associated with which Katz scales. To develop a more precise fix on the scales that should display cross-method correspondence, Radhika Krishnamurthy, David Nichols, and I rated the extent of construct overlap between each MMPI–2 scale and each Katz factor. We based our ratings on "the extent to which constructs" measured by the MMPI–2 scales "should theoretically converge and diverge with certain kinds of extra-test constructs." Ratings were made on a 5-point scale that ranged from –2 (*strong construct correspondence; inverse relationship*) to 0 (*minimal or no construct correspondence*) to +2 (*strong construct correspondence; direct relationship*).

We agreed on the extent of construct overlap in 234 out of the 252 construct pairs (93%), using the criteria that two of us needed to be in exact agreement whereas the third had to be no more than one rating point away. Construct pairs that did not reach this level of agreement were excluded from the subsequent analyses. All coefficients were transformed to Fisher's z prior to analysis. To determine if the validity coefficients could be combined regardless of gender or type of MMPI–2 scale (basic, supplemental, content), a 2 × 3 analysis of variance (ANOVA) was conducted exploring the effects of these variables and their interaction. There was no main effect for gender ($F < .02$) or scale type, $F(2, 462) = 1.47, p = .23$), nor was there an interaction effect ($F < .07$). Consequently, validity coefficients were combined across these groups.

Coefficients for all construct pairs rated as having strong or moderate construct correspondence (disregarding the direction of the expected relationship) were given a positive sign and averaged together. Two procedures can be used to aggregate constructs that should be unrelated. The first is to calculate an average value, disregarding sign. However, this is an imperfect index because large magnitude relationships with opposing signs will cancel each other out. An alternative is to use the absolute value of the relationship, as this more accurately gauges the full association between predictor and criterion. However, it is also a biased estimate when associations should be zero because sampling error fluctuations about the expected value of zero will not be allowed to cancel each other out. Thus, the first procedure provides an underestimate of the true relationship when constructs should be independent, whereas the second procedure provides a slight overestimate. Nonetheless, both results set limits on the true parameter value.

Table 2 presents the findings. As expected, scales that should display strong construct correspondence are more highly correlated with criterion ratings than scales that should display moderate or minimal construct correspondence, $F(2, 465) = 133.14, p < .0001$; all means significantly different by post hoc Scheffé. However, the constructs that should be strongly related do not have correlations of a strong magnitude. Instead, only about 8% of the variance in self-ratings and partner ratings overlap. Thus, mentally healthy couples who live together do not have nearly the same impressions of each other with respect to depression, hostility, sociability, conformity, antisocial behaviors, or anxiety.

Gough (1987, pp. 25–29) presented coefficients of a similar magnitude with adult nonpatients. Self-ratings were made with the California Psychological Inventory while criteria consisted of the average ratings from a panel of at least four judges (a mode of 6 in one study and 10 in another). Gough did not directly match scale constructs across methods, and self-ratings were dichotomously split into type categories. Both of these procedures should decrease the magnitude of observed correlations. Nonetheless, it is worth noting that none of the 1,600 correlations he calculated exceeded a magnitude of |.26|.

Self-Reports and Observer-Ratings: Adolescent Psychiatric Samples

A low degree of cross-method correspondence was also found when normative data were collected for the adolescent form of the MMPI (MMPI–A; Butcher et al., 1992). Using two primarily inpatient clinical samples, one of which was comprised of 420 boys and the other of which was comprised of 293 girls, the MMPI–A manual presents data on the relationship between self-ratings and two sources of external criteria. The first consisted of parent ratings on the Child Behavior Checklist (CBCL). The second set of ratings were made by master's level psychologists,

TABLE 2
Convergent Validity Coefficients for MMPI–2 Self-Ratings and Partner Ratings at Three Levels of Theoretically Expected Concordance

