CHAPTER 8

An Introduction to Rorschach Assessment[1]

GREGORY J. MEYER
DONALD J. VIGLIONE

Introduction

The Rorschach is a performance-based task or behavioral assessment measure[2] that assesses a broad range of personality, perceptual, and problem-solving characteristics, including thought organization, perceptual accuracy and conventionality, self-image and understanding of others, psychological resources, schemas, and dynamics. The task provides a standard set of inkblot stimuli, and is administered and coded according to standardized guidelines. In many respects, the task is quite simple. It requires clients to identify what a series of richly constructed inkblots look like in response to the query, "What might this be?" Despite its seeming simplicity, the solution to this task is quite complex, as each inkblot provides myriad response possibilities that vary across multiple stimulus dimensions. Solving the problem posed in the query thus invokes a series of perceptual problem-solving operations related to scanning the stimuli, selecting locations for emphasis, comparing potential inkblot images to mental representations of objects, filtering out responses judged less optimal, and articulating those selected for emphasis to the examiner. This process of explaining to another person how one looks at things against a backdrop of multiple competing possibilities provides the foundation for the Rorschach's empirically demonstrated validity. Unlike interview- based measures or self-report inventories, the Rorschach does not require clients to describe what they are like but rather it requires them to

provide an in vivo illustration of what they are like by repeatedly providing a sample of behavior in the responses generated to each card. Each response or solution to the task in this overall behavior sample is coded across a number of dimensions and the codes are then summarized into scores by aggregating the codes across all responses. By relying on an actual sample of behavior collected under standardized conditions, the Rorschach is able to provide information about personality that may reside outside of the client's immediate or conscious awareness. Accessing information obtained from observing a client's personality in action can be a considerable and unique asset for clinicians engaged in the idiographic challenge of trying to understand a person in her or his full complexity.

The Rorschach is taught in about 80% of United States doctoral clinical psychology programs (Childs & Eyde, 2002; Hilsenroth & Handler, 1995; Mihura & Weinle, 2002). Internship training directors expect incoming students to have good working knowledge of the Rorschach (Clemence & Handler, 2001), and it ranks third in importance for them after the Wechsler Adult Intelligence Scale (WAIS-III; Wechsler, 1997) and the Minnesota Multiphasic Personality Inventory (MMPI-2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989). Among doctoral students in training, Mihura and Weinle (2002) found the Rorschach was viewed as most useful for understanding a client's personality. Their survey showed students were more satisfied with it and anticipated using it more in the future when they had more didactic and practical experience with it, more familiarity with its empirical literature, and more positive attitudes toward it in their training program. Among clinical psychologists in practice, the Rorschach is typically the third or fourth most commonly used assessment instrument, following the WAIS and MMPI (Camara, Nathan, & Puente, 2000; Watkins, Campbell, Nieberding, & Hallmark, 1995). The same rank ordering has been found internationally in a survey of psychologists in Spain, Portugal, and Latin American countries (Muniz, Prieto, Almeida, & Bartram, 1999). With respect to its research base, the Rorschach has been the second most investigated personality assessment instrument (following the MMPI), with about 7,000 citations in the literature as of the mid-1990s (Butcher & Rouse, 1996).

Although the Rorschach is frequently taught in graduate programs, valued on internship and in clinical practice, and regularly researched, it also has generated notable controversy throughout much of its history. Why is this? Although we cannot provide a definitive explanation, we provide insight into some of the key research relevant to its use as part of evidence based practice. In the process, we address several critical questions that have been raised over the last decade about the Rorschach. These include: (a) What does the evidence show about the reliability of Rorschach scores, (b) what strengths and limitations are present in the evidence for the construct validity

and utility of its scales, (c) does the instrument have a reasonable base of normative data, (d) can it reasonably be applied across cultures, and (e) does the evidence suggest certain modifications should be made to traditional interpretive postulates?

Because it is not possible to learn how to do Rorschach administration, scoring, and interpretation by reading a single book chapter, we assume that readers interested in gaining applied proficiency with the instrument will rely on other resources. As such, even though we provide readers with a general understanding of the Rorschach and how it is administered, scored, and interpreted, our goal in this chapter is to emphasize the psychometric evidence and issues associated with its use.

Theory and Development

The Rorschach consists of inkblot stimuli[3] that were created, artistically refined, and studied by Herman Rorschach from 1917 to 1920. Exner (2003) provides an overview of their development, which we briefly summarize here. The final set of 10 stimuli was first published in 1921 (Rorschach, 1921/1942). Before publication, Rorschach experimented with 40 or more inkblots, many of which appear to be less complex, nuanced, and detailed precursors to the final set. Figure 8.1 is an example of one of these inkblots; it appears to be an early version of what is now the second inkblot. Rorschach developed his task largely as a means to understand and diagnose Bleuler's newly described syndrome of schizophrenia. Rorschach's doctoral dissertation, which did not focus on inkblots, examined hallucinations in schizophrenia and it was directed by Bleuler. In 1917 another of Bleuler's students, Szymon Hens, completed a dissertation that used eight inkblots he created to determine the content-based distinctions observed among 1,000 children, 100 adults, and 100 patients with psychoses. Rorschach was more interested in perceptual processes than content per se and thus pursued a different direction in his own research. Most of Rorschach's research took place with 12 inkblots, though he was forced to give up 2 to secure a publisher. All 10 of the final inkblots appear to have been artistically embellished by Rorschach, who added details, contours, and colors "to ensure that each figure contained numerous distinctive features that could easily be identified as similar to objects stored in the memory traces of the individual" (Exner, 2003, p. 8). Thus, despite common belief to the contrary, the images are not arbitrary, haphazard, or accidental inkblots. Instead, they are purposively altered images that were refined through trial and error experimentation to elicit informative responses. Each inkblot has a white background; five are achromatic (i.e., gray or black) color only, two are in red and achromatic color, and three are in an array of pastel colors without any black. During the initial printing process, gradations in color and shading became accentuated. Although initially dissatisfied,

Figure 8.1 Early inkblot for possible use created by Hermann Rorschach. (Used with permission of the Hermann Rorschach Archives and Museum.)

Rorschach concluded that this unexpected change offered new possibilities for capturing individual differences in perceptual operations.

Rorschach died in 1922, just 7 months after his book was published. Over the next 40 years, different systems of administration, scoring, and interpretation developed. In the early 1970s, Exner (1974, 2003) developed what he called the Rorschach Comprehensive System (CS), which synthesized what he believed were the most reliable and valid elements of the five primary systems in the United States—those developed by Samuel Beck, Marguerite Hertz, Bruno Klopfer, Zygmunt Piotrowski, and David Rapaport. Since that time, the CS has become the dominant approach to administration, scoring, and interpretation in the United States (Hilsenroth & Handler, 1995; Mihura & Weinle, 2002) and it is widely used internationally (e.g., in Argentina, Belgium, Brazil, Denmark, Finland, France, Holland, Japan, Israel, Italy, Norway, Peru, Portugal, Sweden, and Spain; see Butcher, Nezami, & Exner, 1998; Erdberg & Shaffer, 1999).

A wide array of formal variables can be coded on the Rorschach, though clinicians also draw personality inferences based on numerous response features and testing behaviors that are not formally coded (e.g., Aronow, Reznikoff, & Moreland, 1995; Exner & Erdberg, 2005; Fischer, 1994; Peebles-Kleiger, 2002; Weiner, 2003). With respect to coded variables, there are a large number of scales and indexes described in the literature that are not included in the CS, and many of them have accumulated substantial evidence of reliability and validity (see, e.g., Bornstein & Masling, 2005). Not surprisingly, a range of test construction models have influenced the formal coding criteria for these scales, including those in the CS.

Scale development procedures can be considered on a dimension that ranges from purely empirical, in which items are selected based on statistical

relationships with a criterion regardless of whether they make conceptual sense, to fully rational, in which items are selected based on logic and a theoretical understanding of the construct to be measured regardless of whether there is statistical evidence to support that belief. Adopting this framework and applying it to the Rorschach, the empirical end of the continuum would be anchored by some of the actuarial indexes found on the CS, such as the Perceptual Thinking Index (PTI) and the Suicide Constellation (S-CON). Although both indexes were influenced to some extent by theory, they were developed primarily by atheoretical empirical findings using discriminant function analyses in a contrasted groups design (Exner, 2003).

Other indexes were developed using a combined rational and empirical approach. For instance, the developers of the CS-based Ego Impairment Index (EII-2; Perry & Viglione, 1991; Viglione, Perry, & Meyer, 2003) initially identified variables that both had empirical research support and theoretically should be related to impaired object relations and ego functioning. These scores were then refined to create the final scale by using factor analysis and regression-based factor scores to differentially weigh the relative contribution of each variable.

A bit further on the continuum toward the rational end are scores that are largely defined by a theoretical model but that are also refined and specified in such a way that they take into account the unique qualities and limitations associated with the Rorschach inkblot stimuli. The CS Good and Poor Human Representation variables (GHR and PHR; Perry & Viglione, 1991; Viglione, Perry, Jansak, Meyer, Exner, 2003) are good examples. These indexes are founded on object relations theories in which healthy functioning is defined by perceptions of self and others that are complete, accurate, realistic, intact, independent, and generally benevolent or supportive as opposed to partial, distorted, confused, damaged, enmeshed or fused, and generally malevolent or aggressive. From a theoretical perspective, the healthiest object relations are those in which human others are perceived accurately as whole and complete figures that are not embellished with mythic or fictionalized attributes. However, the Rorschach stimuli provide limited opportunities to observe such objects (i.e., there are relatively few places in the ten inkblots where it is conventional to see a complete person). Consequently, the GHR and PHR scoring algorithms take into account instances when it is typical for people to perceive nonhuman or partial human figures in specific inkblot locations.

At the rational end of the empirical versus rational continuum are scales created by theory that do not make special provisions for the stimulus pull of specific Rorschach inkblots. A good example is the Rorschach Oral Dependency scale (ROD; Bornstein, 1996, 1998, 1999; Masling, Rabie, & Blondheim, 1967), which is a well-validated measure of dependency based

on response content. The coding criteria are theoretically derived from the psychodynamic construct of orality (Schafer, 1954) and include imagery such as food sources, oral activity, nurturance, passivity, and helplessness. Another example is Blatt's Concept of the Object Scale (COS; Blatt, Brenneis, Schimek, & Glick, 1976). Like the GHR and PHR scores, the COS is based on object relations theory. However, unlike GHR and PHR, the COS coding criteria are derived entirely from theorizing about developmental processes; they do not make allowances for the stimulus pull of the individual inkblots and the extent to which that pull produces typical responses that do not conform to theory. As a result, some of the things that people typically or normatively see on the Rorschach receive less healthy COS scores than do perceptions that are normatively atypical or unusual. For instance, the stimulus features of Cards IV and IX pull for people to see quasi-human or human-like figures (e.g., a monster or a wizard) rather than ordinary people. Even though these responses are so common they are considered "Popular," the COS assigns them a less than optimal score because the latter is reserved for human beings.

There are at least three other models for understanding types of Rorschach scores; those that are founded on (1) simple classification, (2) clinical observation, and (3) behavioral similarity. The first is the least important. These are response features that are coded primarily to exhaust a coding category. Probably the best examples are some of the content codes in the CS. Every response is coded for the content it contains, though not all of the content categories are interpretively valuable. For instance, the CS has separate categories for household objects, science based percepts, botany as distinct from landscape content, and an idiographic category for not otherwise classifiable objects. None of these distinctions factor into standard interpretation.

