JOURNAL OF PERSONALITY ASSESSMENT, 1996, *67*(3), 598–628

Construct Validation of Scales Derived From the Rorschach Method: A Review of Issues and Introduction to the Rorschach Rating Scale

Gregory J. Meyer
University of Alaska Anchorage

I suggest the main goal of Rorschach validation should be a refined understanding of what each score means. Toward this end, I review general issues in construct validity, hurdles unique to the Rorschach, and general limitations with validation criteria. I then recommend two approaches for improving criteria so they can begin approximating the gold standards that are necessary for a refined understanding of what scores actually measure. The first is a method for improving expert clinical judgment, and the second is a method for aggregating data across diverse judges. Finally, the Rorschach Rating Scale (RRS) is presented as a criterion tool to be used with either of these approaches to validation. The RRS is a fairly comprehensive summary of the constructs thought to be measured by various Rorschach scoring systems. The utility of the RRS for research and training are discussed, as are other practical, theoretical, and psychometric issues in its application.

> So that where some kinds of psychological work require us to know only a little bit about a large number of persons, and other kinds of work require us to know a very great deal about a few persons, construct validation of tests ... will probably require that we know a great deal, and at a fairly intensive or "dynamic" level, about a large number of persons.
>
> Paul E. Meehl (1959, p. 123)

The recent history of the Rorschach has affirmed the reliability and general validity of this method of assessment (e.g., Parker, Hanson, & Hunsley, 1988). This is most notably the case for the Comprehensive System's so-called structural and atheoretical approach to scoring (Exner, 1991, 1993), although there have also been many developments with a number of more theoretically informed scoring systems, such

as the Ego Impairment Index (Perry & Viglione, 1991), Kwawer's (1980) scale to assess object relations (Gacono, Meloy, & Berg, 1992), the Lerner Defense Scales (Lerner & Lerner, 1980, 1982), the Mutuality of Autonomy Scale (Tuber, 1992; Urist, 1977), the Psychoanalytic Rorschach Profile (Burke, Friedman, & Gorlitz, 1988), the Rorschach Defense Scales (Cooper, Perry, & Arnow, 1988), and the Rorschach Oral Dependence Scale (Bornstein, 1993; Masling, Rabie, & Blondheim, 1967), among others. All of these scoring systems have shown promise as tests that can quantify clinically important phenomena using the Rorschach method.

Despite these advances, more data still needed to document Rorschach scores are actually quantifying the latent constructs they are believed to measure. This is a need facing all measures of personality and neuropsychological functioning, not just the Rorschach. However, because Rorschach-derived scales are less "face valid" than many other assessment procedures, they may have to meet a higher standard of validation than other tests in order to quell undue criticism of this approach (e.g., Dawes, 1994; Wood, Nezworski, & Stejskal, 1996).

WHAT IS IMPORTANT IN RORSCHACH VALIDATION?

From a practical standpoint, there are essentially three uses for any clinical test: (a) to aid differential diagnosis; (b) to predict treatment needs, treatment course, or outcome; and (c) to describe abilities, proclivities, traits, states, or dynamics. Many studies indicate scales derived from the Rorschach can yield information relevant to differential diagnosis and a smaller number of studies indicate these scales can yield information relevant to prediction (e.g., Blatt & Ritzler, 1974; Exner, 1993; Garwood, 1977; LaBarbera & Cornsweet, 1985; Perry & Viglione, 1991; Rydin, Asberg, Edman, & Schalling, 1990; Thomas & Duszynski, 1985; Tuber, 1983). However, any refined ability to make differential diagnoses or predictions is fundamentally accomplished because scales have the ability to describe and quantify relevant personality characteristics. Consequently, to further the Rorschach's clinical utility it is most important to focus efforts on validating the descriptive meaning of test scores (Weiner, 1977; Widiger & Schilling, 1980).

GENERAL ISSUES IN CONSTRUCT VALIDATION

To document what a test score means is always a matter of construct validity (Cronbach & Meehl, 1955). Any consideration of construct validity assumes two phenomenal planes, one of abstract, theoretical constructs, and one of messy, imperfect, real-world variables that approximate the abstract constructs. A test score is valid to the extent that it shines a strong, clearly defined beam of light into the theoretical plane and illuminates the specific region of meaning that was antici-

pated. It is invalid to the extent that it casts a diffuse or weak light that partially illuminates the expected region, or shines a focused beam that veers off in an unanticipated direction. To empirically demonstrate a score is valid, it is necessary to "calibrate" the score against a real-world criterion variable. Ideally, the latter would be known to shine a strong, focused light on the same theoretical terrain as the test score. In reality, however, this is rarely the case. Instead, the criterion construct is generally different from the test score construct and the operationally defined variable that measures the criterion construct is also typically inadequate, having its own diffuse or errant illumination. Thus, test validation requires an imperfect, test-derived measure of a latent construct to be compared to a criterion variable that is an imperfect measure of a somewhat different latent construct.

Given the ambiguities and imprecision on both ends of such an empirical relationship, construct validation procedures must be invoked in order to make any headway. A test score must be evaluated against the broad *pattern* of empirical findings obtained from a wide array of less-than-optimal criterion variables (Cronbach & Meehl, 1955). Theory about the test score construct dictates which criterion variables should be associated with the score and which should be independent (Campbell & Fiske, 1959). The extent to which the broad pattern of empirical findings interlock with theoretical predictions about the test variable then forms the "nomological network" that either supports, negates, or refines understanding of the construct actually measured by the score.

Years ago, in an insightful discussion of test validation problems, Meehl (1959) described four levels of practical importance at which a test could be validated. The first or lowest level simply addresses the question: "How accurate are the semantically clear statements which can be reliably derived from the test" (p. 114)? The second level addresses the extent to which the information provided by a test score is a commodity that is not "concurrently and readily" obtained by routine and ordinary clinical observation (p. 114). The third level addresses the more complex question of "*how much earlier in time*" (p. 116) the test allows us to make these accurate statements that are not readily available from other sources. Finally, Meehl's most demanding level addresses the question of "in what way, *and to what extent*, does this incremental advance information help us in treating the patient" (p. 116–117)?

In this hierarchy, Meehl intermingled issues of validity with issues of utility. By the time the final level is reached, the question is primarily one of utility rather than validity, because the accuracy of the construct being measured by the test score is no longer in dispute. The first three levels, however, are focused more squarely on issues of validity. In fact, it is the first level that captures the essence of construct validity because this level is focused on determining the descriptive meaning of a score. Unfortunately, the levels presented by Meehl are also a bit deceptive. They imply validation should proceed in a sequential fashion up the hierarchy, addressing the first question prior to the second, and so on. However, it is impossible to answer

the first question unless one also has in place the methodology to answer the second and third. That is, it is impossible to gauge the accuracy of test derived descriptions (Level 1) unless one already has unassailable validation criteria that significantly surpass the fallible clinical judgments and inferences that may be drawn during extended but routine clinical care (Levels 2 and 3).

The value of all psychological tests lies in their ability to provide quantified information that clinicians may not yet know and which may not be readily apparent to other mental health or medical professionals. This is their purpose. Needless to say, trying to validate that which may not yet be known nor readily apparent to the clinically trained eye is a very difficult enterprise. To perform this feat, researchers need to have at least five things: (a) a well-articulated model that specifies the pattern of empirical findings that should be observed if a test score actually measures its intended construct, (b) understanding of the systematic sources of bias that affect every method of assessment, (c) understanding of how characteristics at various levels of personality organization are related (or not) to conscious awareness and overt behavioral symptoms, (d) appreciation of the limitations inherent in typical validity criteria, and (e) strategies to improve validation criteria so they can more efficiently document whether scores measure what they should (i.e., Meehl's first level).

HURDLES UNIQUE TO RORSCHACH VALIDATION

Over the years a number of reviews have discussed issues pertinent to validating Rorschach-derived scales (e.g., Blatt, 1975; Meehl, 1959; Weiner, 1977; Widiger & Schilling, 1980). However, several factors have not been sufficiently addressed before, even though they make the validation enterprise particularly hazardous for the Rorschach method.

As with other techniques, the science of Rorschach-based assessment is relatively young. However, Rorschach-based assessment aspires to quantify the exceedingly complex phenomena of personality in all its dynamic richness across levels of awareness and degrees of behavioral expression. This is an ambitious undertaking.

A second hurdle arises because scores derived from the Rorschach do not begin with a self-evident meaning. The psychological construct measured by any score is initially unknown. Unlike a self-report or observer-rating scale that can initially define any intended construct in an obvious fashion through easily understood language (e.g., "How depressed are you?"), there is no such simplicity when it comes to metrics derived from the articulation of inkblot images. Those who study the Rorschach have started from scratch in their efforts to ascertain the core meaning associated with the various features of traditional inkblot interpretation, such as shading, dimensionality, reflections, movement, unusual detail locations, types of

content, and so on. Historically, the task of divining how a complexly determined score related to subtle variations in personality was guided more by attentive clinical observation than by research data. This was appropriate because astute clinical inference can often be more trustworthy as a guide to clinical realities than poorly conceived empirical findings (Meehl, 1978). However, because there are also many blind alleys and wrong turns that can be taken during such a nebulous inferential journey, interpretive reasoning requires as many empirical landmarks as possible.