Expected Concordance	*Observed Correlation*	*Number of Construct Pairs*
Minimal or none		
Raw score	.02	280
Absolute value	.12	280
Moderate	.15	134
Strong	.28	54

Note. $N = 1{,}655$. Data were derived from information obtained on the MMPI–2 normative sample (Butcher et al., 1989).

psychiatric nurses, or chemical-dependency counselors who worked with the adolescents in a psychiatric setting (either an inpatient unit or a day-treatment facility). Clinician ratings were made after 14 days of observation using the Devereux Adolescent Behavior Rating Scale (DAB). Prior to calculating convergent correlations, MMPI–A profiles were eliminated on the basis of stringent validity criteria that resulted in the elimination of 119 (14%) of the original sample and left the number of boys and girls listed earlier.

As with the MMPI–2, the MMPI–A manual does not present hypotheses about which MMPI scales should be associated with which criterion ratings. Consequently, Krishnamurthy, Nichols, and I again organized the MMPI–A data by determining the extent to which MMPI–A constructs should be associated with DAB and CBCL constructs. Using the procedures described earlier, we agreed on the extent of construct overlap in 339 out of 416 construct pairs (81%). Only the agreed upon construct pairs were included in the analyses.

The MMPI–A data are slightly confounded because the manual does not report all of the validity coefficients. Rather, it reports those that were statistically significant at a Bonferroni corrected level. Conservatively protecting Type I errors in this fashion presumes that those coefficients not reaching statistical significance are actually zero. Unfortunately, such a procedure also results in an increased number of Type II errors and makes it more difficult to determine parameter values for a theoretically driven analysis like ours. Because we were forced to assign a validity coefficient of 0.0 to all unreported construct pairs, including some that may have fallen just below the criteria for statistical significance, the average values presented here may be slight underestimates of construct correspondence.

As before, coefficients were transformed to Fisher's z prior to analysis. To determine whether the coefficients could be aggregated across gender and source of the criterion rating (parent vs. clinician), a 2 × 2 ANOVA was run. This analysis indicated there was no interaction ($F < .07$) and no main effect for type of criterion measure, $F(1, 541) = .521, p = .47$. However, there was a difference between male and female participants, $F(1, 541) = 4.27, p = .039$. Although statistically significant, the absolute magnitude of this difference, 0.012, was trivial, so the coefficients were collapsed across gender and criterion rating source. Coefficients for all variable pairs rated as having strong or moderate construct correspondence (regardless of direction) were given a positive sign and averaged together. Raw coefficients and absolute values were used for those construct pairs expected to be independent.

Table 3 reports the findings from this analysis. Whether considering the raw coefficients or their absolute value, the means for each level of predicted association were significantly different from each other; for example, for the overall ANOVA using raw scores, $F(2, 542) = 27.28, p < .0005$. However, the magnitude of association for those constructs that should have been strongly associated across methods is remarkably small ($r = .0694$). The coefficient indicates less than 1% of the variance in personality traits is shared across methods. To be reported in the

TABLE 3
Convergent Validity Coefficients Between MMPI–A Self-Ratings and Parent or Clinician Ratings at Three Levels of Theoretically Expected Concordance

Expected Concordance	*Observed Correlation*	*Number of Construct Pairs*
Minimal or none		
Raw score	.00	262
Absolute value	.01	262
Moderate	.02	224
Strong	.07	59

Note. N = 713. Data were derived from information reported on the MMPI–A normative sample (Butcher et al., 1992).

MMPI–A manual, validity correlations had to achieve a minimum value of .18. Even if the wildly generous assumption was made that every unreported coefficient for a "strong" association just missed this cutoff and had a magnitude of .17, the parameter value would only increase to .19.

These data indicate that what adolescents in a clinical setting say about their anxiety, depression, physical problems, social withdrawal, psychotic problems, delinquency, and so on bears little resemblance to what their parents say about them or to what mental health clinicians say about them. These findings again strongly support the notion that, at a nomothetic level, self-ratings cannot be viewed as veridical reflections of a consensual reality.