Clinical observation is a form of empirical keying, in that response features are linked to personality characteristics through clinical experience even if there is no obvious parallel between the response feature and the characteristic that is thought to be indicated by the score. As an example, clinical observation suggested that the perception of moving inanimate objects (an *m* score) is associated with environmental stress, internal tension, agitated cognitive activity, and loss of control, while responses that are prompted by the general shading features in the ink (*Y* scores) are associated with disruptive experiences of anxiety or helplessness. In each example there are nonobvious links between the score and the construct that it is hypothesized to measure. The big difference between scores based on clinical observations and those based on empirical keying is that the former may or may not demonstrate empirical relationships when actually tested. However, both of the example scores (*m* and *Y*) have replicated data supporting their construct validity (e.g., Hartmann, Nørbech, & Grønnerød, 2006; Hartmann, Sunde, Kristensen, & Martinussen, 2003; Hartmann, Wang, Berg, Sæther, 2003;

McCowan, Fink, Galina, & Johnson, 1992; Nygren, 2004; Perry et al., 1995; Sultan, Jebrane, & Heurtier-Hartemann, 2002). As has been the case for *m* and *Y*, other clinical observation scores that garner empirical support over time also typically develop an experiential explanation or theory that links the observed test behavior to the criterion construct. For instance, in hindsight it is now not too difficult to see how at an experiential level a person who feels considerable stress, tension, and agitation may see an elevated number of nonliving objects in motion (e.g., percepts of objects exploding, erupting, falling, spinning, tipping, or shooting).

Finally, many Rorschach scores are rationally constructed "behavioral representation" scores, in that the response characteristic coded in the testing situation closely parallels the real-life behavior that it is thought to measure (Weiner, 1977). That is, what is coded in the microcosm of the test setting is a representative sample of the behavior or experience that one expects to be manifested in the macrocosm of everyday life (Viglione & Rivera, 2003). For instance, the CS morbid score (MOR) is coded when dysphoric or sad affect is attributed to an object or when an object is described as dead, injured, or damaged in some manner. When responses of this type occur fairly often, they are thought to indicate a sense of gloomy, pessimistic inadequacy. Thus, the behavior coded in the testing situation is thought to be representative of the dysphoric, negative, damaged mental set that the person generally uses to interpret and filter life experiences. Similarly, the CS cooperative movement scores (COP) is coded when two or more objects are described as engaging in a clearly cooperative or positive interaction. Higher COP scores are thought to assess a greater propensity to conceptualize relationships as supportive and enhancing.

Probably the most well-known and best-validated behavioral representation scores on the Rorschach are the indicators of disordered thought and reasoning. In the CS these are called the Cognitive Special Scores and they are coded in a number of instances, including when responses are circumstantial or digressive, when objects have an implausible or impossible relationship (e.g., two chickens lifting weights), and when reasoning is strained or overly concrete. In all these examples, the coded test behavior represents the extra-test characteristic it is thought to measure. Thus, behavioral representation scores require relatively few inferential steps to link what is coded on the test to everyday behavior.

Basic Psychometrics

Reliability

Reliability is the extent to which a construct is assessed consistently. Once assessed consistently, it is necessary to establish that what is being measured is actually what is supposed to be measured (validity) and that the measured

information is helpful in some applied manner (utility). We briefly address each issue; more details can be found in Meyer (2004) and Viglione and Meyer (2007).

There are four main types of reliability: internal consistency, split half or alternate forms, test-retest, and interrater. Internal consistency reliability examines item-by-item uniformity in content to determine whether the items of a scale all measure the same thing (Streiner, 2003a, 2003b). Split-half and alternate forms reliability operate at a more global level; they examine consistency in total scores across parallel halves of a test or parallel versions of a full length test. They allow for some item-by-item heterogeneity because they evaluate whether the composite of information on each form of the test produces a consistent and equivalent score. Although there are exceptions (e.g., Bornstein, Hill, Robinson, Calabreses, & Bowers, 1996; Dao & Prevatt, 2006), researchers typically do not investigate split-half and alternate forms reliability with the Rorschach because each Rorschach card and even each location within a card has its own distinct stimulus properties that pull for particular kinds of variables (Exner, 1996). For instance, the cards vary in the extent to which they are unified versus fragmented, shaded, colored, and so on. As a result, each item on the test, whether defined as each response to the test or as the responses to each card on the test, is not equivalent and internal consistency analyses are generally considered inapplicable. The same factors make it impossible to split the inkblots into truly parallel halves or to produce an alternative set of inkblots that have stimulus properties equivalent to the original.

Somewhat different issues affect internal consistency analyses of the CS Constellation Indexes (e.g., Dao & Prevatt, 2006). There are six of these indexes; the Perceptual-Thinking Index (PTI), the Depression Index (DEPI), the Coping Deficit Index (CDI), the Hypervigilance Index (HVI), the Obsessive Style Index (OBS), and the Suicide Constellation (S-CON). These indexes were created as heterogeneous composite measures to maximize validity, not as homogeneous scales of a single construct, which makes internal consistency reliability largely immaterial (Streiner, 2003a). Psychometrically, predictive validity is maximized by combining unique and nonredundant sources of information, so strong validity can occur despite weak internal consistency reliability, even with a short and simple measure.

Test-retest or temporal consistency reliability evaluates the stability of scores over time to repeated administrations of the same instrument. Temporal consistency has been studied fairly often with the Rorschach, and Grønnerød (2003) recently conducted a systematic meta-analysis of this literature. The results show acceptable to good stability for Rorschach scores, including for the CS (also see Meyer & Archer, 2001; Viglione & Hilsenroth, 2001). For the CS and other systems, scores thought to measure more trait-like aspects

of personality have produced relatively high retest coefficients, even over extended time periods, while scores thought to reflect state-like emotional process have produced relatively low retest coefficients even over short time intervals. Grønnerød found that across all types of Rorschach scores and over an average retest interval of slightly more than 3 years (38 months), the average reliability was $r = .65$ using data from 26 samples ($N = 904$). Meyer (2004) organized results from all the meta-analyses of test-retest reliability in psychology, psychiatry, and medicine that had been published through 2001. Grønnerød's results compare favorably to the stability of other characteristics included in that review, including self-reported Big Five personality traits ($r = .73$ over 1.6 years); personality disorder diagnoses (kappa = .44 over 7.1 months); disorganized parent-child attachment patterns ($r = .34$ over 2.1 years); and the extent to which the same professionals in medicine, psychology, business, meteorology, and human resources make consistent judgments over time about the same information ($r = .76$ over 2.9 months).

Although these meta-analytic results indicate the stability of Rorschach scores compares favorably to other variables, a recent well-designed French study examining CS stability found lower than anticipated consistency over a 3-month retest period (Sultan, Andronikof, Réveillère, & Lemmel, 2006). A factor that may influence stability is the overall complexity of a person's protocol when tested on both occasions. The two variables that index the overall richness or complexity of a protocol are *R*, the number of responses, and Lambda (or PureForm%), which indicates the proportion of responses prompted by relatively simple form features rather than other more subtle or complex qualities of the inkblot. In the Sultan et al. (2006) study, stability coefficients for these variables were .75 and .72, respectively. Because these variables are excellent markers of the primary source of variance in Rorschach scores (i.e., the first dimension in factor analysis; see Meyer, Riethmiller, Brooks, Benoit, & Handler, 2000), when they are unstable, most other scores also will be unstable. Indeed, this is what Sultan et al. observed; the median 3-month stability coefficient across 87 ratios, percentages, and derived scores that are emphasized in interpretation was .55. Although lower than expected or desired, this level of stability is similar to that observed with memory tests and job performance measures (Viglione & Meyer, 2007). Perhaps not surprisingly, Sultan et al. found that stability was moderated by *R* and Lambda; it was higher when people had values that did not change much over time and lower among those with values that did change. Although more research on Rorschach stability is needed and Sultan et al.'s findings should be replicated, their results indicate that generally healthy people who volunteer for a study can provide noticeably different protocols when tested by one reasonably trained examiner and again 3 months later by a different reasonably trained examiner.

The final type of reliability is inter-rater reliability, which assesses the consistency of judgments across raters. For the Rorschach, this type of reliability concerns scoring reliability as well as the reliability of interpretation across clinicians. Rorschach scoring reliability has been studied regularly and there are four meta-analyses summarizing this literature. Two of them were related studies addressing CS reliability (Meyer, 1997; Meyer et al., 2002) and the other two addressed the Rorschach Prognostic Rating Scale and the Rorschach Oral Dependency scale (see Meyer, 2004). The meta-analyses indicate that reasonably trained raters achieve good reliability, with average Pearson or intraclass correlations (ICCs) for summary scores above .85 and average kappa values for scores assigned to each response above .80.[4] Meyer (2004) compared Rorschach interrater reliability data to all other published meta-analyses of interrater reliability in psychology, psychiatry, and medicine, and the data showed it compared favorably to a wide range of other applied judgments. For instance, Rorschach raters agree more than supervisors evaluating the job performance of employees ($r = .57$), surgeons or nurses diagnosing breast abnormalities on a clinical exam (kappa = .52), and physicians evaluating the quality of medical care provided by their peers (kappa = .31). For many Rorschach variables, scoring shows the same degree of reliability as when physicians estimate the size of the spinal canal and spinal cord from MRI, CT, or X-Ray scans ($r = .90$); dentists and dental personnel count decayed, filled, or missing teeth in early childhood (kappa = .79); or when physicians or nurses rate the degree of drug sedation for patients in intensive care ($r = .91$, ICC = .84). These comparisons show that Rorschach coding for trained examiners is typically fairly straightforward and agreement is attainable across raters.

At the same time, there are challenges or difficulties associated with Rorschach scoring. Several studies show how the reliabilities for low base rate variables are erratic (e.g., Acklin, McDowell, & Verschell, 2000; McGrath et al., 2005; Meyer et al., 2002; Viglione & Taylor, 2003). Roughly speaking, low base rate variables occur on average once or less often per record (i.e., in < 5% of responses; e.g., sex, reflections, color projection), so that large samples are needed to accurately estimate their reliability. In addition, there are some more common codes that generally show lower reliability and thus appear to be more challenging to code accurately (e.g., types of shading; the extent to which form is primary, secondary, or absent when coded in conjunction with color or shading responses; differentiating botany, landscape, and nature contents; classifying specific types of cognitive disorganization). Viglione (2002) developed a coding workbook that addresses these issues.

Students learning Rorschach assessment also need to realize that interrater reliability is not a fixed property of the score or test instrument. Rather, it is entirely dependent on the training, skill, and conscientiousness of the

examiner. Thus, repeated practice and calibration with criterion ratings are essential for good practice.

Another issue is that most reliability research (for the Rorschach and for other instruments) relies on raters who work or train in the same setting. To the extent that local guidelines develop to contend with scoring ambiguity, agreement among those who work or train together may be greater than agreement across different sites or workgroups. As a result, existing reliability data may then give an overly optimistic view of scoring consistency across sites or across clinicians working independently. Another way to say this is that scoring reliability (i.e., agreement among two fallible coders) may be higher than scoring accuracy (i.e., correct coding).

This issue was recently examined for the CS. In a preliminary report of the data, Meyer, Viglione, Erdberg, Exner, and Shaffer (2004) examined 40 randomly selected protocols from Exner's new CS nonpatient reference sample (Exner & Erdberg, 2005) and 40 protocols from Shaffer, Erdberg, and Haroian's (1999) nonpatient sample from Fresno, California. These 80 protocols were then blindly recoded by a third group of advanced graduate students who were trained and supervised by the second author. To determine the degree of cross-site reliability, the original scores were compared to the second set of scores. The data revealed an across site median ICC of .72 for summary scores. Although this would be considered "good" reliability according to established benchmarks, it is lower than the value of .85 or higher that typically has been generated by coders working together in the same setting.