A third validation hurdle is the considerable skill required for all aspects of Rorschach administration, inquiry, and scoring. Unlike a self-report instrument, sophistication and advanced training are necessary to perform each of these steps in a reliable and valid fashion. This journal has established basic standards for accurate scoring (Weiner, 1991). However, these standards are not uniform across journals and a large body of data has been published without this information. Furthermore, guidelines have not yet been produced to ensure that protocols are administered and inquired in a proper fashion. Examiners may fail to recognize key words and thereby not inquire about relevant determinants, may "chase" determinants in an inappropriate fashion, may not establish sufficient rapport, may over-include or under-include information because responses were not transcribed verbatim, and so on. When these upstream factors are operative, they are not corrected by having raters reliably code the resulting protocol.

A fourth hurdle for Rorschach validation arises from the uniqueness of the method (e.g., Weiner, 1994). The manner in which the Rorschach obtains information from patients is so different from any other method of collecting personality information that it forms its own category of method; a category that is very distinct from the other so-called "projective" tests with which it has historically been categorized. Being a unique method, the Rorschach lacks a powerful advantage bestowed upon the more commonly used measures of personality assessment—namely, the ready ability to demonstrate convergent validity with criteria that are drawn from within the same method family.

The vast majority of tools for assessing personality employ a self-report methodology. Most often, construct validity for these measures is determined by correlating a target scale with another self-report scale of the same construct. In such a design several factors conspire to inflate the magnitude of the resulting correlations (see McClellend, 1980). First, the same participants rate themselves on both measures from the same observational perspective—a perspective of conscious self-representation. As a result, all the limitations inherent to this observational perspective, such as one's degree of self-awareness, pervade both tests to the same extent. Second, scores on both measures are moderated by the same motivational biases for self-presentation, such as efforts to present a globally idealized or pathological view of oneself, efforts to disguise only certain characteristics, and so on. Third, scores on both measures are moderated by the same response biases and proclivities. These include idiosyncratic differences in care-

lessness, preferences for the central regions or extremes of a rating scale, benchmarks for judging the meaning of an item, benchmarks for determining the degree to which an item describes oneself, and so on. Finally, participants rate themselves on both the predictor and criterion using items that are essentially synonymous, even though the instruments may have different names or slightly different formats. For instance, tearfulness may be assessed by an item on both instruments, even though one may be named the Beck Depression Inventory (BDI), the other the Minnesota Multiphasic Personality Inventory–2 (MMPI–2); one may use true–false responses, the other a Likert-type scale; one may present constructs in adjective form, the other in sentence format; and so on.

These kinds of procedures clearly stack the deck in favor of positive convergent "validity" findings simply because so many methodological confounds are intertwined (see Campbell & Fiske, 1959; McClelland, 1980). Psychometrically, such monomethod validation procedures yield results that are more akin to reliability estimates than they are to validity coefficients. Often, this fact is overlooked, even though these features can easily—and erroneously—make it seem like a self-report measure is more "valid" than the Rorschach.

When searching for validation criteria, it is impossible for Rorschach-derived scales to stay within the same method family unless validation criteria are themselves other Rorschach scales. This is typically not the case. Therefore, researchers have consistently had to evaluate the construct validity of Rorschach scores with more demanding heteromethod validation criteria. The importance of this point should not be underestimated. The dramatic impact of method variance on convergent validity coefficients continues to be underappreciated in assessment research (e.g., Meyer, in press), even though it is ubiquitously established that heteromethod validity coefficients are considerably smaller than monomethod coefficients.[1]

The fifth hurdle in Rorschach validation is related to the fourth. Because the Rorschach is so unique, the sources of systematic bias associated with the method are also unique. To date, there has been less attention paid to these confounds than to the confounds associated with the self-rating or observer-rating methods. Nonetheless, the available data indicate the Rorschach is like any other assessment

[1]Because of method confounds, it is difficult to directly compare Rorschach validity coefficients with the validity coefficients found for other personality tests. To level the playing field and make such a comparison in a fair and impartial manner, evidence would need to be restricted to heteromethod validity coefficients. When doing so, however, it is important to recognize that confounds also affect some types of heteromethod analyses. Currently, a structured interview is considered to be the standard for obtaining clinician judgments. In an effort to maximize precision and interrater reliability, structured interviews have drifted in the direction of minimizing the degree to which true clinical judgment and inference are needed to complete an interview topic. To the extent an interview is very structured (or to the extent it fits within the oxymoronic category of "computer administered structured interview") and the derived data become limited to what the patient reports, the interview becomes little more than a verbal self-report scale, sharing considerable method variance with self-report questionnaires.

procedure. Method variance is large relative to desired trait variance (e.g., Meyer, 1992a, 1992b). To conduct optimal validation research, it will be essential to consider method variance, understand its impact on validity coefficients, and design research that takes its influence into account (Meyer, in press).

A sixth hurdle relates to the manner in which information on the Rorschach is aggregated. Most scoring categories are focused entities that assess single blot attributes. For instance, vista responses only capture those occasions when shading features generate a three-dimensional perspective. Although the data within any scoring category are aggregated across all responses, each card stimulus is quite distinct and each participant can give many or few responses to each card. Consequently, card pull and variable *R* within a card ensures that every score does not have an equal chance of occurring on each response across a full protocol. That is, a 20-response protocol does not present 20 equivalent opportunities to observe such things as a vista, diffuse shading, chromatic color, or reflection response (see Gacono & Meloy, 1994, chap. 7 for interesting data on this). As a result, many Rorschach scoring categories function more like single items, rather than like the multi-item scales found in most other personality, intelligence, and neuropsychological tests. A scoring category aggregated across 20 responses is not like 20 items aggregated to form an overarching scale because each response is not equivalent to the presentation of another "item" targeting the same underlying dimension.[2]

Because Rorschach scores operate more like single items than multi-item scales, they are more affected by random error and are less reliable measures of an underlying construct than a multi-item scale would be. The type of reliability being considered here is not captured in the agreement between two raters examining the same transcript of a protocol. Rather, what is being considered are those random factors that interfere with the translation of a patient's implicit and truly scorable blot image into a written transcript that allows the rater to quantify this imagery. These interfering factors can affect the participant's choice of particular words to articulate the image, the examiner's choice of particular inquiry words, the motivational climate during a particular response, omissions or elaborations in the written record, and so on. Because of these factors, psychometric theory predicts many Rorschach scores will have observed validity coefficients that are lower than they optimally could be if the scores could be aggregated in a meaningful fashion (e.g., Allen & Yen, 1979). To appreciate the potential significance of this, consider data presented by Cheek (1982). He found self-ratings and observer ratings of conscientiousness had an average correlation of .27 when assessed by a single item,

[2]In terms of single scores, the form quality percentages and *Zd* are exceptions to this rule because they take into account the unique features of each card as well as the number of responses given. Other exceptions include those scales that are genuine aggregations of items, such as total scores on the Constellation Indices.

although the validity coefficient rose to .37 when assessed by a three-item composite scale. Similarly, the validity coefficient for extraversion was .43 when assessed by a single item, although it jumped to .53 when assessed by a three-item composite.[3] The upshot is that whenever possible, Rorschach scores should be aggregated into meaningful conceptual categories in order to maximize validity coefficients.

A final hurdle that complicates Rorschach validation is uncertainty over the extent to which personality characteristics quantified by Rorschach scores should be manifest in an overt and recognizable fashion. In the context of self-rated personality characteristics being validated by observer judgments, Tellegen (1991) referred to this issue as a problem of *behavioral penetration.* This reflects the extent to which characteristics are expressed in overt behavior as readily identified symptoms or expressive actions, rather than simply experienced on a private, internal level as feelings, motives, needs, or mental templates. All of the latter may have minimal surface indicators and low penetration into the domain of overt behaviors. Characteristics with low penetration are obviously more difficult to validate than characteristics with high penetration (e.g., Harkness, Tellegen, & Waller, 1995).

In the context of Rorschach validation, it is necessary to expand Tellegen's concept by adding the notion of *conscious penetration.* Under optimal circumstances, self-report instruments generate data that are limited to consciously accessible information.[4] In contrast, Rorschach-derived scores may measure constructs that are fully within conscious awareness, within partial awareness, or that reside completely out of consciousness. Of those characteristics that are not in conscious awareness, some may be conflict-free (much like heart rate or blood pressure), whereas some may be dynamically defended against. Of those characteristics that are defended against, some may also be actively countered by overt behaviors or conscious attitudes that oppose the underlying dynamic. Obviously, characteristics that have low conscious penetration will be more difficult to validate than those with high penetration, those that are defended against will be more difficult to validate than those that are not, and those that are kept submerged by antithetical beliefs or behaviors will be the most difficult to validate.

If Rorschach-derived scores genuinely measure characteristics with low behavioral penetration and/or low conscious penetration (e.g., Meyer, in press, this issue), it will be much trickier to find appropriate validation criteria for these scores because they may not be readily observable to others, may remain out of the participant's awareness,

[3]Cheek (1982) actually reported the coefficient for three items and three raters, in addition to one item and one rater. However, he provided sufficient information to calculate the three-item-one-rater validity coefficient by using the formula given in Tsujimoto et al. (1990).

[4]Strictly speaking, this is the best that can be obtained from self-reports; patients can of course report whatever they like. It is also the case that clinicians can make inferences about item responses, scale elevations, or profile configurations that go beyond the participant's conscious awareness.

or may be camouflaged by conspicuous overt behaviors and/or consciously held attitudes that are diametrically opposed to the underlying characteristic.