Approaching this issue from a slightly different angle, Ehrenworth and Archer (1985) published an instructive study on the general accuracy of self-rated personality characteristics with adolescents. They asked therapists to rate the accuracy of statements derived from their patients' MMPI codetype after seeing the patients in treatment for no less than five sessions. Three different interpretive schemes were used to generate codetype descriptors. On the whole, the therapists of the 66 patients in this study rated the codetype descriptors as falling someplace between "Generally inaccurate and incomplete" and "Accurate in some respects, but contains sufficiently inaccurate material that it is of questionable utility."

Self-Reports and Observer-Ratings: Mixed Psychiatric/Nonpatient Child and Adolescent Samples

Achenbach, McConaughy, and Howell (1987) organized the most comprehensive database addressing the relationship between child and adolescent self-ratings and ratings made by others. They conducted a meta-analysis on this topic using 269 samples of data drawn from a total of 45,039 subjects. Table 4 summarizes portions of the Achenbach et al. findings. The self-ratings versus other ratings presented in the first four rows of the table are quite consistent with the literature discussed earlier. Self-ratings across a broad array of behavioral problems, emotional states, and personality characteristics share little overlap with the perceptions held by parents, teachers, peers, or clinicians.

EXCEPTIONS TO THE GENERAL PATTERN

On occasion, studies obtain higher validity coefficients for self-ratings than those reported earlier. These differences are partially a function of statistical procedures (i.e., degree of item or rater aggregation), the match between predictor and criterion constructs, methodological procedures, and intrinsic characteristics of the participants, the raters, and their relationships to each other (e.g., Achenbach et al., 1987; Christensen, Margolin, & Sullaway, 1992; Mabe & West, 1982; McCrae, 1994).

The most consistent exception to the general pattern emerges from the programmatic research conducted by Costa and McCrae (e.g., 1992), who regularly obtain higher validity coefficients. This holds for self-ratings compared to spouse-ratings, which produce correlations in the range between .35 to .65, and for self-ratings compared to peer-ratings, which produce correlations ranging from .25 to .50. Some portion of Costa and McCrae's success is probably due to the high reliability of their predictor and criterion scales, clear matching of cross-informant constructs, and the aggregation of raters whenever possible. However, their robust correlations

TABLE 4
Achenbach, McConaughy, and Howell's (1987) Data on Cross-Informant Convergent Validity

Informant Pair	*Average Correlation*	*Total Number of Subjects*	*Total Number of Samples*
Self and parent	.25	1,381	14
Self and teacher	.20	4,641	21
Self and mental health worker	.27	366	7
Self and aggregation of peers	.26	6,083	31
Parent and teacher	.27	12,853	41
Parent and mental health worker	.24	392	7
Teacher and mental health worker	.34	1,325	8

Note. From "Child/Adolescent Behavioral and Emotional Problems: Implications of Cross-Informant Correlations for Situational Specificity," by T. M. Achenbach, S. H. McConaughy, and C. T. Howell, 1987, *Psychological Bulletin, 101,* pp. 213–232. Copyright 1987 by the American Psychological Association, Inc. Adapted with permission.

are also likely a function of their participants and study procedures. Their core validation sample is a group of older, highly educated, nonpatients who are or were employed in high-functioning professional positions. Furthermore, each participant has volunteered to participate in an extensive longitudinal program of research. Each participant has also recruited his or her spouse and/or several friends to volunteer for the project. Finally, over the course of more than a decade, the participants have regularly completed numerous personality inventories, all of which they knew would be compared to the ratings made about them by others. Not only are these a select group of participants, but the cyclical nature of the research makes it likely that spouses and friends would discuss personality topics or specific perceptions of one another more than is typical (but see Piedmont, 1994).