Findings like this suggest there are complexities in the coding process that are not fully clarified in standard CS training materials (Exner, 2001, 2003). As a result, training sites, such as specific graduate programs, may develop guidelines or benchmarks for coding that help resolve these residual complexities. However, these principles may not generalize to other training sites. To minimize these problems, students learning CS scoring should find Viglione's (2002) coding text helpful and should thoroughly practice their scoring relative to the across-site gold standard scores that can be found in the 300 practice responses in Exner's (2001) workbook and in the 25 cases with complete responses in the basic CS texts (Exner, 2003; Exner & Erdberg, 2005).

Beyond agreement in scoring the Rorschach, an important question is the extent to which clinicians show consistency in the way they interpret Rorschach results. Interclinician agreement when interpreting psychological tests (not just the Rorschach) was studied fairly often in the 1950s and 1960s, though it then fell out of favor (Meyer, Mihura, & Smith, 2005). The reliability of Rorschach interpretation in particular has been challenged, with some suggesting that the inferences clinicians generated said more about them than

about the client being assessed. To examine agreement on CS interpretations, Meyer et al. (2005) had 55 patient protocols interpreted by three to eight clinicians across four data sets. A total of 20 different clinicians participated in the research. Consistency was assessed across a representative set of 29 personality characteristics (e.g., "This person experiences himself as damaged, flawed, or hurt by life."). Substantial reliability was observed across all the data sets, with aggregated judgments having higher agreement ($M\ r = .84$) than judgments to individual interpretive statements ($M\ r = .71$). As Meyer et al. (2005) illustrated, these findings compared favorably to meta-analytic summaries of interrater agreement for other types of applied judgments in psychology, psychiatry, and medicine. For instance, therapists or observers ratings the quality of the therapeutic alliance in psychotherapy produce an average agreement of $r = .78$, while neurologists classifying strokes produce an average agreement of kappa = .51.

At the same time, it was also clear that some clinicians were more reliable than others. For aggregated judgments, the average reliability among the three most consistent judges was $r = .90$ and among the three least consistent judges it was $r = .73$. Thus, the findings indicated that experienced clinicians could reliably interpret CS data; when presented with the same Rorschach data, they drew similar conclusions about patients. However, some clinicians were clearly more consistent than others, which highlights how one needs to conscientiously learn principles of interpretation and then carefully and systematically consider all relevant testing data when conducting an idiographic clinical assessment.

Validity

Construct validity refers to evidence that a test scale is measuring what it is supposed to measure. It is determined by the conglomerate of research findings related to both convergent and discriminant validity. Convergent validity refers to expected associations with criteria that theoretically should be related to the target construct, while discriminant validity refers to an expected lack of association with criteria that theoretically should be independent of the target construct. Evaluating the validity of a complex, multidimensional measure like the Rorschach is challenging because it is difficult to systematically review the full historical pattern of evidence attesting to convergent and discriminant validity for every test score. As such, we focus primarily on results from meta-analytic reviews.

Thousands of studies from around the world have provided evidence for Rorschach validity (e.g., for narrative summaries of specific variables see Bornstein & Masling, 2005; Exner & Erdberg, 2005; Viglione, 1999). Meyer and Archer (2001) summarized the available evidence from Rorschach meta-analyses, including four that examined the global validity of the test

and seven that examined the validity of specific scales in relation to particular criteria. The scales included CS and non-CS variables. For comparison, they also summarized the meta-analytic evidence available on the validity of the MMPI and IQ measures. Subsequently, Meyer (2004) compared the validity evidence for these psychological tests to meta-analytic findings for the medical assessments reported in Meyer et al. (2001).

Although the use of different types of research designs and validation tasks makes it challenging to compare findings across meta-analyses, the broad review of evidence indicated three primary conclusions. First, psychological and medical tests have varying degrees of validity, ranging from scores that are essentially unrelated to a particular criterion to scores that are strongly associated with relevant criteria. Second, it was difficult to distinguish between medical tests and psychological tests in terms of their average validity; both types of tests produced a wide range of effect sizes and had similar averages. Third, test validity is conditional and dependent on the criteria used to evaluate the instrument. For a given scale, validity is greater against some criteria and weaker against others.

Within these findings, validity for the Rorschach was much the same as it was for other instruments; effect sizes varied depending on the variables considered but, on average, validity was similar to other instruments. Thus, Meyer and Archer (2001) concluded that the systematically collected data showed the Rorschach produced good validity coefficients that were on par with other tests:

> Across journal outlets, decades of research, aggregation procedures, predictor scales, criterion measures, and types of participants, reasonable hypotheses for the vast array of Rorschach … scales that have been empirically tested produce convincing evidence for their construct validity (Meyer & Archer, 2001, p. 491).

Atkinson, Quarrington, Alp, and Cyr (1986) conducted one of the earliest meta-analytic reviews of the Rorschach and found good evidence for its validity. They noted that the test is regularly criticized and challenged despite the evidence attesting to its validity. To understand why, they suggested that "deprecation of the Rorschach is a sociocultural, rather than scientific, phenomenon" (p. 244). Meyer and Archer (2001) reached a similar conclusion about the evidence base and concluded that a dispassionate review of the evidence would not warrant singling out the Rorschach for particular criticism. However, they also noted that the same evidence would not warrant singling out the Rorschach for particular praise. Its broadband validity appears both as good as and also as limited as that for other psychological tests.

Robert Rosenthal, a widely recognized and highly regarded expert in meta-analysis, was commissioned to conduct a comparative analysis of Rorschach

and MMPI validity for a Special Issue of the journal *Psychological Assessment*. He and his coworkers (Hiller, Rosenthal, Bornstein, Berry, & Brunell-Neuleib, 1999; Rosenthal, Hiller, Bornstein, Berry, & Brunell-Neuleib, 2001) found that on average the Rorschach and MMPI were equally valid. However, they also identified moderators to validity for each instrument. Moderators are factors that influence the size of the validity coefficients observed across studies. The Rorschach demonstrated greater validity against criteria that they classified as objective, while the MMPI demonstrated greater validity against criteria consisting of other self-report scales or psychiatric diagnoses.[5] The criteria they considered objective encompassed a range of variables that were largely behavioral events, medical conditions, behavioral interactions with the environment, or classifications that required minimal observer judgment, such as dropping out of treatment, history of abuse, number of driving accidents, history of criminal offenses, having a medical disorder, cognitive test performance, performance on a behavioral test of ability to delay gratification, or response to medication. Viglione (1999) conducted a systematic descriptive review of the Rorschach literature and similarly concluded that the Rorschach was validly associated with behavioral events or life outcomes involving person-environment interactions that emerge over time. In general, these findings are consistent with the types of spontaneous behavioral trends and longitudinally determined life outcomes that McClelland, Koestner, and Weinberger (1989) showed were best predicted by tests measuring implicit characteristics, as opposed to the conscious and deliberately chosen near-term actions that were best predicted by explicit self-report tests (also see Bornstein, 1998).

In the most recent Rorschach meta-analysis, which was not considered in the previous reviews, Grønnerød (2004) systematically summarized the literature examining the extent to which Rorschach variables could measure personality change as a function of psychological treatment. The Rorschach produced a level of validity that was equivalent to alternative instruments based on self-report or clinician ratings. Grønnerød also examined moderators to validity and, consistent with expectations from the psychotherapy literature, found that Rorschach scores changed more with longer treatment, suggesting that more therapy produced more healthy change in personality. Grønnerød also noted that effect sizes were smaller when coders clearly did not know whether a protocol was obtained before or after treatment but larger in studies that clearly described scoring reliability procedures and obtained good reliability results using conservative statistics.

Overall, the meta-analytic evidence supports the general validity of the Rorschach. Globally, the test appears to function as well as other assessment instruments. To date, only a few meta-analyses have systematically examined the validity literature for specific scales in relation to particular criteria. The

evidence has been positive and supportive for the ROD, the Rorschach Prognostic Rating Scale (RPRS), and the precursor to the PTI, the Schizophrenia Index (SCZI), though it has not been supportive of the CS Depression Index (DEPI) when used as a diagnostic indicator. As is true for other commonly used tests, such as the MMPI-2, Personality Assessment Inventory (PAI; Morey, 1991), Millon Clinical Multiaxial Inventory (MCMI-III; Millon, 1994), or Wechsler scales (e.g., Wechsler, 1997), additional focused meta-analytic reviews that systematically catalog the validity evidence of particular Rorschach variables relative to specific types of criteria will continue to refine and enhance clinical practice.

Utility

In general, the utility of an assessment instrument refers to the practical value of the information it provides relative to its costs. The Rorschach takes time to administer, score, and interpret. To make up for these costs, the Rorschach needs to provide useful information that cannot be obtained from tests, interviews, or observations that are readily available and less time consuming. One way to evaluate this issue in research is through incremental validity analyses (see Hunsley & Meyer, 2003), where the Rorschach and a less time intensive source of information are compared statistically. To demonstrate incremental validity, the Rorschach would need to predict the criterion over and above what could be predicted by the simpler method. Such a finding demonstrates statistically that the Rorschach provides unique information.

Although utility cannot be equated with statistical evidence of incremental validity, the latter is one commonly obtained form of evidence that can attest to utility. Utility also can be demonstrated by predicting important real-world behaviors, life outcomes, and the kind of ecologically valid criteria that are important in the context of applied practice with the test. Research reviews and meta-analyses show that the Rorschach possesses utility in all of these forms, such that Rorschach variables predict clinically relevant behaviors and outcomes and have demonstrated incremental validity over other tests, demographic data, and other types of information (Bornstein & Masling, 2005; Exner & Erdberg, 2005; Hiller et al., 1999; Meyer, 2000a; Meyer & Archer, 2001; Viglione, 1999; Viglione & Hilsenroth, 2001; Weiner, 2001).

We do not have the space to review more than a sampling of utility findings. With respect to incremental validity, recent studies published in the United States and Europe show the Rorschach yields important information that is not attainable through simpler, less time consuming methods. The criteria include predicting future success in Norwegian naval special forces training (Hartmann et al., 2003), future delinquency in Swedish adolescents and adults based on clinician ratings of ego strength from childhood Rorschach protocols (Janson & Stattin, 2003), future psychiatric relapse among previously

hospitalized United States children (Stokes et al., 2003), future improvement across a range of interventions in United States adults (Meyer, 2000a; Meyer & Handler, 1997), future benefit from antidepressant medication in adult United States inpatients (Perry & Viglione, 1991), previous glucose stability levels in diabetic French children (Sultan et al., 2002), and future emergency medical transfers and drug overdoses in United States inpatients during a 60-day period after testing (Fowler, Piers, Hilsenroth, Holdwick, & Padawar, 2001). In these studies, the Rorschach demonstrated incremental validity over various alternative data sources, including self-report scales, collateral reports, *DSM* diagnoses, and intelligence tests.

Studies have repeatedly shown that Rorschach and self-report scales have minimal correlations even when they purportedly measure similar constructs (e.g., Bornstein, 2002; Krishnamurthy, Archer, & House, 1996; Meyer & Archer, 2001; Viglione, 1996). Although this lack of association was unexpected, it suggests that the Rorschach should display incremental validity over self-report scales. If both types of measures are related to a criterion but not to each other, each should maintain a unique association to the criterion and thus provide incremental validity over the other. At this point, more research has documented the limited associations between these two data sources than their combined value.