Surprisingly, these issues have not received much attention in the Rorschach literature. Currently, there is no consensus about which scores measure constructs that should be easily observed, which should have few surface indications, which should be expressed in overt behaviors but only under particular circumstances, which should be clearly represented in consciousness, which should have no conscious representation but be conflict-free, which should have no representation in consciousness but be consistently defended against, or which should be consciously represented but only under particular conditions. In the absence of cogent, empirically guided theoretical reasoning that addresses these issues for each score, I believe validation efforts should adopt a conservative approach. In general, Rorschach scores should be seen as providing information about implicit problem-solving strategies, tacit dynamics, mental templates, and underlying propensities, all of which have only mild to moderate behavioral penetration and low conscious penetration. No doubt, there are exceptions to this suggestion. However, it seems wise to err on the conservative side and not expect scores to permeate behavior or consciousness when these scores are considered in isolation from other scores or sources of information. Making such an assumption does not diminish the importance of Rorschach-derived variables. Rather, it recognizes the inherent complexity of personality and aspires to maximize the research yield by ensuring validation criteria are selected which are not excessively dependent on superficial behavioral observations or conscious self-awareness.

THE CONSTRUCT VALIDITY OF CRITERIA

When trying to document the construct validity of a test score, one must ask whether the operationally defined criterion variable can be considered an unerring quantification of a relevant criterion construct. To answer this question, the criterion should be evaluated on three interdependent dimensions, each of which are subject to different kinds of systematic error. Gold standard criteria must be in an optimal range on each dimension. However, criteria can be compromised on any dimension, and inadequacies on one are not rectified by strengths on another.

The first dimension of criterion construct validity (CCV) is one of construct relevance and fidelity. It focuses on the theoretical level and deals with the correspondence between the criterion and predictor constructs. A good criterion must have a psychological meaning that directly corresponds to the presumed descriptive meaning of the test score it will validate. The more synonymous the better. For instance, a test score construct of "painful self-evaluation" should be validated with a criterion construct of "painful self-evaluation," rather than a more tangential construct like "emotional distress." For the criterion to illuminate the

same conceptual terrain as the test score, the criterion construct must be as clearly defined, parallel in scope, and conceptually differentiated as the predictor construct (see Campbell & Fiske, 1959; Cook & Campbell, 1979).

The second dimension is concerned with operational definitions, or how well a messy, real world variable defines a theoretical construct. The relevance of the construct is no longer at issue. Rather, this dimension is only concerned with the correspondence between the abstract, theoretical plane and the practical plane. The abstract construct must be defined in all its relevant parameters by an observable tool that is amenable to measurement. The operationally defined variable must be sufficiently representative of the construct, must not introduce excessive information from sister constructs through faulty definition, and must have gradations that correspond to gradations in the criterion construct in a tightly linked fashion.

The third dimension is concerned with data points, or the quality of the actual numbers that are generated to quantify an operationally defined variable. Although any operational definition necessarily implies a particular source of data, the quality of the data is a function of forces that are quite distinct from how a construct is formally defined. In particular, to determine the accuracy of statements derived from a test score (Meehl's first level), it is essential to have criterion data that approximate truth and permanence rather than being encumbered by the bias of superficial observation, transient accuracy, or inherently limited points of view. For instance, to determine if a Rorschach score actually assesses splitting, a criterion must be able to document whether a narcissistic, borderline, or bipolar patient has a self-structure that oscillates between poles of idealization and inferiority, even though criterion data may be gathered while the patient is hospitalized for the first time in the midst of a major depressive episode and temporarily identified with only the negative pole of this dichotomy. The third dimension directs the researcher to consider whether the activities associated with obtaining, classifying, or entering data into the operationally defined criteria cause the data to be confounded by situational bias or perspective-associated error.

In brief, the first dimension of CCV is focused on the similarity between predictor and criterion constructs. Systematic error is present when the criterion construct deviates from the predictor construct in either scope or intensity. The second dimension is focused on accurately defining an observable structure to track the target construct in all its relevant features. Systematic error is present when the operationally defined variable inadequately encompasses the construct or when it includes irrelevant components. The third dimension is focused on obtaining data points that accurately quantify the variable. Systematic error is present when factors associated with a particular time, place, or perspective bias the data.

To exemplify these dimensions, consider criteria to validate a test construct of depression. To be high on the first dimension, the criterion construct must also be depression. The construct of neuroticism has a broader scope that encompasses irrelevant features, whereas the construct of anger lies in adjacent conceptual

terrain, making either of these less optimal. Depression can be operationally defined by many different but equally reliable tools, including the dexamethasone suppression test (DST), a count of self-critical statements made during a performance task, self-ratings on the BDI, or observer ratings on the Hamilton Depression Rating Scale. However, each of these variables define the construct in a different manner. For illustration, only the BDI and DST will be discussed. The BDI would be high on the second dimension of CCV. Although the items in the scale may include some minor irrelevancies, the conceptual terrain illuminated by the BDI's operational definition is fairly precise and consistent with an accepted definition of the depression construct. The DST, on the other hand, defines depression as the failure to inhibit cortisol secretion after the administration of dexamethasone. Although there is a very explicit operational definition about how to measure cortisol nonsuppression, the construct is over-inclusive and a single administration of the DST is not specific to depression per se (e.g., Ribeiro, Tandon, Grunhaus, & Greden, 1993). Therefore, the DST does not define the criterion construct in all its relevant parameters and is low on the second dimension of CCV.

However, when considering the third dimension of CCV, the DST and BDI reverse positions because the data that enter into the DST are more able to transcend systematic bias than the data that enter into the BDI. Although the results of the DST can be sensitive to situational factors associated with neuroendocrine functioning, the human tasks associated with administering dexamethasone, drawing blood at specified time intervals, and determining cortisol levels are well-specified and not very sensitive to judgment error, perspective-dependent observational bias, or other sources of situational influence. Thus, even though the DST does not provide a sound definition of depression, the data that quantify the variable are fairly robust and accurately quantify the operationally defined variable. In contrast, BDI data are much more subject to systematic bias from factors such as limited self-awareness, poor comprehension of the item content, motivated distortions, global judgments, and so on. Consequently, even though the BDI provides a sound operational definition of the depression construct, the BDI is quite compromised on the third dimension of CCV because the obtained data points may well quantify something other than what is contained within the formal operational definition.

JUDGMENT-DEPENDENT AND NONJUDGMENT-DEPENDENT CRITERIA

Criterion variables can also be roughly generalized into two types: those that utilize data requiring a minimal degree of subjective judgment, and those that utilize data points requiring considerable subjective judgment. Such a classification is useful because it helps clarify a way in which validation criteria systematically differ on the dimensions of CCV. Within this classification there tends to be an inverse relationship between the first and third dimensions. Criteria that require minimal

judgment tend to assess less relevant constructs, although they tend to do so with less bias introduced at the stage of data collection (i.e., they are lower on the first dimension of CCV but higher on the third). In contrast, criteria requiring considerable judgment tend to be buffeted by bias at the stage of data collection, although they can also operationally define highly relevant constructs (i.e., lower on the third dimension, although also capable of being higher on the first). Recognizing the strengths and limitations associated with these broad types of criteria is an initial step toward improving validation research.

Criteria requiring relatively little judgment can be manipulated independent variables, naturally occurring classification variables, or simply other variables that will be associated with the test score. Examples of some broad categories would include the following: (a) affects, beliefs, or motivations that are experimentally manipulated on a short-term basis, such as through instructions to malinger, induced expectations regarding social desirability, the production of anxiety, helplessness, or arousal, and so on (see Bornstein, Rossner, Hill, & Stepanian, 1994; Exner, 1980; McCown, Fink, Galina, & Johnson, 1992; Perry & Kinder, 1990; Perry, Sprock, et al., 1995); (b) longer term interventions that should differentially effect particular test scores, such as through pharmacological treatment, various forms of psychotherapy, and so on (see Exner & Andronikof-Sanglade, 1992; Weiner & Exner, 1991); (c) historical life events, such as having experienced sexual abuse, having parents who recently divorced, having been placed in many foster homes, and so on (see Exner, 1993; Leifer, Shapiro, Martone, & Kassem, 1991; Spigelman & Spigelman, 1991); (d) naturally occurring criterion events, such as being in a severe storm on a relatively small ship, awaiting an initial trial of electroconvulsive therapy (ECT), anticipating an initial parachute jump, being sentenced to indefinite incarceration, receiving a strict fluid-only diet, and so on (see Exner, 1993; Shalit, 1965); (e) outcome on some future life event, such as reductions in medication, length of time in psychotherapy, symptom remission, eventual suicide, and so on (see Blatt & Ritzler, 1974; Exner, 1993; Garwood, 1977; LaBarbera & Cornsweet, 1985; Perry & Viglione, 1991; Rydin et al., 1990; Thomas & Duszynski, 1985; Tuber, 1983); (f) behaviors quantified in the laboratory or under controlled field observations, such as the amount of time spent gazing at oneself in a mirror, the frequency with which one sorts objects by texture, the manner in which one interacts with a marital partner, whether one helps an examiner pick up dropped books, the physical distance maintained between oneself and experimental confederates, the number of errors made during a game of "Simon Says," and so on (see Alexander, Farber, Sacco, & Trungold-Apter, 1995; Blake, Humphrey, & Feldman, 1994; Exner, 1993; Marsh & Viglione, 1992); (g) test score changes and stability over time, such as found in studies of maturational development or test–retest stability (see Blatt, Brenneis, Schimek, & Glick, 1976; Exner, 1980, 1993; Perry, McDougall, & Viglione, 1995; Wenar & Curtis, 1991); (h) performance on ability tasks, such as problem-solving speed, problem-solving accuracy, visual perception, verbal intel-

ligence, and so on (see Exner, 1993; Meyer et al., 1993); and (i) psychobiological phenomena, such as serotonin or dopamine levels, localized brain damage, and so on (see Rydin, Schalling, & Asberg, 1982).