Another situation that will generate higher than average cross-method correlations is when observer ratings derived from structured interviews are used as criteria. As Seligman (1995, p. 972) noted, self-reports form the "blood and guts" of interview-based judgments. Consequently, self-ratings and interview-derived ratings share considerable method variance. This is particularly the case when interviews are highly structured, as then the interview becomes little more than an "oral" self-report. Not surprisingly, when patients are given a paper-and-pencil task that asks about things like tearfulness, problems with anger, phobic fears, and physical complaints, and are then interviewed and orally asked the same kinds of questions, some of the highest "cross-method" validity coefficients can be obtained (e.g., Basham, 1992; Beck, Steer, & Garbin, 1988; Kobak, Reynolds, Rosenfeld, & Greist, 1990; Rosenfeld, Dar, Anderson, Kobak, & Greist, 1992).

IMPLICATIONS FOR RORSCHACH AND MMPI DISAGREEMENT

Even before reviewing all of this literature, most researchers would recognize the difficulty of trying to use self-ratings to validate the constructs measured by performance-based cognitive tests. For instance, most psychologists would not really expect answers to the question "How well can you flexibly shift your mental set while sequencing information in a visuo-motor task?" to be a suitable criterion for validating Part B of the Trail Making Test. If such criteria were employed, we would quickly have to conclude the cognitive measures were invalid.

Although it could be argued that patients know themselves better than anyone else and therefore should be in a position to validate these kinds of scores, it is generally accepted that: (a) these abilities do not have a high degree of representation in consciousness, or, to the extent that they do enter awareness; (b) patients have no sound yardstick by which to accurately judge themselves on these qualities against others; and (c) ratings are sure to be biased by motivational sets and other factors.

Similar reasoning is warranted when it comes to validating measures derived from a performance-based personality test like the Rorschach. One should not expect patients to be accurate judges of questions like: "To what extent do you have a proclivity to make hasty judgments without full consideration of all relevant information?" (i.e., *Zd* scores); or "How often do you misperceive environmental events under the press of angry or oppositional feelings?" (*S* – %); or "To what extent do you imbue other people with imagined or fantasized qualities?" (*H*:[*H*]+*Hd*+[*Hd*]); and so on.

At a minimum, patients can provide accurate self-ratings only when they are asked about qualities with high conscious penetration, when they are willing to be open and frank, when they are quite insightful and psychologically minded, and when they share a psychological yardstick that makes them capable of accurately describing themselves relative to other people. In general, the people for whom clinical tests need to be validated—patients with problematic clinical conditions—are unlikely to posses the requisite abilities. This is particularly true for Rorschach scores, because many of these variables address qualities that theoretically should be outside of conscious representation.

How then should we proceed with respect to the Rorschach and MMPI? Archer and Krishnamurthy (1993a, 1993b) and Meyer (in press) have established that Rorschach and MMPI constructs do not converge on a common universe of information in unrestricted heterogeneous samples. This finding is so robust that additional efforts to find cross-method correlates in heterogenous samples would be redundant. As such, it would be wise to invest research energy in other assessment questions.

A second step is to revise Rorschach and MMPI constructs. The data indicate Rorschach scores do not typically measure constructs that reside within conscious awareness. Thus, to refine understanding of the Rorschach's domain of differential sensitivity, it would help to strip scores of self-report connotations. To the extent that Rorschach constructs are not viewed as consistently tapping conscious and deliberately reported phenomena, clinical interpretations will be more accurate and research will be more appropriately focused on validation studies that use subtle behavioral criteria that do not depend on conscious awareness (also see Meyer, this issue).

Currently, there is enough research to conclude the Rorschach does not consistently or globally measure self-reported characteristics. However, there is relatively little theory and research that can be used as a guide for understanding what "level of personality" or type of information the procedure effectively taps. Clearly, Rorschach scores are not unrestricted MRIs of the psyche. However, the locus of the Rorschach's sensitivity has not yet been sufficiently demarcated and data are critically needed in this area.