There are exceptions, however. For instance, studies have shown how it is the combined interaction of Rorschach-assessed and self-reported dependency that affords the optimal prediction of certain kinds of dependent behavior (Bornstein, 1998). In addition, the CS scales of psychotic symptoms (i.e., PTI or SCZI) have shown incremental validity over MMPI-2 scales of psychotic symptoms when predicting psychotic disorders (e.g., Dao, Prevatt, & Horne, in press; Meyer, 2000b; Ritsher, 2004). Rubin and Arceneaux (2001) recently illustrated this phenomenon with a case study.

A recent series of studies examining obese patients in Sweden demonstrated the utility of the Rorschach by predicting practical behavioral and life outcome criteria. Rorschach scores predicted the rate of consumption during an experimental meal, atypical acceleration in consumption during that meal, eventual weight loss in an obesity treatment program, and a positive response to weight loss medication (Elfhag, Barkeling, Carlsson, Lindgren, & Rössner, 2004; Elfhag, Barkeling, Carlsson, & Rössner, 2003; Elfhag, Carlsson, & Rössner, 2003; Elfhag, Rössner, Carlsson, & Barkeling, 2003; Elfhag, Rössner, Lingren, Andersson, & Carlsson, 2004).

Two other recent Swedish studies examined the Rorschach in relation to psychotherapy considerations. Bihlar and Carlsson (2001) documented how particular CS scores obtained before treatment predicted whether therapists would have to alter their initial plans for treatment over time, suggesting that the Rorschach scores identified characteristics that were not obvious from

interview and history information. Nygren (2004), using a selected set of hypothesized variables, found CS scores (a) differentiated patients who were selected versus not selected for intensive, long-term psychoanalytic therapy, and (b) were associated with clinician ratings of ego strength and capacity to engage in dynamic therapy.

Lundbäck et al. (2006) studied Swedish patients who had recently attempted suicide. They examined cerebrospinal fluid (CSF) concentrations of 5-hydroxyindoleacetic acid (5-HIAA), a serotonin metabolite, because previous research indicated low CSF 5-HIAA was associated with more violent and severe suicide attempts. As expected, the S-CON was negatively correlated with 5-HIAA levels (r_S = –.39). Post hoc analyses showed that responses in which shading gives rise to depth or dimensionality (vista) and the extent to which the form of objects perceived is secondary to their color (color dominance index; CF + C > FC) were the strongest individual predictors among the S-CON variables. In this study, 5-HIAA was unrelated to scores on the DEPI (r_S = –.21) and the Coping Deficit Index (CDI; r_S = .26). These results echo Fowler et al.'s (2001) United States findings, where the S-CON predicted subsequent suicidal behavior but the DEPI and CDI did not. Both sets of results provide evidence for both the convergent and discriminant validity of the S-CON.

As a final example, many studies have examined the ROD as an index of dependency. These have been systematically reviewed and meta-analyzed (Bornstein, 1996, 1999), with results showing that ROD scores validly predict help-seeking behavior, conformity, compliance, suggestibility, and interpersonal yielding in laboratory and clinical settings. Results also show the ROD has discriminant validity by being unrelated or minimally related to scales of alternative constructs like social desirability, IQ, and locus of control.

Our brief summary of recent studies addressing utility is limited in several ways. Although the authors for all of these studies carefully articulated hypothesized associations, some of the samples were small and the findings need to be replicated. There also were negative findings where the results did not support the hypothesized variables. For instance, Elfhag, Rössner et al. (2004) did not find support for the ROD in relation to eating behavior and Nygren (2004) did not find support for several anticipated variables as predictors of who would be selected for intensive psychotherapy (e.g., inanimate movement, distorted or arbitrary form quality, dimensionality based on form).

Nonetheless, based largely on the kinds of findings reviewed in this section, the Board of Trustees of the Society for Personality Assessment (2005) synthesized the available evidence and issued an official statement on the scientific foundation for using the Rorschach in clinical and forensic practice. They concluded "the Rorschach possesses reliability and validity similar to

Quick Reference

- The Rorschach can evaluate personality and problem solving in psychiatric, medical, forensic, and nonclinical settings.
- It is used with children, adolescents, and adults in any language or culture.
- The task is individually administered in a collaborative two-step process that elicits responses with the prompt, "What might this be?", and then clarifies the what, where, and why of each percept.
- Responses are recorded verbatim. The CS requires a minimum of 14; data and cost benefit considerations support prompting for at least two per card but obtaining no more than four.
- Proper administration, scoring, and interpretation require considerable training.
- Computer-assisted scoring is recommended and likely will become increasingly important.

that of other generally accepted personality assessment instruments and its responsible use in personality assessment is appropriate and justified" (p. 219).

Administration and Scoring

The Rorschach is used across a wide range of settings where questions of personality and problem solving are relevant, including inpatient and outpatient psychiatric settings, inpatient and outpatient medical settings, and forensic contexts. It can also be used to assess normal range personality functioning and to assist generally healthy people with goals for professional development or life enhancement. Because reading skills are not required, the Rorschach can be used as readily with children and adolescents as with adults, and as readily with people from the United States as with people from other countries around the world. Indeed, the International Society for the Rorschach boasts 20 member countries and more than 3,000 individual members from the African, Asian, European, North American, and South American continents.[6]

The CS provides guidelines for standardized administration and scoring, as well as reference data for children (in 1-year age increments from 5 to 16), adults (age 19 to 86), and several patient groups (see Exner, 2001, 2003; Exner & Erdberg, 2005). Practitioner surveys indicate that the CS takes about 45 minutes to administer and about 40 minutes to score (Camara et al., 2000).

Administration

The Rorschach is typically administered in the context of other assessment measures and the adequacy of any personality assessment depends on the

quality of the collaborative working relationship established between the examiner and client (see Fischer and Finn, chapter 10, this volume). Rorschach testing is not different and should not be attempted "cold" without first establishing decent rapport. Administration requires three tools: the inkblot stimuli, recording utensils (either notepaper with a pen or pencil or a laptop computer), and a location sheet that provides miniature inkblot images for recording where the key features of each response are located. Standardized CS administration takes place with the examiner seated next to the client to minimize visual cues from the examiner and to help him or her see what the client perceives, with the location sheet out of sight, and the inkblots face down on a table. The task is generally introduced as "the inkblot test" and because many people have heard of it the examiner typically asks the client what he or she knows about the test and if it was ever taken before. If the client has questions about the test or why it is being used, the examiner responds in a straightforward manner (e.g., "It's a test that provides some information about personality characteristics." or "No, there are no right or wrong answers.").

The administration itself is a two phase process consisting of the Response and Inquiry phases. In the Response phase, the client is sequentially handed each inkblot in order and at the outset is asked the standardized question, "What might this be?" The examiner numbers each response and records it verbatim, along with all additional commentary by the client. Once the Response phase is complete for all ten cards, the examiner introduces the Inquiry phase by explaining to the client that they will go through the responses a second time to ensure that the examiner sees each response in the same way that the client perceived it. The goal of this stage is not to elicit new information but to gather sufficient information to accurately score each response. The examiner primarily wants to know three things: what is being perceived (i.e., the content), where it is in the inkblot (i.e., the location), and how particular inkblot features contribute to or help determine the response (i.e., the so-called determinants of the response). The Inquiry begins with the examiner explaining that he or she wishes to briefly go through each response again to "see the things you saw and make sure I see them like you do." The examiner elaborates by saying, "I want you to show me where it is in the blot and then tell me what there is there that makes it look like that to you so I can see it just like you did." The somewhat awkwardly worded instructions to "tell me what there is there that makes it look like that" emphasize how the goal is not just to know what objects are seen where but also what aspects of the inkblot contribute to the perception. The examiner initiates the inquiry for each response by reading the verbatim portion from the Response phase and again records verbatim the further elaborations and examiner questions that emerge during the Inquiry phase. As the Inquiry

proceeds, the examiner completes the location sheet by roughly outlining the location of each numbered response and identifying its key features in sufficient detail so that another examiner will readily recognize the correct response location.

The first two inquiry goals (content and location, or what and where) are often obvious from the Response phase and may not need further clarification during the Inquiry. If they do, it is typically accomplished easily. The last goal (determinants or how inkblot features contribute to the percept) can be more complex, as clients often use indirect key words or phrases that suggest but do not confirm certain determinant scores. In the CS, determinant scores are related to the perception of movement (coded as human [M], animal [FM], or inanimate [m]), symmetry [reflection images, Fr or rF or paired objects, 2), shading (diffuse [*Y*] or involving a tactile impression [T]), color (chromatic [C] or achromatic [C']), and depth (based on shading [V] or on form [FD]). Determining whether movement and symmetry are present is typically straightforward and most often these features are coded without the examiner asking any additional questions during Inquiry. However, clients may not so clearly describe whether the shading, color, or depth contributed to their perception.

As such, to obtain the information that will allow for accurate scoring, the examiner must be alert to key words or phrases in the response suggesting these features and then generate a query to clarify the ambiguity. For instance, "a pretty flower" suggests that color may be an important determinant of the response; "trees on the horizon" suggests that depth may be important in forming the response; "it looks like a soft and furry rug" or "it's a wispy rain cloud" suggests that shading features may be important for the response. In each of these examples, the proper coding is uncertain, so the examiner has to formulate a question that will efficiently clarify how to code. What constitutes an effective and efficient question will depend on the context, including the quality of the relationship between the examiner and client and the kinds of Inquiry questions that already have been asked. At times, an efficient question may be quite general (e.g., "I'm not sure I see that like you; can you help?"), though more often the examiner would strive to ask a question that is focused directly on the key word or phrase (e.g., "You said it looks pretty?"; "On the horizon? I'm not sure what makes it look like that."; "What about the inkblot makes it look soft and furry?"), rather than being nonspecific (e.g., "Can you say more?" or "Help me see it like you"), tangential (e.g., "I'm not sure I see the flower" or "Where is the flower?"), or "double-barreled" and referring to multiple response elements (e.g., "Help me see the pretty flower," which would allow the client to address location or form features without necessarily addressing the prettiness that suggested color may be involved).

Standard CS administration requires a client to give at least 14 responses to the 10 inkblot stimuli and, although there are procedures in place to limit excessive responding, there is not a fixed limit to the upper end of the range. CS normative data indicate that an average protocol contain 22 or 23 responses, with 80% in the range from 18 to 27 responses. Because the CS norms are most applicable to protocols with 18 to 27 responses, it is desirable for all protocols to be in this range. However, existing administration guidelines (Exner, 2003) often produce protocols that fall outside of this range in clinical settings. Recent evidence (Dean, Viglione, Perry, & Meyer, in press; Sultan, 2006; Sultan et al., 2006) shows that the number of responses in a protocol moderates the test-retest stability and validity of scores, and that both are maximized when R is in the optimal range. Consequently, we have recommended simplified administration guidelines to maximize the prospect that examiners will obtain records of an optimal length (see Dean et al., in press). Specifically, this R-optimized administration uses a "prompt for two, pull after four" guideline. To ensure an adequate minimum, if only a single response is offered to any card, examiners should prompt for a second. To ensure the maximum number of responses is not excessive, examiners would remove any card after four responses. In preliminary work, when the impact of these revised administration guidelines was modeled on normative reference data, the score means were essentially unchanged but their variability decreased, suggesting a potentially better ability to discriminate typical from problematic functioning.