Once these criterion constructs have been operationally defined, there is generally little judgment required to determine a participant's score on the criterion. For instance, a participant is either in the group instructed to malinger or is not, has either experienced a recent divorce or has not, either sits 2 ft or 5 ft away from a confederate, is either in the long-term or the short-term psychotherapy group, has a Trails B time of either 45 or 102 seconds, is either awaiting his first ECT trial or is not, and so on. Although the data points that enter into some of these criteria can be bound by situational constraints, the data are generally not confounded by other sources of systematic influence. As a result, when research is conducted in a sophisticated fashion, these criteria tend to share strengths on the third dimension of CCV. However, these criterion variables are also typically low on the first dimension of CCV. In their final form, they tend to reference constructs that are not synonymous with the conceptual terrain believed to be measured by Rorschach test scores.

Clearly, there are some exceptions to the last assertion. For instance, in an exemplary study McCown et al. (1992) developed narrowly defined experimental criteria of stress and helplessness that paralleled the hypothesized meaning of *m* and *Y* scores; Exner's (1993) criterion of suicide is synonymous with the construct targeted by the Suicide Constellation; Exner's (1993) criterion of errors made during a game of "Simon-Says" has a meaning that comes fairly close to the hypothesized meaning of the *Zd* score; and Alexander's (Alexander et al., 1995) criterion of spontaneously helping an examiner pick up her books comes fairly close to one element of what the *COP* response is supposed to mean. Despite these exceptions, when a researcher aspires to use data that require relatively little judgment, it will be quite difficult to find criterion variables that illuminate the same conceptual terrain as the test score.

It should not be surprising that significant problems of construct relevance and fidelity are encountered when using these kinds of criterion variables. In fact, Cronbach and Meehl's (1955) great insight was to recognize criterion variables always have excessive, incomplete, or irrelevant meanings. Consequently, they proposed the goal of validation research was to find many different kinds of imperfect criteria that should share some degree of conceptual (and empirical) association with the test score (Cronbach & Meehl, 1955; Weiner, 1977; Widiger & Schilling, 1980). Only after obtaining a network of empirical findings with suboptimal criterion variables can one examine the overall pattern to determine if it is theoretically compatible with the intended construct.

If Cronbach and Meehl's (1955) supposition is true and all validation criteria are plagued by construct limitations, a large number of studies will need to be conducted on each test score before gaining confidence about the meaning of that

score. For a task like the Rorschach, which can generate an almost infinite number of scores, this requirement translates into an extensive body of empirical data. A reasonable question then is whether there is any way to facilitate this process by finding criterion variables that are more able to approximate construct relevance and fidelity.

To maximize the first dimension of CCV and obtain a criterion variable that defines the same construct as that believed to be measured by the test score, it is necessary to create a criterion that captures the target construct—and nothing more—in all its essentials. The most efficient way to capture a theoretical construct in its essentials is to utilize language. In turn, if language is the tool that illuminates a criterion construct, this tool can only be powered with data points generated by subjective judgment. Unfortunately, such an approach maximizes problems encountered with the third dimension of CCV. Questions about the adequacy of the criteria now shift almost exclusively to the ability, skill, knowledge base, biases, and clinical sophistication of the individuals who determine the ratings, be they self-ratings or the ratings of teachers, parents, peers, spouses, or clinicians. As such, although there is little doubt the rated constructs can have a high degree of relevance and fidelity, there are many doubts about the ability of any rater to generate data points that are not situationally dependent, clouded by a particular observational perspective, moderated by response sets and motivations, or influenced by other sources of systematic error.

Nonetheless, if situation-dependent and perspective-dependent sources of bias could be minimized, it is possible these criteria could become more optimal. Two options are available. The first makes use of methods to enhance expert judgment, whereas the second makes use of aggregated data from diverse sources.

IMPROVING JUDGMENT-DEPENDENT VALIDATION CRITERIA

Longitudinal Expert Evaluation of All Data

In 1983 Spitzer briefly articulated a method for generating judgments that approximate gold standard validity criteria. This approach, the Longitudinal Expert evaluation of All Data (LEAD), proposes that an expert clinician evaluate a patient over time and utilize all available sources of information before formulating a diagnosis or making a description of the patient's personality. Although demanding, the LEAD method has been successfully applied in several studies of personality disorders (e.g., Pilkonis, Heape, Ruddy, & Serrao, 1991) and, when properly implemented, obtains criterion data that come about as close to clinical reality as possible.

The success of the LEAD approach depends on at least three factors. The first is the degree of expertise held by the clinicians who will provide ratings. Obviously, it is necessary to select skilled, experienced, and conceptually sophisticated clini-

cians. Because the Rorschach is believed to measure constructs that may not be fully conscious, it is also important to select clinicians who have a dynamic understanding of psychopathology (yet are not rigidly wed to a single theoretical approach). Also, because data indicate that credentials or years of experience alone do not aid clinical judgment (e.g., Dawes, 1994; Kleinmuntz, 1990), it would probably be best to use some procedure to select clinicians who are perceived by others to be competent and clinically sophisticated. Finally, because clinicians should have expertise in the rating tool they will use, they would need to be trained to understand the constructs they are quantifying and to properly use the criterion measure.

Second, clinicians must review "all data" in an effort to transcend the limits of situation-dependent behavior and the peculiarities of their observational perspective. Both the depth and the breadth of the data used to formulate judgments are important. In terms of depth, it would be optimal for clinician-raters to adopt an insight-oriented or interpersonally focused approach to interviewing patients so there is an opportunity to understand the patient's developmental experiences, central conflicts, defensive strategies, functional capacities, and interpersonal relationships. In terms of breadth, clinician-raters should obtain information from a variety of different sources, consulting medical, military, and/or psychiatric records as appropriate, and consulting with family members, spouses, employers, coworkers, friends, and/or any other mental health professionals who have treated the patient. Furthermore, clinician raters need to have their observations cover as many distinct settings as possible, so should optimally sample one-to-one, group, peer, and family interactions.

The third important factor is the length of time raters have to observe or gather relevant information about the target patients. Obviously, a gold standard understanding is not obtained in six 1-hr meetings nor in a single lengthy interview. Instead, to help maximize the third dimension of CCV and counter the biases inherent in judgments, a longitudinal design is necessary. Relevant research indicates that knowing a participant for longer periods of time will increase the validity coefficients between personality test scores and observer ratings (e.g., Paunonen, 1989; Tett, Jackson, & Rothstein, 1991). A lengthy period of contact ensures there has been sufficient time for the clinician to observe the patient across a range of situations and affective states and it allows for relevant information and the course of illness to unfold. Although somewhat arbitrary, an observation/information gathering period of about 30 to 40 hr seems to be a reasonable minimum cutoff in this regard (i.e., about 8 to 10 months if each patient is allocated an hour per week of the expert's time, or 4 to 6 weeks if allocated an hour per day).

Aggregated Ratings

The second approach to refining judgments is psychometrically based. Averaging across multiple judges substantially improves the validity that can be demonstrated

between a test score and criterion (Tsujimoto, Hamilton, & Berger, 1990; also see Cheek, 1982). Tsujimoto et al. presented a useful formula that specifies the typical validity coefficient one can anticipate when aggregating across raters. This expected coefficient (subject to sampling error) is a function of three things: the magnitude of the effect size typically found with single raters, the number of raters who will contribute data to the aggregated ratings, and the average pairwise correlation between raters (i.e., interrater agreement). Tsujimoto et al. demonstrated the validity coefficient one can expect hinges on the average agreement between raters. Specifically, the validity coefficient will increase when agreement among raters decreases. In other words, it is when independent judgments are nonredundant that validity coefficients are maximized. Therefore, this approach would encourage researchers to obtain ratings from people who have quite distinct perspectives of the target patient. This would be accomplished by sampling individuals who know the ratee from across a range of life settings and have qualitatively different relationships with him or her (e.g., spouses, therapists, friends, etc.). Although self-ratings are unrelated to Rorschach scores in heterogeneous samples (Archer & Krishnamurthy, 1993a, 1993b; Meyer, in press), to the extent that self-ratings are distinct from the ratings made by others, the aggregation procedure indicates they should also be included within the pool of ratings.

How practical is the aggregation approach? Consider for instance a Rorschach construct that has low conscious penetration but moderate behavioral penetration. It would be reasonable to expect such a construct to generate a medium effect size in nature, as Cohen (1988) defined this as a phenomenon that is likely to be visible to the naked eye of a careful observer. A medium effect would be reflected in an average correlation of about .30 between the test score and the judgment of any single rater. This would then be the typical validity coefficient for the target construct without aggregation. However, rather than using a single rater, consider what happens when the judgments of four raters are pooled. If each rater has a relatively nonredundant understanding of the target subject, say with an average pairwise correlation among raters of .25, then the expected validity coefficient for aggregated ratings jumps from .30 to .45; clearly a substantial gain. Alternatively, if the raters tend to agree with each other, having an average pairwise correlation of say .75, the expected validity coefficient for four raters would be only increase from .30 to .33.