Just as Rorschach scores should be trimmed of connotations that necessarily imply a direct representation in consciousness, MMPI scores should be stripped of connotations that necessarily imply a construct has been measured completely or thoroughly. A more scientifically defensible position is that these scores reflect self-knowledge as it has been filtered through whatever desires a patient may have to convey a particular message to the assessment clinician. MMPI scores may occasionally provide a fairly accurate and fairly complete picture of personality. However, the data suggest this would be true for only a small proportion of patients. In general, the MMPI does not provide a veridical and complete picture of personality functioning.

IMPLICATIONS FOR A MORE REFINED SCIENCE OF PERSONALITY ASSESSMENT

It bears repeating that my purpose for highlighting the intrinsic limitations of the self-report method was not to suggest that the method is invalid. There are many studies which document that self-rated characteristics, despite their limitations, can provide useful information. The data only indicate self-report scales are invalid if the method is conceived to be a complete and exhaustive way of measuring the full spectrum of personality. If researchers and clinicians hold this belief, implicitly or explicitly, the data clearly indicate such a conceptualization is wrong. Thus, self-report scales should not be seen as easy to administer, easy to score, cost-effective ways to obtain accurate information about personality in its full complexity (Kagan, 1988).

Furthermore, the data do not support the notion that self-rated personality characteristics can serve as screening tools to determine when more extensive assessment is warranted. To function as a screening tool, a test must have a high degree of specificity, or accuracy identifying a condition as "absent" when it is truly absent (e.g., Kraemer, 1992). Lower levels of sensitivity, or accuracy identifying a condition as "present" when it is truly present, are tolerated under the assumption that further testing will rule out false positive cases. Given that self-rated personality characteristics have little nomothetic association with spouse, teacher, peer, or clinician ratings, and virtually no association with scores obtained from other personality assessment methods, the data do not suggest self-report scales can function as effective screening tools simply because there is so much personality variance that remains unaccounted for by self-ratings.

With all this said, the question remains, how should a science of personality assessment proceed? Several pieces of data from within the personality assessment literature suggest the approach to assessment used in astronomy may be a viable model to emulate. The issue of cross-method agreement/disagreement has been dealt with fairly extensively in the assessment of psychopathology with children and adolescents because it is common to obtain ratings from sources other than just the patient. As a result, clinicians and researchers in this area have regularly confronted the realities presented in Table 4. What a youth says about himself or herself does not correspond to what others say about him or her. Furthermore, as the last three rows of Table 4 indicate, teachers, parents, and clinicians generally do not agree with each other either.

Importantly, Achenbach et al. (1987) and other clinical researchers do not believe the data in Table 4 reflect test invalidity. Rather, it is considered natural for different kinds of people, in different kinds of settings, who have different kinds of relationships with the target participant, to have different perspectives on the same core phenomena—be it somatic complaints, delinquent behaviors, or affective distress. In other words, each measurement perspective is understood as being differentially sensitive to certain clinical realities. At least for the CBCL, Achenbach (1994) used the nomothetic information presented in Table 4 to generate indices of typical cross-informant agreement. He then made ideographic determinations for a given patient that indicate whether different sources of information agree with each other more than usual or disagree with each other more than is typical.

Although it is believed that instances of atypical agreement can provide insight into the child being evaluated, Achenbach has not yet taken this line of reasoning as far as the astrophysicists. That is, he does not specify which conditions in nature should lead to particular patterns of cross-informant agreement. I believe the science of assessment would advance considerably if clinical researchers began to theorize and experiment in such a fashion. For instance, if the dynamics of delinquency could be spelled out sufficiently, one might anticipate that highly delinquent children have a narcissistic core. If so, their self-appraisals should be overly positive and this should cause an unusual extent of disagreement between their self-report and the more negative reports of their parents or teachers. At the same time, because the dynamics of delinquency lead to easily observed behavior across a range of situations, this should result in unusually high agreement between parents and teachers. Thus, the condition in nature could be identified by particular scale elevations in combination with a specific pattern of agreement across these multiple perspectives, just as the gas helium can be identified by its distinctive pattern of readings across infrared, optical, and ultraviolet telescopes.