These modified guidelines are consistent with the evidence and also with cost-benefit principles. Short protocols tend to provide insufficient information and they lead to false negative errors of inference (i.e., incorrectly concluding that the client does not possess a characteristic). Lengthy protocols tend to provide unnecessarily redundant information and they lead to false positive errors of inference (i.e., incorrectly concluding that the client does possess a characteristic; one which is often unhealthy or pathological). In addition, both short and long protocols can be time consuming and frustrating for examiners and their clients. Under current CS guidelines examiners must administer the test a second time starting from scratch when less than 14 responses are obtained. This effectively doubles the testing time and often leaves clients confused about whether they should repeat initially offered responses. At the other end of the spectrum, lengthy protocols of 40 or more responses are time consuming to administer and score, and their complexity is often draining or exhausting for both the examiner and client.

Scoring

To score the Rorschach, codes are typically applied to each response and then aggregated across all responses. In the CS the codes assigned to each response

form what is known as the Sequence of Scores and the tally of codes across all responses is known as the Structural Summary. The scoring process can be fairly simple for single construct scoring systems, like the ROD, or fairly complex for multidimensional scoring systems, like the CS. However, scoring according to any system requires the same ingredients: a clearly articulated set of scoring guidelines, an understanding of those guidelines by the coder, and the coder's repeated practice of scoring against gold standard example material until proficiency is obtained. For a multidimensional system like the CS, fairly substantial training is required for proficiency. Table 8.1 provides a brief list of the standard CS codes that can be assigned to each response to generate the Sequence of Scores. These scores are then summed across responses and form the basis for about 70 ratios, percentages, and derived scores that are given interpretive emphasis on the Structural Summary. Because of the complexity of this material, we do not provide a detailed description. However, a full guide to interpretation can be found in standard interpretive texts (Exner, 2003; Exner & Erdberg, 2005; Weiner, 2003). These sources make it clear that formal coding is only part of the data that contributes to an interpretation. There are behaviors expressed during the testing, themes associated with response imagery, and perceptual or content based idiosyncrasies that are not captured by the formal scores but that may nonetheless be very important for helping to develop an idiographic and unique understanding of the client (e.g., Peebles-Kleiger, 2002).

The requirements for competent administration and interpretation are similar to the requirements for coding. In order to perform an adequate administration the examiner must first understand scoring in order to formulate suitable Inquiry questions. Like with scoring, developing proficient administration skills requires practice and accurate feedback about errors or problems. The latter can be accomplished most adequately when a thoroughly trained supervisor is physically present to observe and correct the student's practice administrations as they are occurring, though supervisory feedback on videotaped administrations also can be quite helpful. The least optimal training occurs when supervision feedback is only provided on hand written or typed protocols, as many nuances of nonverbal interaction are not captured by this written record and it is not possible for the supervisor to see how adequately the written record captured what actually transpired during the administration.

Interpretation

Not surprisingly, Rorschach interpretation is the most complex or difficult activity, as proficiency requires knowledge and skills in multiple areas. These include:

Table 8.1 A Brief Summary of Rorschach Comprehensive System Scores

Location and space	The client either makes use of the *whole* inkblot (W), one or more of its *commonly perceived detail* (D) locations, or one or more of its *small or rarely used detail* (Dd) locations. The background white *space* (S) can also be incorporated with each location (i.e., WS, DS, or DdS).
Developmental quality	The object(s) perceived either have definite or *ordinary* form demands (o) or they are characteristically formless or *vague* (v). When more than one object is identified they also are designated as either being *synthesized* in a meaningful interaction (o becomes +; v becomes v/+) or not.
Determinants	• *Movement* is scored when an object is perceived as being in motion or in a state of tension and it is designated separately for human activity (M), species appropriate animal activity (FM), or inanimate motion (m). Each type of movement is further designated as *active* (a) or *passive* (p). • *Color* scores can be of two types. Use of *chromatic color* is scored when the red or pastel colors are important to a response. Like all the remaining determinants, scores are differentiated by the extent to which form is also an important feature to the response, such that form can be primary and color secondary (FC), color can be primary and form secondary (CF), or form can be nonexistent (C). Use of *achromatic color* (FC', C'F, C') is scored when the white, black, or gray colors are important to a response. • *Shading* is scored in three ways. *Diffuse shading* (FY, YF, Y) is scored when the light and dark gradations of ink contribute to a response. *Texture from shading* (FT, TF, T) is scored when the light and dark gradations of ink give rise to a tactile quality, such as soft, furry, wet, or cold. *Vista from shading* (FV, VF, V) is coded when the light and dark gradations of ink give rise to a perception of depth or dimensionality. • *Form Dimensional* scores (FD) refer to instances when just the outline or form of an object generates a perception of depth or dimensionality. By definition form dominates this kind of response, so form is never scored as secondary or not present. • *Reflections* (Fr, rF) are scored when one side of the inkblot is a reflected or mirror image of the other. Form is considered inherent in such a response, so it is never coded as absent. • *Pure Form* (F) responses are assigned when it is only the shape or outline of an object that is salient. It is also a default score; it should be assigned when no other determinants are present and not assigned when other determinants are present. • *Blends* are instances when more than one determinant is present in a response; each is separated by a period. For instance, the score M^a.FC.C'F indicates the response contains active human movement, form dominated chromatic color, and form secondary achromatic color.

(*continued*)

Table 8.1 Continued

Form quality and popular responses	These scores characterize whether it is conventional to see an object in a particular location on a given card. Responses with at least some form are classified as *ordinary* (o; or + if thoroughly described) if they are commonly seen, *unusual* (u) if they are infrequent but consistent with the blot contours, and *minus* (–) if they are arbitrary, distorted, or impose nonexistent lines to define the object. To assign these codes the examiner consults an extensive table derived from more than 200,000 responses from 9,500 protocols. These tables document percepts perceived in W, D, or Dd locations to each card. In addition to the codes noted above, objects that were seen in at least one third of the 9,500 protocols are separately coded as *Popular* (P).
Pairs	A *pair* (2) is coded when the same object is identified on each side of the blot. This is a symmetry based score, like the reflection response.
Contents	Each object perceived is classified into a content based category. • There are four types of human or animal objects that are differentiated on two dimensions: whole versus partial and realistic versus fictional or mythological. The *human* codes are H versus Hd, for realistic whole objects versus realistic partial objects, and (H) versus (Hd), for fictional whole objects versus fictional partial objects. The *animal* codes are A versus Ad and (A) versus (Ad), respectively. In addition, *human experiences* (Hx) are coded when human emotions or sensory experiences are described. • Another class of content addresses body related imagery, including internal *anatomy* (An), *X-ray* or MRI-type images (Xy), *blood* (Bl), and *sexual organs or activity* (Sx). • A number of content codes relate to the physical environment, including *botany* (Bt), *landscape* (Ls), *nature* (Na), *clouds* (Cl), maps and *geography* (Ge), *fire* (Fi), and *explosions* (Ex); or to human creations, including *household* objects (Hh), products of *science* (Sc), *art* objects (Art), or cultural/historical images (Ay for *anthropology*). There is also a category for *food* items (Fd) and for percepts that are unique to the client or not otherwise classifiable (Id for *idiographic*)
Organizational activity	*Organizational Activity*, or Z scores, are coded for their *frequency* (Zf) and for the *degree of synthesis* evident in the response (Z-value or ZSum). The degree of synthesis is determined separately for each blot as a function of whether the response uses the whole inkblot (ZW), describes meaningful relationships between adjacent (ZA) or distant (ZD) objects, or integrates white space (S) with the rest of the blot (ZS).

Cognitive special scores	Six codes index disrupted or illogical thought processes. These include use of mistaken or inappropriate words (DV for *Deviant Verbalization*), circumstantial responses or use of inappropriate phrases (DR for *Deviant Responses*), describing one object with implausible or impossible attributes (INCOM for *Incongruous Combination*), describing two objects in an implausible or impossible relationship (FABCOM for *Fabulized Combination*), seeing two objects superimposed on each other and merged into a single percept (CONTAM for *Contamination*), and showing highly strained or overly concrete reasoning (ALOG for *autistic logic*).
Other special scores	The remaining codes identify a mix of notable features in a response. • Several of the codes are representational scores related to thematically defined images, including *aggressive* interactions (AG), *cooperative* interactions (COP), and *morbid* (MOR) perceptions where objects are broken, damaged, dead, spoiled, or imbued with dysphoric affect. • Other codes quantify instances when percepts are fixed, rigid, or *perseverative* (PSV); deal with symbolic, intellectualized, or *abstract* content (AB); imbue cards with color even though none is present (CP for *color projection*); or justify perceptions based on authority derived from *personal* knowledge (PER). • Two final codes provide an indication of object relations, though they are not independently assessed. Rather the *Good* and *Poor Human Representation* variables (GHR and PHR) summarize other scored information in the protocol, drawing upon determinants, content, form quality, cognitive special scores, and the COP, AG, and MOR special scores.

- an understanding of interpretive postulates associated with the various scores obtained from the test;
- an understanding of the kind of information the Rorschach can and cannot provide (i.e., its locus of effectiveness);
- knowledge of the psychometric research literature on the types of systematic bias that can affect Rorschach scores;
- knowledge of the psychometric research literature on the reliability and validity of the test scores to be interpreted;
- a thorough understanding of personality and psychopathology, particularly of the condition(s) being assessed;
- recognition of the kind of judgment errors that can adversely influence clinical inferences;

- the capacity for disciplined reasoning to rule in and rule out inferences; and
- the ability to integrate Rorschach-based inferences with inferences obtained from other tests, from observed behavior, and from history as reported by the client and other sources of collateral information.

Of course, to adequately perform the last step of integration, the examiner must also have parallel forms of knowledge about the other tests and sources of information that are contributing inferences. That is, for each non-Rorschach data source, the clinician must understand the interpretive postulates associated with the observation, understand the kind of information that the data source can and cannot provide, know what forms of systematic bias influence the data source, and know the reliability and validity evidence for the alternative data source. To become proficient with the idiographic task of correctly interpreting a complex array of personality test results, including Rorschach scores, requires considerable closely supervised clinical experience with a well-trained individual.

Computerization

Although computerized administration has been used in Rorschach research, standard CS test administration does not lend itself to automated, computer-adapted administration or to computer automated scoring. However, computer-*assisted* scoring and interpretation for the CS is quite common, with the two primary software programs being the Rorschach Interpretive Assistance Program (RIAP), which is now in its 5th edition and authored by John Exner and Irving Weiner, and ROR-SCAN, which is now in its 6th edition and authored by Philip Caracena. Reviews of each program can be found in Acklin (2000; for the 4th edition of RIAP) and Smith and Hilsenroth (2003; for the 6th edition of ROR-SCAN).

Because the CS Structural Summary tabulates many different scores and then generates numerous other ratios or derived scores, we strongly recommend computer-assisted scoring to minimize the prospect of computational errors. For computer-assisted scoring, the examiner manually assigns codes to each response on the sequence of scores, but allows the computer algorithms to generate the final Structural Summary. Doing so has a number of benefits. First, it allocates the clinician's time and expertise where it is required, which is with judging what codes should be assigned to each response, and it leaves the mundane (but error prone) mathematical operations to a machine that is perfectly suited to these clerical tasks. Second, computer-assisted scoring would allow all users to obtain CS-based variables like the Ego Impairment Index (EII-2; Perry & Viglione, 1991; Viglione, Perry, & Meyer, 2003) that are too complex for hand scoring.

Third, although commercial programs currently do not do so, they can be programmed to generate complex scores that will facilitate clinical interpretation. For instance, programs could provide scores that are adjusted for the overall complexity of the protocol (i.e., first factor variance) or they could provide congruence coefficients that empirically show how well a client's pattern of scores fit with the average scores from a criterion group (e.g., patients diagnosed with schizophrenia or borderline personality disorder). Future computerization also could enable users to maximize information at the level of individual responses or cards. Currently, scores are summarized at the protocol level, aggregating equally across all responses and cards. However, because of card pull, responses that occur to specific cards and location areas may have differential validity that should be taken into account during interpretation.