To the benefit of test validation efforts, raters drawn from diverse settings tend not to agree with each other. For instance, Achenbach, McConaughy, and Howell (1987) conducted a meta-analysis of this issue with respect to child and adolescent psychopathology. Incorporating data from 269 samples and more than 44,000 participants, they found the following average pairwise correlations between different sets of raters: parent and teacher, $r = .27$; parent and mental health worker, $r = .24$; parent and self, $r = .25$; teacher and mental health worker, $r = .34$; teacher and self, $r = .20$; and mental health worker and self, $r = .27$. Not surprisingly, they

also found cross-informant coefficients were significantly lower for internalizing personality problems than for externalizing conduct problems. Given these findings, an aggregation procedure should prove to be quite beneficial for Rorschach validation.

Also of note, the extent to which expected validity coefficients increase with aggregation follows a deaccelerating course, such that more and more raters provide smaller and smaller gains. For instance, using the previous example when interrater agreement was .25 and any single rater had an expected validity coefficient of .30, the expected validity coefficient for aggregated groups of two through seven raters would be .38, .42, .45, .47, .486, and .50, respectively. Given this pattern of gains, one can anticipate substantial validity increments while maintaining a relatively cost-effective research design when data are gathered from three or four raters who have distinct perspectives on the target patient.

The aggregation procedure empirically synthesizes information and does not place trust in the judgment of any single rater above any other. In contrast, Spitzer's (1983) approach places stock in a single expert clinician to synthesize information under the assumption that training and expertise allow the clinician to accurately recognize subtle qualities that may well be missed by aggregating across diverse raters. It is unclear which approach will ultimately prove to be the most accurate. In terms of practical implementation, the aggregation procedure requires the time to obtain data from a larger number of raters per participant, but much less professional time to actually gather and synthesize relevant information and generate ratings. Consequently, for most researchers, the aggregation procedure may be a more practical strategy.

To implement either of these methods for test validation purposes, a suitable rating tool is also required. To validate the Rorschach, it is necessary to have a criterion that contains clearly written and easy-to-understand items that accurately describe each construct the Rorschach is believed to measure. The remainder of this article discusses the initial development and potential uses of one such scale.

SCALE DEVELOPMENT

The Rorschach Rating Scale (RRS) has been developed over the past several years by Cheri Adrian, Robert Bornstein, Walter Burke, Greg Friedman, Carl Gacono, Paula Gorlitz, Nancy Kaser-Boyd, Paul Lerner, Barry Ritzler, Steven Tuber, Donald Viglione, Jr., and myself. All items on the scale have been subjected to extensive review and revision to ensure they are clearly written and accurately reflect the construct thought to be measured by different scores.

Items for the Comprehensive System were systematically reviewed on six occasions. On each occasion, items were generated or revised and then judged for adequacy. I generated the first set of items and had them judged for conceptual

clarity and readability by six clinical psychology graduate students.[5] Based on their ratings, items were dropped or revised and new items were generated. For the second step, the revised items were submitted to two experts (Viglione and Ritzler), who judged how well each item described the construct believed to be measured by the target Rorschach variable. Based on their ratings, items were again dropped or revised. At this point, the item pool was also expanded to cover the Rorschach's primary factors (Meyer, 1992a, 1992b), the Ego Impairment Index (see Perry & Viglione, 1991), and the Complexity Index. The Complexity Index is still being developed, although it is quite similar to the Rorschach's first factor, being defined by high *R*, high *Zf*, high *DQ+*, many secondary contents, low *Lambda*, and few pure *A* contents.

As the third step, these items were submitted to a panel of seven clinician reviewers[6] who were asked to judge: (a) the clarity of the concept (without respect to any Rorschach construct), (b) the clarity of the language used in the item, and (c) the ease with which a clinician could rate a patient after seeing the patient in weekly psychotherapy for 1 year. Items were revised if two of the judges agreed an item's construct or language was only *fairly clear*, or if one of the judges felt an item was *unclear*. On the basis of these conservative criteria, 28 of the 167 items (17%) were revised and then resubmitted to the same clinicians for a second evaluation. By and large, the revisions were successful, with the consensus among judges being the revised items were *much better than the original*. Seven of the 28 items met the conservative criteria used earlier to enact a revision. Because there was general agreement 3 of these items continued to be rather poor, they were dropped from the rating scale. Regarding the ease with which clinicians could rate a patient on each item, 4 of the 164 remaining items (2.5%) were considered to be *difficult* by one of the seven judges, although no judge considered any item to be *very difficult* for a therapist to use. Consequently, the difficult items were revised and 3 new items were developed, bringing the total once again to 167.

Because the first and third stage in scale construction dealt with the clarity of items, the fourth stage was again designed to ensure that the items were accurate descriptions of their intended Rorschach constructs. The revised items were considered by a panel of Comprehensive System experts consisting of Adrian, Kaser-Boyd, Ritzler, and Viglione, who judged how accurately each item captured the construct believed to be measured by each Rorschach score. These ratings were made on a 5-point scale that ranged from 0 (*very inaccurate/discard*) to 4 (*very*

[5]Thanks to James Aikens, Kerry Aikman, Bill Divane, Fred Fischer, Deborah Gross, and Rebecca Weisenthal for their help with these ratings.

[6]These clinicians were either interns in clinical psychology or residents in psychiatry. They had an average of 4.3 years of clinical experience and were primarily from either psychodynamic or cognitive behavior orientations. Thanks are extended to Jennifer Bleak, Mary Daly, Pat Mumby, Deanne Orput, Susan Ries, John Roberts, and Michelle Rosen for their help with this stage of the project.

accurate/hard to make better). If an item was less than optimal, the judges suggested revisions. Through a process of successive approximations, problematic items were refined. Initially, most items fared well. Three of the judges considered 9 items to be fair or inaccurate. A number of other items were considered accurate or very accurate by two of the judges but fair or inaccurate by the other two. Subsequently, a total of 39 new or revised items were considered for the next round of the review. At the fifth stage, three judges assigned a rating of fair or below to just 1 item, although 3 other items were somewhat problematic, being considered fair by two judges and accurate or very accurate by the other two. These 4 items were revised and reviewed again. All were judged to be accurate or very accurate following this final revision.

To make the data pool more comprehensive, a number of individuals were asked to contribute items for other scoring systems that they had developed or used extensively. Most decided to do so, such that the extended pool is heterogeneous and covers most of the current approaches to Rorschach scoring. A similar but less formal process of item generation and revision was used with these scoring systems. Gacono generated items for his impressionistic response and aggression scores (Gacono, 1990; Gacono & Meloy, 1994), for the Rorschach Defense Scales (Cooper et al., 1988; Cooper, Perry, & O'Connell, 1991), and for Kwawer's description of primitive object representations (Gacono et al., 1992; Kwawer, 1980). Friedman, Burke, and Gorlitz generated items for the Psychoanalytic Rorschach Profile (Blake et al., 1994; Burke et al., 1988), P. M. Lerner generated items for his defense scales (Lerner & Lerner, 1980, 1982; Lerner, 1990), Tuber created items for the Mutuality of Autonomy Scale (Tuber, 1992; Urist, 1977; Urist & Shill, 1982), and Bornstein generated items for the Rorschach Oral Dependence Scale (Bornstein, 1993; Masling et al., 1967). All of these items were then reviewed and revised by the initial authors and myself in an iterative process of up to five steps.

Currently, the rating scale consists of 263 items arranged in a 16-page packet. The first page contains a series of questions to be completed by the rater and instructions for completing the form. Following the instructions are 185 items assessing single scores or focused constructs. Within this section, 4 items are designed to assess random responding and another 5 are repeated items that will help assess response consistency. This section is then followed by a section devoted to the Constellation Indices. The information section and instructions differ for observer and self-ratings. As would be expected, the items and instructions are phrased in the third person for observer ratings and the first person for self-ratings.

Table 1 presents about 10% of the 185 items focused on single constructs. As can be seen from the table, all RRS items have been organized into rough conceptual categories, and a representative item or two is included from each category. Placement was based on my general sense of how the written constructs fit together, without consideration of which Rorschach variable was being targeted. Although classification into these categories was not rigorous, it provides a meaningful

TABLE 1
Selected Rorschach Rating Scale Items and the Scores They Target

Self Concept

5. At least below the surface, this person is very self-critical and has painful feelings about him/herself. [*SumV* (high, > 0)]

Emotional Experience

23. This person has emotional experiences that s/he finds very confusing. For example, s/he frequently feels both positively and negatively about the same thing and is unable to resolve this ambivalence. [Color-Shading Blends (high, > 0)]

Problem Solving and Coping

32. This person is flexible and has multiple ways of coping with stressful circumstances. [Complexity Index (high) with EII (low)]

35. This person is quite economical in his/her approach to tasks and rarely becomes engrossed or emotionally caught up in activities. [*R* (low; < 18) and *Lambda* (high, > .85), experimental; or Factor 1 (low); or Factor 2a (low)]

Cognitive Style

52. This person's style of thinking is holistic, impressionistic, and lacking in specific detail. [Impressionistic Response (high, Gacono); or Factor 3 (high)]

57. This person has difficulty shifting attention, thinking flexibly, or understanding events from more than one perspective at a time. [*a:p* ratio (imbalanced, $a + p \geq 4$ and [$a \geq 2p$ or $p \geq 2a$]), ideational interpretation; or *PSV* (high; > 0)]

Internal Dynamics and Defensive Operations

64. This person relies on logic, knowledge, and objectivity in order to avoid feelings. [Intellectualization Index (high, > 4); or Intellectualization (high, RDS)]