In essence, this model takes the relatively common configural approach to MMPI interpretation (e.g., Archer, 1992; Greene, 1991) and extends it across assessment methods. Within this model, validation efforts would be required to shift away from traditional strategies that just focus on the correlates of scales from a single method or that use "unconfigured" data in multiple regression techniques as a way to assess incremental validity. These are nomothetic validation strategies that seek to validate scales rather than to identify uniquely configured people and they do not adequately address the ideographic considerations required by this model. Instead, validation would begin when an articulated theory about clinical phenomena generates expectations regarding within-method scale elevations as well as specific patterns of cross-method findings. Subsequently, research using a spectrum of measurements would determine whether the observed findings matched the expected heteromethod pattern.

Meyer (in press) successfully applied the rudiments of this approach in a study specifically designed to explore MMPI and Rorschach patterns, although the clinical value of MMPI and Rorschach discrepancies have been retrospectively identified in other studies as well (e.g., Ganellen, 1994; Meyer, 1993). Shedler, Mayman, and Manis (1993) also employed the rudiments of this approach in a series of studies using self-report scales and clinical judgments derived from earliest recollections. The authors demonstrate how people with a cross-method pattern of healthy self-ratings and disturbed earliest memories have much higher levels of physiological reactivity under stressful conditions. Thus, the differential sensitivities of these two assessment methods identify people with a form of "illusory" mental health who may be at risk for subsequent medical illness. A similar pattern of cross-method disparities (between self and other ratings) has been identified in children (Weinberger, 1996) and adults (Colvin, Block, & Funder, 1995), with the latter predicting adverse longitudinal outcomes.

However, the most differentiated and advanced theorizing about cross-method disparities can be found in the work of McClelland (1980, 1989; McClelland, Koestner, & Weinberger, 1989). McClelland believed TAT-based measures of motivation assess "implicit" aspects of personality, whereas self-report questionnaires assess self-attributed motivations. The distinctions between these constructs are many. Implicit motivations are viewed as being more unconscious and physiologically related, as developing earlier in life and not requiring language and verbal mediation to solidify, and as being more strongly associated with long-term spontaneous behavioral trends. In contrast, self-attributed motives are understood as having different historical antecedents and as being better predictors of conscious choices and immediate, situationally defined behaviors. Perhaps most important, McClelland articulated expectations that identify the kind of people who will produce specific cross-method patterns of data. Depending on the criterion variable and the environmental factors influencing behavior, these individuals will act in ways that could not be predicted directly from just one source of personality data. Like astrophysicists, McClelland has a specific theory about the differential sensitivities of each assessment tool. Furthermore, he believes the natural world is populated with complexly organized people who have dynamics that leave a distinctive pattern or signature across TAT and questionnaire scales of motivation. Cross-method disagreement is thus not a question of test invalidity. Rather, it is a phenomena that can lead to a more refined identification of people and more accurate behavioral predictions.

McClelland's hypotheses about TAT-derived measures of achievement motivation, self-reported motivation, and their interaction have been supported in a meta-analysis conducted by Spangler (1992) using 383 correlations drawn from 28,289 participants. Not only was TAT validity demonstrated, but TAT motives had somewhat larger validity coefficients than self-rated motives, particularly when predicting longer term, spontaneous behaviors. Importantly, the analysis also documented the expected cross-method interactions between TAT and self-rated motivations.

In my view, the sophisticated and differentiated theoretical landscape articulated by McClelland (1980, 1989; McClelland et al., 1989; Spangler, 1992) should be the destination we strive to reach as we embark on a scientific exploration into the combined use of the MMPI, Rorschach, and other assessment tools.