With these potentials in mind, reliability, validity, and utility can be maximized by more fully harnessing computer resources. At the same time, users should be cautious when considering computer generated interpretative reports. These can certainly be helpful but their ready accessibility can tempt less experienced or proficient clinicians to cut-and-paste material into a final report without sufficiently considering idiographic contextual issues or the nature and limitations of Rorschach-based scores.

Applications and Limitations

As noted above, the Rorschach can be used in a wide range of settings, including inpatient and outpatient psychiatric and medical settings, in forensic contexts, and in nonclinical situations for professional development, personal

Just the Facts

Ages:	5 or 6 to elderly
Purpose:	To assess personality and problem solving characteristics using a sample of spontaneously generated behavior and imagery collected under standardized conditions.
Strengths:	Provides an in vivo demonstration of personal characteristics, many of which may reside outside of conscious awareness.
Limitations:	Many assessed characteristics are implicit and independent of self-reported characteristics, which make it risky to interpret test scores in isolation.
Time to Administer:	about 45 minutes
Time to Score:	about 40 minutes for the CS

enhancement, or counseling. With minimal extra-test modifications, it can also be used in the same form with children, adolescents, and adults, regardless of culture, language, or nationality.

Clinicians may choose to use the Rorschach for many different reasons. However, it is often selected precisely because it is an office based procedure that provides a unique source of information—one that differs considerably from the self-reported characteristics that form the basis for the many inventories or structured interviews[7] available for assessing personality (e.g., those described in other chapters of this text).

A number of authors have described important distinctions between self-report scales and Rorschach measures (Meyer, 1997; Meyer & Archer, 2001; Viglione & Rivera, 2003). Self-report measures require clients to determine the extent to which verbal statements, adjectives, or symptoms are characteristic of their personality. Although there is some variability from instrument to instrument, because of how the task is structured, the information obtained from a self-report measure is dependent upon the client's conscious understanding of himself or herself, ability to accurately characterize himself or herself relative to others when determining if a characteristic is or is not self-descriptive, and willingness to convey information in an accurate and forthright manner. Under optimal conditions, self-reported data is particularly adept at addressing and quantifying the presence and severity of specific, consciously recognized preferences, affective states, and symptoms.

In contrast, the Rorschach task requires clients to identify and articulate images in response to a set of complex and novel stimuli. Although subject to its own sources of bias and error, as a sample of actual behavior obtained under standardized conditions, the information obtained from the Rorschach does not depend on the client's consciously represented self-image or ability to accurately evaluate him or herself. Under optimal conditions then, this allows Rorschach data to provide information about problem solving styles and implicit or tacit personal qualities that may reside outside of consciousness, even though these characteristics may regularly guide and motivate behavior or provide the schematic templates that filter and interpret experiences.

One way to understand the distinction between these methods of assessment is to consider them in the context of assessing intelligence. It certainly can be informative to directly ask people how intelligent they are or how they compare to peers in their specific abilities, such as capacity to solve verbal problems, to identify visuospatial relationships, to quickly and easily process information, or to mentally transform and manipulate information in short-term memory stores. However, most people do not have a clear awareness or understanding of their cognitive abilities, are uncertain how they stack up against their peers, and/or are motivated to describe their abilities in an overly positive light (or overly negative light, depending on

the circumstances). Consequently, when it is important to have an accurate understanding of someone's actual intelligence, psychologists typically administer a standardized intelligence test that provides a behavioral sample and in vivo demonstration of problem solving, information processing, verbal ability, and so on. Not surprisingly, this performance based information is quite different than self-reported results. Depending on the ability construct and sample considered, research reveals the correlation between self-reported and performance based methods of assessing cognitive ability range from about $r = .00$ to $r = .30$ (Meyer et al., 2001; Paulhus, Lysy, & Yik, 1998).

Returning to personality assessment, self-reported information from a cooperative client can provide critical information about many clinical conditions, personal experiences, and normative characteristics. For example, when assessing depressive suicidality, self-report measures can quantify specific symptoms and warning signs, such as consciously experienced and persistent depressed mood, diminished interest or pleasure in almost all activities, excessive or inappropriate guilt, and deliberate suicidal ideation with intention and means. No matter how many responses are available for consideration, one simply is not able to assess these specific characteristics with the Rorschach. In contrast, however, the Rorschach can measure the extent to which experiences are filtered through a depressively biased schema, whether underlying affect is chaotic or modulated, and the extent to which implicit coping resources are disorganized and unavailable, all of which are personality features associated with variables on the CS S-CON. Although these characteristics are not readily assessed by self-report and although there is no correlation between the S-CON and self-rated depressive symptoms or suicidality (Meyer, 1997; Meyer et al., 2000), as noted above, research has consistently documented that the S-CON predicts self-harm behavior.

The issues are different for clinical conditions in the psychotic spectrum. Here, although self-reports can be useful to understand some specific symptoms (e.g., hearing voices, identifying whether seemingly nonsignificant events feel imbued with personal meaning, beliefs that one is being plotted against by others), many of the most relevant symptoms are based on observable behavior, including the accuracy or conventionality of one's perceptions, faulty and overly personalized or concrete logic, fluid and disorganized thinking, or a difficulty maintaining conceptual distinctions among events, experiences, and images of self and other. The latter are not readily assessed by direct questions or self-reported endorsement of specific characteristics. However, they often can be readily observed in, or distilled from, the in vivo sample of behavior obtained with the Rorschach. As a standardized behavioral task that requires visual processing, problem solving, and verbal expression, the Rorschach is adept at identifying atypical or distorted perceptions and disrupted thought processes.

There are a number of limitations associated with using the Rorschach in applied practice. For instance, it is time intensive to learn proper administration, scoring, and interpretation. This can be a particular limitation in increasingly crowded graduate curricula, where less-than-adequate time may be devoted to teaching students how to conduct idiographic and in-depth personality assessment and students may be inadequately prepared to use the instrument in a competent and useful manner. Another limitation is that even though the CS is the dominant system used in the United States and abroad, the validity evidence for some scales that are not included in the system (e.g., ROD, RPRS, or Mutuality of Autonomy Scale [MOA; Urist, 1977]) has eclipsed the evidence for some scales that are part of the system (e.g., Isolation Index, Obsessive Style Index, active to passive movement ratio, the PSV score).

Several limitations associated with scoring also can be noted. First, some of the CS scoring distinctions are of dubious value (e.g., the distinction between botany, landscape, and nature content categories; the household and science content categories; instances when different form quality codes are assigned to similarly shaped objects), particularly because they make the system more difficult to learn, consume teaching resources and scoring time, and contribute to unreliability.

Second, some CS scoring principles are not optimally refined to assess a targeted construct. For instance, the Isolation Index is thought to assess a sense of isolation or remoteness from others and it is formed by considering the number of responses containing content codes for botany, landscape, nature, clouds, or geography. However, each of these scores can co-occur with content codes for human or human-like objects, which would suggest an interest in others rather than a sense of isolation or remoteness from others. Thus, the overall Isolation Index can be elevated even when every response in a protocol contains perceptions of human characters.

Third, most CS scoring criteria are based on abstract principles that do not offer specific guidance for applying those principles to the inkblot stimuli that are most likely to elicit them. For instance, out of the 10,512 responses that make up the 450 protocols in the current CS normative sample (Exner & Erdberg, 2005), shading generated a sense of texture most often on Card VI (302 responses; 66% of all texture responses), followed by Card IV (102 responses; 22% of all texture responses), and then rarely on the remaining eight cards (all < 13 responses; < 3% of all texture responses). Given this, and assuming this patterning generalizes to other types of samples (which our data indicates it does), it would be desirable to have scoring guidelines that are specifically tailored to the types of responses that are typically found on Cards VI and IV.

It also would be desirable to have specific guidelines for instances when

abstract coding criteria are challenging to apply to commonly given responses. For instance, the D1 area on Card VII is very commonly described as a girl or woman's head. Typically, the object is also described as having her hair sticking up in the air and coders would benefit from specific guidelines for when inanimate movement should be coded in this common response (e.g., Viglione, 2002).

Finally, in many instances there is a degree of irreducible uncertainty associated with scoring because of the ambiguity that is inherent in a verbalized response. Much like a reversible figure or Necker cube, even after being adequately inquired, some responses can be interpreted in two notably different and mutually exclusive ways. This allows for reasonably trained people to disagree on what exactly was perceived and described by the client, and thus will lead reasonably trained people to disagree on scoring. At times, coders also can disagree on what is included in a response. For example, clients sometimes change their perception from the Response to the Inquiry phase, or examiners may be unsure when multiple objects are identified if they constitute one combined response or several distinct responses. Such ambiguities need to be addressed in the future to increase reliability in the test.

Despite these limitations, the Rorschach offers clinicians a rich sample of behavior on which to base carefully considered, disciplined, and synthesized

Important References

Exner (2003), Viglione (2002), Exner and Erdberg (2005), and Weiner (2003). Together these four resources provide the basic information needed to learn standard CS administration, scoring, and interpretation. Exner also provides an overview of evidence for each CS score, Viglione elaborates on and clarifies basic scoring principles, Exner and Erdberg review relevant research in the context of an interpretive guide that addresses particular referral questions, and Weiner complements the latter by providing an easy to read general interpretive guide.

Meyer (1999b) and Meyer (2001c). These citations reference a special series of eleven articles in the journal *Psychological Assessment*. The authors in the series participated in a critical, structured, sequential, evidence based debate that focused on the strengths and limitations of using the Rorschach for applied purposes. The debate took place over four iterations, with later articles building upon and reacting to those generated earlier. This series gives an overview of all the recent criticisms of the test.

Bornstein and Masling (2005). This text provides an overview of the evidence for seven approaches to scoring the Rorschach that are not part of the CS. Scores that are covered include the ROD for assessing dependency, as well as scales to measure thought disorder, psychological defenses, object relations, psychological boundaries, primary process thinking, and treatment prognosis.

Society for Personality Assessment (2005). Drawing on the recent literature, this document is an official statement by the Board of Trustees of the Society for Personality Assessment concerning the status of the Rorschach in clinical and forensic practice. Their primary conclusion was that the Rorschach produces reliability and validity that is similar to other personality tests, such that its responsible use in applied settings is justified.

inferences about personality. In the applied arena, the meta-analyses and individual studies reviewed above have shown it can predict important and clinically relevant behaviors, predict subsequent treatment outcome, identify qualities associated with good and poor treatment prognosis, quantify change in personality as a function of treatment, and assist in differential diagnosis, particularly for psychotic disorders.

Research Findings

In earlier sections we described the evidence base for the Rorschach in some detail. We documented how meta-analyses have shown its scores can be reliably assigned, are reasonably stable, and, when evaluated globally, are as valid as those obtained from other personality assessment instruments. We also documented how the Rorschach can validly assess a range of personal characteristics that have meaningful utility for applied clinical practice, including diagnosing psychotic difficulties, planning treatment, and monitoring the outcome of intervention. Here we focus on some of the relatively unique challenges that are associated with documenting the construct validity of its scores and validly interpreting them in clinical practice.