79. This person contends with emotional conflict or stress by compartmentalizing and separating experiences on the basis of how they feel. Although s/he can be aware of different feelings at different times, s/he is not able to experience positive and negative emotions at the same time. As a result, s/he fails to integrate positive and negative qualities into a cohesive picture of him/herself and other people. [Splitting (high, LDS or RDS)]

Reality Testing

87. When this person becomes angry or oppositional, s/he begins to perceive other people or external events in a less accurate fashion. [*S*–% (high, > .39)]

Thought Process

94. Without clear external structure, or under the press of strong feelings, this person's thinking is loose, tangential, rambling, or flighty. [*DR* (high, Level 1 > 2 or Level 2 > 0)]

Thought Content and Preoccupations

99. This person admires aggressive people or objects of strength and power. [Aggressive content (high, GM–AGGR)]

Interpersonal Behaviors

110. This person has a sturdy ability to relate to others. S/he feels autonomous, supports the autonomy of others, and recognizes other people may have different interests and needs than s/he. [MOA (low)]

129. This person establishes relationships that have a merged quality. S/he seems to lose touch with other people's individual distinctiveness, identity, and personal motivations. [Object Relations, Differentiation (low, PRP); or POR (high)]

Interpersonal Beliefs, Representations, and Expectations

132. This person sees him/herself as powerless and ineffectual. S/he believes others are stronger and have more control of how situations turn out. [ROD (high), cognitive interpretation]

(Continued)

TABLE 1 *(Continued)*

145. This person tends to perceive other people in unrealistic ways, such that his/her understanding is based primarily on imaginative or fantasized qualities, rather than upon a complex understanding of their actual characteristics. [*H : (H) + Hd + (Hd)* ratio (low, ≤ 1:1), perception of other people interpretation]

Interpersonal Experiences and Feelings

151. This person has permeable psychological boundaries and experiences his/her personal thoughts and feelings as transparent to others and vulnerable to their influence. [Ego Structure, Boundary (low; PRP)]

152. This person has difficulty distinguishing his/her personal feelings from the emotions of other people. Without recognizing it, s/he generates his/her own conflictual feelings in others, but then believes other people have caused him/her to feel the way s/he does. [Projective Identification (high, LDS or RDS)]

Other Personality Characteristics

166. This person has an "observing ego" which allows him/her to step back from events and take a detached perspective on his/her experience. [*FD* (high, > 1), experimental]

Note. Variables are from the Comprehensive System unless otherwise noted. GM–AGGR = Gacono and Meloy's Aggressive Response; LDS = Lerner and Lerner's Defense Scales; MOA = Urist's Mutality of Autonomy Scale; POR = Kwawer's Primitive Object Relations Scale; PRP = Burke, Friedman, and Gorlitz's Psychoanalytic Rorschach Profile; ROD = Masling and Bornstein's Rorschach Oral Dependence Scale; and RDS = Cooper, Perry, and Arnow's Rorschach Defense Scales.

organization that should assist research. Typically, items within a conceptual category begin with content that is more healthy or adaptive and then move to more problematic characteristics. Items produced in Table 1 are written in a mixed gender format. Although this leaves them suitable for rating male or female participants, it also leaves the wording of many items cumbersome. Consequently, gender specific forms have also been developed.[7]

In Table 1, each item is followed by brackets that indicate the score(s) the item was designed to measure. Although these are the targeted scores for each item, it should be recognized that many items have connotations that apply to other scores as well. Table 1 also provides an indication of the variable's interpretive slant if it can be interpreted in more than one way. For example, the ratio of whole pure human to other human contents (*H* : [*H*] + *Hd* + [*Hd*]) can be interpreted as an indication of a person's self-representation or as an indication of her perception of

[7]Copies of the complete rating scales and a complete table of target constructs can be obtained from any of the authors. To assist research and/or educational goals with the RRS, a table is also available from any of the authors that organizes the items by scoring system.

other people. To facilitate ease of use, these are considered separately in the rating scale. A few items do not correspond to the typical interpretation for a Comprehensive System variable and are listed as "experimental" constructs.

Figure 1 presents the first page of ratings for the Constellation Indices (illustrated with the observer form), which cover the instructions for this section, as well as items for the Coping Deficit Index. As can be seen, three kinds of ratings are made

Global Ratings

In this section, you will evaluate this person in six global areas. First, you will be asked to make a general rating of him/her in each area, then you will be asked to check specific criteria that may apply to his/her condition, and finally you will be asked to make a second general rating based on the specific criteria. For each rating, use the same scale as before.

0	1	2	3	4	T
very uncharacteristic/ definitely false	uncharacteristic/ mostly false		characteristic/ mostly true	very characteristic/ definitely true	tentative (if warranted, use *with* your best estimate)

1.1 This person has social and emotional limitations that make it hard for him/her to cope with the everyday problems of life. These limitations may be expressed in a depressive sense of helplessness and ineffectiveness, or in social difficulties where s/he either relies excessively on others or else disregards and avoids relationships. 0 1 2 3 4 T

Please check which of the following characteristics apply to this person.

a. _____ has limited internal resources which make it hard for him/her to take deliberate and consistent steps when it comes to solving problems.
b. _____ is chronically ineffective in his/her ability to meet the usual demands of daily life.
c. _____ is typically passive rather than active in his/her overall approach to life.
d. _____ does not recognize or use emotions to help orient and guide his/her behavior.
e. _____ has a very limited capacity to express feelings directly.
f. _____ is generally uncomfortable with feelings.
g. _____ deliberately withdraws from situations that are emotionally arousing.

h. _____ has an impoverished understanding of interpersonal interactions.
i. _____ experiences relationships as bland and unmotivated (e.g., sees interactions as neither supportive and enhancing nor hostile and conflictual).
j. _____ is unable to identify easily with other people.
k. _____ are not interested in people.
l. _____ is distant from others and socially isolated.

m. _____ is unable to find the close and emotionally comforting relationships s/he desires, which leads to strong feelings of loneliness or deprivation.
n. _____ is passive in relationships and relies on others to provide him/her with direction, initiative, and security.
o. _____ is psychologically needy or "hungry" for nurturing support.

1.2 Considering just the specific characteristics listed above, rate the extent to which this person has limited coping resources. 0 1 2 3 4 T

FIGURE 1 Example of Rorschach Rating Scale items targeting the constellation indices (observer form, mixed gender format).

for each index. To begin, the rater assigns a score to the global construct. Next, the rater is asked to indicate which constructs from the specific constellation criteria apply to the target person. Finally, the rater is asked to make a second rating of the general construct while only considering the specific criteria enumerated in the preceding section. Table 2 lists the scores that were targeted by each of the items in Figure 1.

A number of criteria within the Constellation Indices have been given a slightly different interpretation from that given when the variable was considered in isolation. This reflects our belief that the exact connotation for a score will vary as a function of other protocol elements. It also reflects our belief that a score's connotation will be slightly different when the score is considered in light of its ability to predict particular symptomatic behaviors. Perhaps a good example is the variable *Zd*. Item 49 in the RRS captures the typical information processing connotations for this variable (i.e., This person processes information in a careful, detailed, and meticulous fashion). Although the core meaning for this variable does not change, we believe that its interpretation should be altered slightly when it is considered as a criterion within the Suicide Constellation or the Hypervigilance Index. When a person is high on *Zd* and also high on other Suicide Constellation variables, it is likely that *Zd* is no longer simply a cognitive/information processing variable. Rather, this variable probably now also reflects a more affectively loaded obsessionality and a propensity to ruminatively pick apart a host of emotionally troubling experiences. In an analogous fashion, the value of the *Zd* for identifying a hypervigilant interpersonal stance may go beyond the standard information

TABLE 2
Rorschach Rating Scale Items Targeting the Coping Deficit Index and Its Subcomponents

1.1 Coping Deficit Index
a. $EA < 6$
b. $AdjD < 0$
c. $p > (a + 1)$
d. $WSumC < 2.5$
e. $WSumC < 2.5$
f. $WSumC < 2.5$
g. Affective Ratio < .46
h. $(COP + AG) < 2$
i. $(COP + AG) < 2$
j. Pure $H < 2$
k. Pure $H < 2$
l. Isolation Index > .24
m. $SumT > 1$
n. $p > (a + 1)$
o. Food > 0
1.2 Coping Deficit Index

processing connotations for this variable. When considered in light of other hypervigilant characteristics, the careful and meticulous processing of the high *Zd* individual is probably motivated by efforts to maintain control and preemptively identify potential threats in the environment.

ISSUES IN THE APPLICATION OF THE RRS

When creating the RRS, the goal was to be inclusive and encompass as many constructs from as many different scoring systems as possible. This approach has several consequences. One disadvantage is that the RRS is currently a rather lengthy scale. As a result, researchers may not wish to use all items in a single study, particularly because there is a degree of redundancy among some of the items and because some items are also reasonable descriptions of scores other than the one initially targeted. Additionally, the extent to which data support the interpretation for each RRS item varies considerably. Some items are much more likely to be wrong than others. In particular, items that require more inferential steps from the response characteristics to the meaning attributed to the response are more likely to contain erroneous assumptions. This applies equally to interpretations that have been inductively generated from observed perceptual qualities (e.g., whole responses, vista responses, movement responses, etc.) and to scoring criteria that have been deduced from theoretical constructs (e.g., idealization responses, denial responses, etc.), although the latter may generally require more inferential steps (Weiner, 1977). On the positive side, the RRS in its current form can serve as a useful tool for generating research hypotheses (e.g., Holaday & Terrell, 1994). The scale is also a useful device for teaching students about the constructs believed to be obtained from the Rorschach method, as it provides a convenient summary of constructs from most of the scoring systems in use today.