ACKNOWLEDGMENTS

The ideas in this article were initially organized for a round table discussion co-chaired by Radhika Krishnamurthy and Gregory J. Meyer at the annual convention of the Society for Personality Assessment, Denver, CO, March 9, 1996.

I thank Radhika Krishnamurthy and David Nichols for their very generous assistance with the MMPI data reported here. I would also like to extend my deep gratitude to Christopher Pierce and Jeff Tysinger who, in the course of their other research, collected the attentional ratings reported here.

REFERENCES

Achenbach, T. M. (1994). Child Behavior Checklist and related instruments. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (pp. 517–549). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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.

Archer, R. P. (1992). *MMPI–A: Assessing adolescent psychopathology.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Archer, R. P., & Krishnamurthy, R. (1993a). Combining the Rorschach and the MMPI in the assessment of adolescents. *Journal of Personality Assessment, 60,* 132–140.

Archer, R. P., & Krishnamurthy, R. (1993b). A review of MMPI and Rorschach interrelationships in adult samples. *Journal of Personality Assessment, 61,* 277–293.

Basham, R. B. (1992). Clinical utility of the MMPI research scales in the assessment of adolescent acting out behaviors. *Psychological Assessment, 4,* 483–492.

Beck, A. T., Steer, R. A., & Garbin, M. A. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review, 8,* 77–100.

Benton, A. L., Sivan, A. B., Hamsher, K. deS., Varney, N. R., & Spreen, O. (1994). *Contributions to neuropsychological assessment: A clinical manual* (2nd ed.). New York: Oxford University Press.

Brickenkamp, R. (1981). *Test d2: Aufmerksamkeits-Belastung-Test* (Handanweisung, 7th ed.) [Test d2: Concentration-Endurance Test (Manual, 7th ed.)]. Göttengin, Germany: Verlag.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Manual for the restandardized Minnesota Multiphasic Personality Inventory: MMPI–2. An administrative and interpretive guide.* Minneapolis: University of Minnesota Press.

Butcher, J. N., Williams, C. L., Graham, J. R., Archer, R. P., Tellegen, A., Ben-Porath, Y. S., & Kaemmer, B. (1992). *Manual for administration, scoring, and interpretation of the Minnesota*

Multiphasic Personality Inventory for Adolescents: MMPI–A. Minneapolis: University of Minnesota Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Christensen, A., Margolin, G., & Sullaway, M. (1992). Interpersonal agreement on child behavior problems. *Psychological Assessment, 4,* 419–425.

Colvin, C. R., Block, J., & Funder, D. C. (1995). Overly positive self-evaluations and personality: Negative implications for mental health. *Journal of Personality and Social Psychology, 68,* 1152–1162.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory: Professional manual.* Odessa, FL: Psychological Assessment Resources.

Dawes, R. M. (1994). *House of cards: Psychology and psychotherapy built on myth.* New York: Free Press.

Ehrenworth, N. V., & Archer, R. P. (1985). A comparison of clinical accuracy ratings of interpretive approaches for adolescent MMPI responses. *Journal of Personality Assessment, 49,* 413–421.

Ganellen, R. J. (1994). Attempting to conceal psychological disturbance: MMPI defensive response sets and the Rorschach. *Journal of Personality Assessment, 63,* 423–437.

Gass, C. S., Russell, E. W., & Hamilton, R. A. (1990). Accuracy of MMPI-based inferences regarding memory and concentration in closed-head-trauma patients. *Psychological Assessment, 2,* 175–178.

Gough, H. G. (1987). *California Psychological Inventory: Administrator's guide.* Palo Alto, CA: Consulting Psychologists Press.

Greene, R. L. (1991). *The MMPI–2/MMPI: An interpretive manual.* Boston: Allyn & Bacon.