Foundation for Interpretive Postulates

Authors over the years have discussed challenges associated with validating Rorschach-derived scales (e.g., Bornstein, 2001; Meehl, 1959; Meyer, 1996; Weiner, 1977; Widiger & Schilling, 1980). One challenge arises because some scores do not have an obvious or self-evident meaning. In other words, the behavioral or experiential foundation for the response is not completely obvious. Examples of these scores include diffuse shading (*Y*), use of the white background (S), or the extent to which form features are primary versus secondary in determinants (e.g., FC vs. CF; see Table 8.1 for score descriptions). These are largely the scores we described above as being based on clinical observation. Historically, these response characteristics have been observed and studied in psychiatric settings with disturbed individuals where the base rates of serious symptoms and failures in adaptation are high. As a result, the standard interpretive algorithms (Exner, 2003) may be skewed or biased toward negative and pathological inferences rather than toward the positive or healthy inferences that may be relevant when such responses are present in nonpsychiatric settings.

Unique Assessment Methodology

Another challenge relates to the uniqueness of the method itself. Because of its uniqueness, the correlation between one Rorschach scale and another Rorschach scale is rarely put forward as evidence for validity. For instance,

both the MOA (Mutuality of Autonomy Scale) and the HRV (Human Representation Variable) assess the quality of object relations and theoretically should be related to each other. However, researchers have not tried to validate either scale by showing that they are correlated. Although this type of research is rare with the Rorschach, it is a pervasive practice with other assessment methods, where, for example, the correlation between two self-report scales or two performance tasks of cognitive ability are regularly put forward as validity evidence.

Instances when two scales from the same assessment method (e.g., two Rorschach scales or two self-report scales) are correlated with each other are known as monomethod validity coefficients (Campbell & Fiske, 1959) and they are contrasted with the heteromethod validity coefficients obtained when scales from two different assessment methods are correlated (e.g., when a Rorschach scale is correlated with ratings of observed behavior). It has been well-documented for the past half-century that monomethod validity coefficients are substantially larger than heteromethod coefficients. This is because method-specific sources of systematic error inflate the monomethod coefficients (Campbell & Fiske, 1959; Meyer, 2002b).

For instance, consider self-report questionnaires to assess depression. To document convergent validity, depression scales on the MMPI-2 and PAI have been correlated with each other and scales on both instruments have been correlated with the Beck Depression Inventory (BDI; Beck, Steer, & Brown, 1996). Several factors conspire to artificially inflate these correlations, and these factors are forms of systematic error. First, and most importantly, there is an issue of what is known as criterion contamination in these studies. Standard psychometric texts (e.g., Anastasi & Urbina, 1997) define criterion contamination as instances in which knowledge of a predictor variable can potentially influence the criterion variable (e.g., IQ scores are to be validated by teacher ratings of intelligence but teachers see their students' scores before making their ratings). These texts also document how it is essential to avoid this problem in validity research to ensure validity coefficients are not falsely inflated. In the case of two self-report scales, not only can knowledge of what is reported on one scale influence what is reported on the other, but in fact the same person—the respondent—determines the scores that will be present on both the predictor scale and the criterion scale. This circularity where the same person determines the data on all measures is a serious methodological confound. Exacerbating the difficulty, people also strive for consistency when answering similar items on two different inventories. Thus people will strive to give consistent answers regarding sadness, tearfulness, or lack of energy on two different depression scales.

It is also the case that self-ratings on two measures of depression (or any other construct) are artificially equated by virtue of psychological defenses,

by genuine limitations in self-knowledge, by an inability to realistically appraise oneself relative to others, and by intentional or unintentional desires to create an overly positive or an overly negative impression. All of these processes artificially inflate convergent correlations because so many methodological confounds are intertwined (see Campbell & Fiske, 1959; McClelland, 1980).

Psychometrically, this kind of monomethod research produces results that are more like estimates of alternate forms reliability than of actual validity (Meyer, 2002b). Because monomethod coefficients are rarely presented as validity evidence for Rorschach scales, a casual or unsophisticated review of the research literature that fails to appreciate these issues can readily but erroneously lead one to believe that self-report scales produce higher validity coefficients than Rorschach scales.

The Rorschach method elicits a sample of problem-solving behavior in the verbal descriptions of what the blots might be, which is then coded by the examiner on a range of structural and thematic dimensions. Although this is a unique method for assessment, the Rorschach is like other assessment procedures in that its method variance is large relative to desired trait variance (e.g., Meyer et al., 2000). For the Rorschach, a primary source of method variance can be seen in the way scores on the test rise and fall in tandem with the number and complexity of the responses that a person gives. This can have a dramatic impact on many final scores, particularly for protocols that fall at either extreme of the simplicity-complexity dimension[8] (Viglione & Meyer, 2007). Validation research is needed to more fully understand this dimension of response complexity and its implications for personality, coping resources, and test-taking defensiveness. In addition, in many situations researchers should control for its impact when attempting to validate specific scales derived from the test.

Implications of Methodology for Interpretation and Research

Given the methodology of Rorschach assessment, there is no aspect of the data collection and scoring process that requires or even suggests that the behaviors coded from the task should quantify consciously represented or consciously experienced personal characteristics. These characteristics may be in consciousness; however, this is not required. Indeed, one of the most pervasive and consistent findings in the literature is that that Rorschach and self-report scales with similar names tend to be minimally correlated (e.g., Krishnamurthy et al., 1996; Meyer et al., 2000). Part of this may be due to the fact that the Rorschach task begins with visual perception. Compared to the solely verbal expression and processing required to complete a self-report inventory, the Rorschach response process likely involves somewhat different filters or censoring processes, as well as inadvertent or unself-con-

scious expressions of personal characteristics. In either case, the Rorschach's methodological uniqueness has implications for both research and clinical interpretation.

With respect to research, validation criteria have to be selected so they are consistent with the type of information the Rorschach can provide. This includes focusing on spontaneously chosen behaviors observed over time. One promising but untried approach is with experience sampling methodology, in which participants record over a period of days or weeks what activities and experiences are occurring at the moment when they are electronically prompted (e.g., McAdams & Constantian, 1983). This kind of methodology should be particularly well suited for some of the representational scores described earlier (e.g., MOR, COP). In addition, Rorschach researchers will need to begin taking fuller advantage of methodological procedures that are used in the social-cognitive literature for validating implicit measures of personality, mood, and attitudes, including experimental procedures that induce particular affective states or prime particular thematic material (see Bornstein, 2001; as well as Balcetis & Dunning, 2006; Long & Toppino, 2004; Payne, Cheng, Govorun, & Stewart, 2005).

Considering Rorschach data from a behavioral representation model adds another dimension to consider when evaluating the Rorschach's locus of effectiveness. When generalizing from test problem-solving behaviors to everyday life, we need to consider functional equivalence (Foster & Cone, 1995), or the extent to which behaviors in the microcosm of the Rorschach environment generalize to particular external environments. More specifically, this perspective should help researchers to conceptualize the discriminative stimuli, antecedents, consequences, and environmental conditions to which we should be able to most assuredly generalize Rorschach behaviors.

With respect to clinical interpretation, the Rorschach's methodological uniqueness has important implications for the extent to which clients are aware of Rorschach assessed characteristics. We bring this issue up in part because there are times when the language used in standard interpretive texts could be misunderstood. For example, an elevated number of diffuse shading responses are typically interpreted as being associated with feelings of helplessness or anxiety. But an elevated number of *Y* scores does not also imply these feelings are consciously recognized. The client who describes how the shading in the ink was influential in his perception may or may not also say he is anxious or feeling helpless. To confidently draw inferences about the conscious experience of anxiety or helplessness a clinician would have to consider the Rorschach data in light of other sources of information (e.g., self-reported, observer-rated, behavioral observation).

So, even though a Rorschach score may be associated with a conscious experience, that may not be the case, as people fail to recognize their internal

states and experiences for various reasons (e.g., because they lack intrapersonal sophistication and insight or because they have defenses that push these threatening feelings from awareness). The notion that clinicians should not infer that a score necessarily implies a conscious and self-reportable experience applies to a long list of constructs often considered in the course of CS interpretation (Exner, 2003), including affective distress, depression, sadness, stress, overloaded coping resources, inability to concentrate, needs for closeness, loneliness, introspectiveness, self-criticism, emotional deprivation, emotional confusion, interest in or discomfort with affective stimuli, oppositionality, hypervigilance, suicidality, passivity, dependence, inflated sense of personal worth, negative self-esteem, bodily concerns, pessimism, interest in others, or the expectation that relationships will be cooperative and/or aggressive. Even though validity data indicate Rorschach variables actively influence perception, behavior, and thought, research also indicates these experiences may not be consistently accessible in consciousness and available to self-report. Recognizing this constraint when interpreting data and writing test reports will help ensure inferences are consistent with the Rorschach's methodology and the evidence about its locus of effectiveness.

The Implications of Card Pull for Summary Scales

With respect to interpretation, we note another caution that can be overlooked when following the standard approach found in textbooks. An average protocol contains about 23 responses. However, each response is given to a specific card and uses one or more specific locations. Each location and card has unique stimulus properties that pull for certain kinds of perceptions, including content categories and determinant scores. Thus, even though summary scores are formed by aggregating codes across all responses, for many scores, only a portion of the responses would be relevant for a particular score (e.g., color responses are impossible to obtain on half the cards). Consequently, a summary score derived from a 23-response Rorschach is not equivalent to the kind of summary score that would be obtained from a 23-item scale on most other personality or cognitive ability tests. Because each Rorschach response is not like a test item that consistently evaluates the same underlying dimension, psychometrically most CS summary scores should be viewed as being derived from relatively brief scales (i.e., fewer than 20 relevant items; at times perhaps just several items), which results in many scores having a truncated distribution where most participants obtain scores of just 0, 1, or 2.

To illustrate this point, we mentioned earlier that the vast majority of texture scores occur to two of the inkblots (in the CS reference sample almost 90% of these scores occur on Cards VI and IV). Because most people generate two responses to each of these cards, for most people there is a reasonable

opportunity to observe a texture response just four times in a protocol. Thus, the stimulus features of the inkblots limit the opportunities to observe a score and result in a summary scale with a truncated range (e.g., 97% of the people in the CS reference sample have 0, 1, or 2 texture scores).

Such truncated scales are particularly sensitive to a form of random error that is not captured by scoring reliability coefficients. Rather, this type of error concerns the factors that interfere with the examiner's ability to transcribe and score what the client actually sees and tries to articulate. These factors include the client's choice of particular words to describe the percept, the examiner's attentiveness to key words or phrases, the sophistication of the examiner's inquiry questions and choice of particular inquiry words, the client's speech, which at times may be inaudible or too rapid for an accurate verbatim transcript, the examiner's misperception of what was said, and so on. These factors can negatively impact all Rorschach scores, but relatively speaking their impact will be more pronounced on those with a small range.

As a result, while keeping in mind the overall complexity of a protocol, we encourage clinicians to focus interpretation on global scores that either are assigned to every response and thus aggregate information across all responses (e.g., form quality, organizational activity, cognitive special scores) or incorporate multiple response features (e.g., the EII-2 or HRV, which combine information from determinants, form quality, contents, and special scores), because these tend to be the most reliably measured variables. In addition, clinicians should cautiously and conservatively interpret Rorschach summary scores with truncated distributions. This means that clinicians should mentally impose fairly wide confidence intervals around observed scores on the test. For instance, even though a client may have produced one texture response, there is enough potential random error in the administration, recording, and scoring process that the savvy clinician will keep in mind how the client's "true" score actually may be 0 or 2.