Data have not yet been systematically obtained on the "user friendliness" of the RRS. Items were initially written with clinician raters in mind, rather than with the expectation they would be used with peer, spouse, parent, teacher, or self-ratings. Consequently, a number of the items are complex and require some knowledge of psychiatric terminology. Although considerable effort was made to use language that was easy to understand by a wide range of clinicians, it is unclear what level of reading comprehension and/or clinical exposure will be needed to grasp the meaning of each item and complete the form in an accurate fashion. Until such research has been undertaken, users should be cognizant of the potential difficulty nonclinicians may have with some items. In a related fashion, some items almost require an external perspective for accurate judgment. Consequently, it may be unwarranted to expect accurate self-ratings on these items. Nonetheless, they have been retained on the self-report form for the sake of completeness and to ensure uniformity in content across the various forms of the scale.

From a psychometric perspective, researchers using the RRS will need to pay attention to the distribution of test scores and criterion scores to decide what statistics are most appropriate for data analysis (see Exner, 1995). Of equal importance in this regard are theoretical questions about which Rorschach scores should be considered to provide information about continuous traits and states (e.g., extent of depression) and which should be considered as discontinuous indicators of a latent taxon or category (e.g., considering narcissism to be a delimited latent category of personality, like raspberries are a discrete category of fruit; see Meehl, 1995). Indicators of a latent taxon require a critical cutoff score, below which score variability means nothing other than the condition is absent and above which score variability means nothing other than the condition is present. For continuous constructs, correlations can be optimal for quantifying predictor and criterion associations. However, correlations are poor statistics for categorical constructs because all the score variance except for the critical shift from below the cutoff to above the cutoff is measurement error. If this variance is erroneously treated as degrees of trait variance, it may obscure rather than clarify the predictor-criterion relationship.

For methodological reasons, this issue may also present itself at the point of score assignment. Like any other method, the Rorschach task has particular qualities that limit its practical ability to measure psychological constructs. These are determined in part by the nature of the task, in part by the idiosyncracies of articulateness, intelligence, or motivation that participants bring to the task, and in part by scoring rules. Because of some of these qualities, there is likely to be an asymmetrical relationship between the presence of a score and its absence (Weiner, 1977). In general, the presence of a score probably says something more definitive about the presence of a characteristic than the absence of a score says about the absence of a characteristic (particularly if a protocol is coarctated). For instance, a vista response does more to indicate a person is prone to be a critical judge of himself, than no vista does to indicate he is not. In the language of diagnostic efficacy, this would mean individual scores have higher positive predictive power (PPP) than negative predictive power (NPP). If the Rorschach method generally obtains scores that have high PPP but low NPP, statistics that assess a continuous linear association between a test predictor and criterion rating are not optimal—particularly for those scores that do not aggregate well across cards.

Most items on the RRS have been written to target Rorschach scores considered as either continuous variables or as variables with discrete cutoff points. For each item in Table 1, two designations are given within the brackets. The first simply indicates "high" or "low" and reflects which end of the test score continuum the construct captures (e.g., Item 37 reflects the construct believed to be measured when *Lambda* is high, whereas Item 51 reflects the construct believed to be measured

when *Lambda* is low). The second designation is only present for Comprehensive System items and it provides the critical cutoff suggested by Exner (1993) beyond which the construct articulated in the item should be applicable.

Although any written item in the RRS is still an approximation to the full construct and any set of ratings will not capture the "truth" about a patient in an ultimate sense, using the RRS with a LEAD or aggregation strategy should significantly improve validation research. In particular, these procedures will provide strong descriptive data that are capable of falsifying erroneous theory regarding test scores. As such, they should lead directly to a refined understanding of the constructs that are actually quantified by various scores.

Having such criterion information will allow researchers to understand a number of important issues related to the Rorschach. For instance, strong criterion data can be used to determine whether validity coefficients fluctuate in magnitude depending on how engaged patients are with the task (i.e., as a function of the Rorschach's first factor; see Meyer, in press). This can be accomplished by determining whether the validity coefficients are different when scores are obtained from coarctated, dilated, or average complexity protocols. Alternatively, RRS criterion data could be used to determine if controlling first factor variance by requiring a set number of responses per card improves the validity coefficients for some scores while impairing them for others.

Another frequently debated issue is whether Rorschach variables can be considered in isolation or if they must be considered in relation to each other and in the context of the overall protocol. Criterion data generated with the RRS should allow this issue to be settled. One way would be to develop specific hypotheses about what secondary scores should be considered in order to maximize the validity coefficient for a primary target. For instance, one could determine if the validity coefficient for pure color responses and Item 22 of the RRS is higher when the *D Score* is negative than when no consideration is given to the *D Score.* An alternative approach would entail generating both predictor and criterion ratings on the RRS. The criterion ratings would be generated by the LEAD or aggregation procedures discussed earlier. However, the predictor ratings would be generated by an experienced and skilled Rorschach clinician after she considered the Rorschach protocol in its totality. Having this data, one could then determine which predictors—the clinician-generated ratings or the actual Rorschach scores—display higher validity coefficients with the criterion ratings.

A variant on this approach would be to use the RRS items in a Q-sort task. The cognitive benchmarks that raters use may shift as they complete a lengthy scale. In addition, ratings generated across a string of items may not adequately characterize the unique and idiosyncratic features of personality that are so essential to quantify if we are to evaluate the Rorschach's utility as an ideographic clinical tool. Therefore, it may be productive to have judges use the RRS items in a Q-sort, rather

than as a rating scale, by having the items sorted into normally distributed piles of predetermined size along a continuum ranging from *most characteristic* to *least characteristic* of the target subject.

Either of the last two methodologies could also be used to evaluate how important additional sources of information are to the optimal prediction of RRS criterion ratings. For instance, it would be profitable to determine whether validity coefficients are higher when predictor ratings are generated after reviewing Rorschach protocols alone, Rorschach protocols and MMPI–2 profiles, Rorschach protocols and history information, all three data sources in combination, and so on.

Meehl's quote at the outset of this article aptly states the requirements of test validation as being lots of information, with lots of depth, about lots of people. Using the LEAD or aggregation methods will require a considerable commitment if one plans to generate a database that has a sufficiently large subject-to-variable ratio to allow specific validity questions to be answered with confidence. To enable this kind of necessary, high quality research to be carried to completion, the "solution may lie in the coordinated effort of a group of Rorschach researchers" working together in multi-site collaboration (Widiger & Schilling, 1980, p. 458). Overall, those of us who have designed this rating scale hope it will encourage this kind of collaboration and play a useful role in the more refined validation research being conducted on the Rorschach.

ACKNOWLEDGMENT

An earlier version of the rating scale was presented at the meeting of the Society for Personality Assessment, San Francisco, March 1993.

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.

Alexander, S. E., Farber, B. A., Sacco, M., & Trungold-Apter, S. (1995, March). *Utilizing the Rorschach as a measure of positive interpersonal relatedness.* Paper presented at the meeting of the Society for Personality Assessment, Atlanta, GA.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks/Cole.

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.

Blake, S. E., Humphrey, L. L., & Feldman, L. (1994). Self-delineation and marital interaction: The Rorschach predicts Structural Analysis of Social Behavior. *Journal of Personality Assessment, 63,* 148–166.

Blatt, S. J. (1975). The validity of projective techniques and their research and clinical contribution. *Journal of Personality Assessment, 39,* 327–343.

Blatt, S. J., Brenneis, C. B., Schimek, J. G., & Glick, M. (1976). Normal development and psychopathological impairment of the concept of the object on the Rorschach. *Journal of Abnormal Psychology, 85,* 364–373.

Blatt, S. J., & Ritzler, B. A. (1974). Suicide and the representation of transparency and cross-sections on the Rorschach. *Journal of Consulting and Clinical Psychology, 42,* 280–287.

Bornstein, R. F. (1993). *The dependent personality.* New York: Guilford.

Bornstein, R. F., Rossner, S. C., Hill, E. L., & Stepanian, M. L. (1994). Face validity and fakability of objective and projective measures of dependency. *Journal of Personality Assessment, 63,* 363–386.

Burke, W. F., Friedman, G., & Gorlitz, P. (1988). The psychoanalytic Rorschach profile: An integration of drive, ego, and object relations perspectives. *Psychoanalytic Psychology, 5,* 193–212.

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

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.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* San Diego, CA: Academic.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston: Houghton Mifflin.

Cooper, S., Perry, J., & Arnow, D. (1988). An empirical approach to the study of defense mechanisms: I. Reliability and preliminary validity of the Rorschach defense scale. *Journal of Personality Assessment, 52,* 187–203.

Cooper, S. H., Perry, J. C., & O'Connell, M. (1991). The Rorschach defense scales: II. Longitudinal perspectives. *Journal of Personality Assessment, 56,* 191–201.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

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

Exner, J. E., Jr. (1980). But it's only an inkblot. *Journal of Personality Assessment, 44,* 562–577.

Exner, J. E., Jr. (1991). *The Rorschach: A comprehensive system: Vol. 2. Interpretation* (2nd ed.). New York: Wiley.

Exner, J. E., Jr. (1993). *The Rorschach: A comprehensive system: Vol. 1. Basic foundations.* (3rd ed.). New York: Wiley.

Exner, J. E., Jr. (Ed.). (1995). *Issues and methods in Rorschach research.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Exner, J. E., Jr., & Andronikof-Sanglade, A. (1992). Rorschach changes following brief and short-term therapy. *Journal of Personality Assessment, 59,* 59–71.