Gronwall, D. (1977). Paced auditory serial-addition task: A measure of recovery from concussion. *Perceptual and Motor Skills, 44,* 367–373.

Herrmann, D. J. (1982). Know thy memory: The use of questionnaires to assess and study memory. *Psychological Bulletin, 92,* 434–452.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Hyler, S. E., Rieder, R. O., Williams, J. B. W., Spitzer, R. L., Lyons, M., & Hendler, J. (1989). A comparison of clinical and self-report diagnoses of DSM–III personality disorders in 552 patients. *Comprehensive Psychiatry, 30,* 170–178.

Kagan, J. (1988). The meaning of personality predicates. *American Psychologist, 43,* 614–620.

Kobak, K. A., Reynolds, W. M., Rosenfeld, R., & Greist, J. H. (1990). Development and validation of a computer-administered version of the Hamilton Depression Rating Scale. *Psychological Assessment, 2,* 56–63.

Kraemer, H. C. (1992). *Evaluating medical tests: Objective and quantitative guidelines.* Newbury Park, CA: Sage.

Mabe, P. A., III, & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67,* 434–452.

McClelland, D. C. (1980). Motive dispositions: The merits of operant and respondent measures. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 1, pp. 10–41). Beverly Hills: Sage.

McClelland, D. C. (1989). Motivational factors in health and illness. *American Psychologist, 44,* 675–683.

McClelland, D. C., Koestner, R., & Weinberger, J. (1989). How do self-attributed and implicit motives differ? *Psychological Review, 96,* 690–702.

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

Meyer, G. J. (1993). The impact of response frequency on Rorschach constellation indices and on their validity with diagnostic and MMPI–2 criteria. *Journal of Personality Assessment, 60,* 153–180.

Meyer, G. J. (in press). On the integration of personality assessment methods: The Rorschach and MMPI. *Journal of Personality Assessment.*

Millon, T. (1994). *Millon Clinical Multiaxial Inventory–III Manual.* Minneapolis: National Computer Systems.

Piedmont, R. L. (1994). Validation of the NEO–PI–R observer form for college students: Toward a paradigm for studying personality development. *Assessment, 1,* 259–268.

Reitan, R. M., & Wolfson, D. (1985). *The Halstead–Reitan neuropsychological test battery: Theory and clinical interpretation.* Tucson, AZ: Neuropsychology Press.

Rosenfeld, R., Dar, R., Anderson, D., Kobak, K. A., & Greist, J. H. (1992). A computer-administered version of the Yale–Brown Obsessive–Compulsive Scale. *Psychological Assessment, 4,* 329–332.

Seligman, M. E. P. (1995). The effectiveness of psychotherapy: The Consumer Reports study. *American Psychologist, 50,* 965–974.

Shedler, J., Mayman, M., & Manis, M. (1993). The illusion of mental health. *American Psychologist, 48,* 1117–1131.

Shrauger, J. S., & Osberg, T. M. (1981). The relative accuracy of self-predictions and judgements by others in psychological assessment. *Psychological Bulletin, 90,* 322–351.

Spangler, W. D. (1992). Validity of questionnaire and TAT measures of need for achievement: Two meta-analyses. *Psychological Bulletin, 112,* 140–154.

Turner, R. G., & Gilliland, L. (1977). Comparison of self-report and performance measures of attention. *Perceptual and Motor Skills, 45,* 409–410.

Wechsler, D. (1981). *WAIS–R manual: Wechsler Adult Intelligence Scale–Revised.* San Antonio, TX: Psychological Corporation.

Weinberger, D. A. (1996). Distorted self-perceptions: Divergent self-reports as statistical outliers in the multimethod assessment of children's social-emotional adjustment. *Journal of Personality Assessment, 66,* 126–143.

Gregory J. Meyer
Department of Psychology
University of Alaska Anchorage
3211 Providence Drive
Anchorage, AK 99508–8224

Received May 9, 1996