Gacono, C. B. (1990). An empirical study of object relations and defensive operations in antisocial personality. *Journal of Personality Assessment, 54,* 589–600.

Gacono, C. B., & Meloy, J. R. (1994). *The Rorschach assessment of aggressive and psychopathic personalities.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Gacono, C. B., Meloy, J. R., & Berg, J. L. (1992). Object relations, defensive operations, and affective states in narcissistic, borderline, and antisocial personality disorder. *Journal of Personality Assessment, 59,* 32–49.

Garwood, J. (1977). A guide to research on the Rorschach Prognostic Rating Scale. *Journal of Personality Assessment, 41,* 117–119.

Harkness, A. R., Tellegen, A., & Waller, N. (1995). Differential convergence of self-report and informant data for Multidimensional Personality Questionnaire traits: Implications for the construct of negative emotionality. *Journal of Personality Assessment, 64,* 185–204.

Holaday, M., & Terrell, D. (1994). Resiliency characteristics and Rorschach variables in children and adolescents with severe burns. *Journal of Burn Care and Rehabilitation, 15,* 455–460.

Kleinmuntz, B. (1990). Why we still use our heads instead of formulas: Toward an integrative approach. *Psychological Bulletin, 107,* 296–310.

Kwawer, J. S. (1980). Primitive interpersonal modes, borderline phenomena, and Rorschach content. In J. Kwawer, H. Lerner, & A. Sugarman (Eds.), *Borderline phenomena and the Rorschach test* (pp. 89–106). New York: International Universities Press.

LaBarbera, J. D., & Cornsweet, C. (1985). Rorschach predictors of therapeutic outcome in a child psychiatric inpatient service. *Journal of Personality Assessment, 49,* 120–124.

Leifer, M., Shapiro, J. P., Martone, M. W., & Kassem, L. (1991). Rorschach assessment of psychological functioning in sexually abused girls. *Journal of Personality Assessment, 56,* 14–28.

Lerner, H., & Lerner, P. (1980). Rorschach assessment of primitive defenses in borderline personality structure. In J. Kwawer, H. Lerner, P. Lerner, & A. Sugarman (Eds.), *Borderline phenomena and the Rorschach test* (pp. 257–274). New York: International Universities Press.

Lerner, H., & Lerner, P. (1982). A comparative study of defensive structure in neurotic, borderline, and schizophrenic patients. *Psychoanalysis and Contemporary Thought, 5,* 77–113.

Lerner, P. M. (1990). Rorschach assessment of primitive defenses: A review. *Journal of Personality Assessment, 54,* 30–46.

Marsh, A., & Viglione, D. J. (1992). A conceptual validation study of the texture response on the Rorschach. *Journal of Personality Assessment, 58,* 571–579.

Masling, J. M., Rabie, L., & Blondheim, S. H. (1967). Obesity, level of aspiration, and Rorschach and TAT measures of oral dependence. *Journal of Consulting Psychology, 31,* 233–239.

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.

McCown, W., Fink, A. D., Galina, J., & Johnson, J. (1992). Effects of laboratory-induced controllable and uncontrollable stress on Rorschach variables *m* and *Y*. *Journal of Personality Assessment, 59,* 564–573.

Meehl, P. E. (1959). Some ruminations on the validation of clinical procedures. *Canadian Journal of Psychology, 13,* 102–128.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Meehl, P. E. (1995). Bootstraps taxometrics: Solving the classification problem in psychopathology. *American Psychologist, 50,* 266–275.

Meyer, G. J. (1992a). The Rorschach's factor structure: A contemporary investigation and historical review. *Journal of Personality Assessment, 59,* 117–136.

Meyer, G. J. (1992b). Response frequency problems in the Rorschach: Clinical and research implications with suggestions for the future. *Journal of Personality Assessment, 58,* 231–244.

Meyer, G. J. (in press). On the integration of personality assessment methods: The Rorschach and MMPI–2. *Journal of Personality Assessment.*

Meyer, G. J., Murphy, S. G., Kiolbasa, T., Daly, M., Orput, D., Zelko, F. A. J., Remer-Osborn, J., & Zillmer, E. A. (1993, March). *Neuropsychological factors and Rorschach performance in children.* Paper presented at the meeting of the Society for Personality Assessment, San Francisco.

Parker, K. C. H., Hanson, R. K., & Hunsley, J. (1988). MMPI, Rorschach, and WAIS: A meta-analytic comparison of reliability, stability, and validity. *Psychological Bulletin, 103,* 367–373.

Paunonen, S. V. (1989). Consensus in personality judgments: Moderating effects of target-rater acquaintanceship and behavior observability. *Journal of Personality and Social Psychology, 56,* 823–833.

Perry, G. G., & Kinder, B. N. (1990). The susceptibility of the Rorschach to malingering: A critical review. *Journal of Personality Assessment, 54,* 47–57.

Perry, W., McDougall, A., & Viglione, D. J. (1995). A five-year follow-up on the temporal stability of the Ego Impairment Index. *Journal of Personality Assessment, 64,* 112–118.

Perry, W., Sprock, J., Schaible, D., McDougall, A., Minassian, A., Jenkins, M., & Braff, D. (1995). Amphetamine on Rorschach measures in normal subjects. *Journal of Personality Assessment, 64,* 456–465.

Perry, W., & Viglione, D. J. (1991). The Ego Impairment Index as a predictor of outcome in melancholic depressed patients treated with tricyclic antidepressants. *Journal of Personality Assessment, 56,* 487–501.

Pilkonis, P. A., Heape, C. L., Ruddy, J., & Serrao, P. (1991). Validity in the diagnosis of personality disorders: The use of the LEAD standard. *Psychological Assessment, 3,* 46–54.

Ribeiro, S. C. M., Tandon, R., Grunhaus, L., & Greden, J. F. (1993). The DST as a predictor of outcome in depression: A meta-analysis. *American Journal of Psychiatry, 150,* 1618–1629.

Rydin, E., Asberg, M., Edman, G., & Schalling, D. (1990). Violent and nonviolent suicide attempts: A controlled Rorschach study. *Acta Psychiatrica Scandinavica, 82,* 30–39.

Rydin, E., Schalling, D., & Asberg, M. (1982). Rorschach ratings in depressed and suicidal patients with low levels of 5-hydroxyindoleacetic acid in cerebrospinal fluid. *Psychiatry Research, 7,* 229–243.

Shalit, B. (1965). Effects of environmental stimulation on the M, FM, and m responses in the Rorschach. *Journal of Projective Techniques and Personality Assessment, 29,* 228–231.

Spigelman, A., & Spigelman, G. (1991). Indications of depression and distress in divorce and nondivorce children reflected by the Rorschach test. *Journal of Personality Assessment, 57,* 120–129.

Spitzer, R. L. (1983). Psychiatric diagnosis: Are clinicians still necessary? *Comprehensive Psychiatry, 24,* 399–411.

Tellegen, A. (1991). Personality traits: Issues of definition, evidence, and assessment. In D. Cicchetti & W. Grove (Eds.), *Thinking clearly about psychology: Essays in honor of Paul Everett Meehl* (Vol. 2, pp. 10–35). Minneapolis: University of Minnesota Press.

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44,* 703–742.

Thomas, C. B., & Duszynski, K. R. (1985). Are words of the Rorschach predictors of disease and death? The case of "whirling." *Psychosomatic Medicine, 47,* 201–211.

Tsujimoto, R. N., Hamilton, M., & Berger, D. E. (1990). Averaging multiple judges to improve validity: Aid to planning cost-effective clinical research. *Psychological Assessment, 2,* 432–437.

Tuber, S. B. (1983). Children's Rorschach scores as predictors of later adjustment. *Journal of Consulting and Clinical Psychology, 51,* 379–385.

Tuber, S. (1992). Empirical and clinical assessments of children's object relations and object representations. *Journal of Personality Assessment, 58,* 179–197.

Urist, J. (1977). The Rorschach test and the assessment of object relations. *Journal of Personality Assessment, 41,* 3–9.

Urist, J., & Shill, M. (1982). Validity of the Rorschach mutuality of autonomy scale: A replication using excerpted responses. *Journal of Personality Assessment, 46,* 450–454.

Weiner, I. B. (1977). Approaches to Rorschach validation. In M. A. Rickers-Ovsiankina (Ed.), *Rorschach psychology* (pp. 575–608). Huntington, NY: Krieger.

Weiner, I. B. (1991). Editor's note: Interscorer agreement in Rorschach research. *Journal of Personality Assessment, 56,* 1.

Weiner, I. B. (1994). The Rorschach inkblot method (RIM) is not a test: Implications for theory and practice. *Journal of Personality Assessment, 62,* 498–504.

Weiner, I. B., & Exner, J. E., Jr. (1991). Rorschach changes in long-term and short-term psychotherapy. *Journal of Personality Assessment, 56,* 453–465.

Wenar, C., & Curtis, K. M. (1991). The validity of the Rorschach for assessing cognitive and affective changes. *Journal of Personality Assessment, 57,* 291–308.

Widiger, T. A., & Schilling, K. M. (1980). Toward a construct validation of the Rorschach. *Journal of Personality Assessment, 44,* 450–459.

Wood, J. M., Nezworski, M. T., & Stejskal, W. J. (1996). The Comprehensive System for the Rorschach: A critical examination. *Psychological Science, 7,* 3–10.

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

Received February 5, 1996
Revised February 28, 1996