Cross Cultural Considerations

In this section we address both the cross-cultural applications of the test as well as normative issues more generally. As suggested by some of the data reviewed above, the Rorschach appears to be as valid when administered in other countries and with other languages as it is in the United States with English. In addition, considerable research shows that scoring can be done reliably on an international basis, with the scores that are more challenging to reliably code in the United States also being more challenging in other countries (Erdberg, 2005). Three fairly recent studies directly examined cross-cultural issues with the CS (Meyer, 2001a, 2002a; Presley, Smith, Hilsenroth,

& Exner, 2001). In addition, Allen and Dana (2004) provided a thorough review of existing evidence, as well as a detailed discussion of methodological issues associated with cross-cultural Rorschach research.

Presley et al. (2001) compared CS data from 44 African Americans (AA) to 44 European Americans (EA) roughly matched on demographic background using the old CS nonpatient reference sample norms. They examined 23 variables they thought might show differences, though found only 3 that differed statistically (the AA group used more white space, had higher SCZI scores, and had fewer COP scores). While preparing this chapter, we examined ethnic differences in the new CS reference sample of 450 adults (Exner & Erdberg, 2005). This sample contains data from 39 AAs and 374 EAs, with the remaining 37 participants having other ethnic heritages. We could not replicate the findings of Presley et al. Although there were small initial differences on the number of responses given by each group (AA $M = 21.4$, $SD = 3.5$; EA $M = 23.8$, $SD = 5.9$), once we controlled for overall protocol complexity, ethnicity was not associated with any of the 82 ratios, percentages, or derived variables on the Structural Summary (i.e., the variables found in the bottom half of the standard CS structural summary page). Across these 82 scores, ethnicity did not produce a point biserial correlation larger than |.09|.

Meyer (2002a) compared European Americans to a sample of African Americans and to a combined sample of ethnic minorities that also included Hispanic, Asian, and Native American individuals using a sample of 432 patients referred to a hospital based psychological assessment program. He found no substantive association between ethnicity and 188 Rorschach summary scores, particularly after controlling for Rorschach complexity and demographic factors (gender, education, marital status, and inpatient status). In addition, CS scores had the same factor structure across majority and minority groups and in 17 validation analyses there was no evidence to indicate the test was more valid for one group than the other.[9] These data clearly support using the CS across ethnic groups.

Meyer (2001a) contrasted Exner's (1993) original CS adult normative reference sample to a composite sample of 2,125 protocols taken from nine sets of adult CS reference data that were presented in an international symposium (Erdberg & Shaffer, 1999). Although the composite sample included 125 (5.8%) protocols collected by Shaffer et al. (1999) in the United States, the vast majority came from Argentina, Belgium, Denmark, Finland, Japan, Peru, Portugal, and Spain. Despite diversity in the composite sample due to selection procedures, examiner training, examination context, language, culture, and national boundaries, and despite the fact that the original CS norms had been collected 20–25 years earlier, relatively few differences were found between the two samples. Across 69 composite scores, the average difference was about four tenths of a standard deviation (i.e., equivalent to about

4 *T*-score points on the MMPI or 6 points on an IQ scale). Also, preliminary analyses using the initial participants in Exner's new normative sample indicated that it differed from the old reference data by about two tenths of a standard deviation, such that the international sample was more similar to the new norms. These data suggested that the CS norms were generally adequate even for international samples. However, there are caveats to this conclusion because, as we discuss next, there are issues associated with the application of the CS norms in the United States as well.

Wood, Nezworski, Garb, and Lilienfeld (2001a, 2001b) criticized the CS normative reference sample for being unrepresentative of the population and for causing healthy people to be considered pathological or impaired. The research that inspired their critique was the study conducted by Shaffer, et al. (1999), who used graduate students to collect a reference sample of 123 nonpatients from the Fresno, California area. For most scores, the values reported by Shaffer et al. were consistent with the CS normative reference group. However, there were also some surprising divergences. Most striking was the lack of complexity in the Shaffer et al. sample. Their participants gave fewer responses and more responses where no determinants were articulated. As a result, their protocols looked more simplistic or constricted relative to the CS reference sample (and relative to a number of other reference samples as well). Building on this research, Wood et al. (2001a) selected 14 scores to examine in a review of the literature. Depending on the score, they compared the CS reference values to values derived from between 8 and 19 comparison samples. They reported small to very large differences, all of which suggested the comparison samples had more difficulties or problems relative to the CS norms.

There were many problems with the samples Wood et al. included in their analyses, which is why Meyer (2001a) contrasted Exner's (2001) old adult normative sample to the composite international sample. As noted above, most scores in the international sample were similar to Exner's values. However, people in the composite international sample used more unusual location areas, incorporated more white space, had less healthy form quality scores, made less use of color, tended to see more partial rather than full human images, and showed a bit more disorganization in thinking.

To more fully understand these differences and to determine whether they may have resulted from changes in the population over time, Exner collected a new adult normative reference group from 1999 to 2006. Although he did not complete data collection before his death, Exner and Erdberg (2005) provide the reference data for 450 new participants. Relative to the old CS norms, the new reference sample also looks less healthy. People in the contemporary norms incorporated more white space into their responses, had less healthy form quality scores, made less use of color, tended to see more

partial rather than full human images, and showed a bit more disorganization in thinking.

As such, changes seen within the CS norms over time are very similar to the differences that had been found when comparing the original CS norms to the composite international sample. However, the new CS reference sample does not eliminate differences with the composite international sample. In particular, the current CS norms continue to show less use of unusual detail locations, better form quality, and more color responding than is seen in the reference samples collected by others.

To understand the factors that may account for this, we compared the quality of administration and scoring for protocols in Exner's (Exner & Erdberg, 2005) CS norms relative to Shaffer et al.'s (1999) sample from Fresno, CA (FCA; preliminary findings were reported in Meyer, Viglione, Erdberg, Exner, & Shaffer, 2004). Two sets of results are notable. First, the FCA protocols were less adequately administered and inquired, with more instances when examiners failed to follow up on key words or phrases. This is not surprising given that graduate student examiners collected all the protocols, though it does indicate that some of the seeming simplicity in the FCA records was an artifact of less thorough inquiry. Second, we found that many of the seeming differences between the FCA and CS samples were reduced or eliminated when 40 protocols from each sample were rescored by a third group of examiners. This indicates that the Shaffer et al. records and Exner protocols were coded according to somewhat different site-specific scoring conventions. In general, the new scoring split the difference between the CS and Shaffer et al. samples, making the CS protocols look a bit less healthy than before and making the Shaffer et al. protocols look a bit more healthy than before. There were two exceptions to this general trend. For complexity, the rescored protocols resembled the CS norms more than the FCA scores. In contrast, for form quality the rescored protocols resembled the FCA scores more than the CS norms. The overall findings suggest that site-specific administration and coding practices may contribute in important and previously unappreciated ways to some of the seeming differences across normative approximation samples (also see Lis, Parolin, Calvo, Zennaro, & Meyer, in press).

Although this research has been conducted with adults, the issues appear to be similar with children. For instance, Hamel, Shaffer, and Erdberg (2000) provided reference data on 100 children aged 6 to 12. Although rated as psychologically healthy, a number of their Rorschach scores diverged from the CS reference norms for children; at times dramatically. Many of the differences were similar to those found with adults (e.g., lower form quality values, less color, more use of unusual blot locations, less complexity), though the values Hamel et al. reported tended to be more extreme. At least in part, this appears due to the fact that all protocols were administered

and scored by one graduate student who followed atypical procedures for identifying inkblot locations. This in turn led to a very high frequency of unusual detail locations and consequently to lower form quality codes (see Viglione & Meyer, 2007). However, other child and adolescent samples in the United States, France, Italy, Japan, and Portugal (Erdberg, 2005; Erdberg & Shaffer, 1999) suggest clinicians should be cautious about applying the old CS norms for children. The CS normative data for children have not been updated recently like they have for adults.

Based on the available evidence, we recommend that examiners use the new CS sample as their primary benchmark for adults, but adjust for those variables that have consistently looked different in international samples, including form quality, unusual locations, color, texture, and human representations (for specific recommendations see Table 8.2). The Shaffer et al. sample can be viewed as an outer boundary for what might be expected from reasonably functioning people within the limits of current administration, inquiry, and scoring guidelines.

For children, we recommend using the available CS age-based norms along with the adjusted expectations given in Table 8.2 for adults. Although we do not recommend using the Hamel et al. sample as an outer boundary for what could be expected for younger United States children, the data for that sample illustrate how ambiguity or flexibility in current administration and scoring guidelines can result in one obtaining some unhealthy looking data from apparently normal functioning children. Besides Hamel et al. (2000), child and adolescent reference samples have been collected by other examiners in the United States, France, Italy, Japan, and Portugal (Erdberg & Shaffer, 1999; Erdberg, 2005). Although these samples vary in age, they also show unexpected variability in a number of scores, particularly Dd (small or unusual locations), Lambda (proportion of responses determined just by form), and form quality scores. These scores differ notably from sample to sample. It is unclear if these differences reflect genuine cultural differences in personality and/or in childrearing practices or if they are artifacts due to variability in the way the protocols were administered, inquired, or scored. However, the composite of data suggest that the adjustments offered above for adults should be made for children too.

In addition, clinicians working with children should consider developmental trends. Wenar and Curtis (1991) illustrated these trends for Exner's (2001) child reference data across the ages from 5 to 16. Although limited, the available international data suggest similar developmental trends are present, including age-based increases in complexity markers like DQ+, Blends, and Zf, as well as increases in M and P. In addition, as children age there is a decrease in WSum6 and to a lesser extent in DQv. Unlike Exner's CS reference samples, however, the alternative reference samples for children generally show that as children get older there is a decrease in Lambda and

Table 8.2 Recommended Adjustments to Adult CS Normative Expectations

Variable	New guidelines based on international samples	Old guidelines based on the current CS reference Sample[a]
Location and form quality		
Dd	3 or 4	1 or 2
X-%	.15–.25	.09–.14
X+%	.45–.60	.65–.70
XA%	.70–.90	.80–.95
WDA%	.80–.90	.85–.95
Avoidant style (Lambda > .99)	2 or 3 of 10 people	1 of 10 people
Human representations		
Pure H	2 or 3	3 or 4
H : Non pure H	H+1 = Non pure H	H > Non pure H
COP	1	2
AG	1 in 2 people	1 per person
GHR to PHR ratio (HRV)	Between 3:2 and 1:1 ratio	2:1 ratio
Color and associated variables		
FC: CF+C	FC = or < CF+C	FC > CF+C +1
WSumC	2.5–3.5	4.5
Afr	.45–.55	.55–.65
Extratensive	1 or 2 of 10 people	3 of 10 people
Ambitent	3 or 4 of 10 people	2 of 10 people
EA	6–8	9
Texture		
T = 0	5 to 7 of 10 people	2 of 10 people
T = 1	2 or 3 of 10 people	6 of 10 people
T ≥ 2	1 or 2 of 10 people	2 of 10 people

Note: [a] Exner & Erdberg, 2005, N = 450

an increase in healthier form quality scores. The field would benefit from additional carefully designed studies that examine developmental processes as expressed on the Rorschach.

Although the research evidence reviewed in this section supports the validity of the Rorschach across ethnic groups in the United States and across languages and cultures around the world, this does not mean that culture and ethnicity are unimportant when using the Rorschach. To the contrary, it is important for clinicians to recognize the ways in which culture and acculturation influence the development, identity, and personality of any particular individual. It is as important to take these issues into account when interpreting the Rorschach as it is with any other personality test.

Current Controversies

The Rorschach has been controversial almost since its publication. Historically, clinicians have found it useful for their applied work, while academic psychologists have criticized its psychometric foundation and suggested that clinical perceptions of its utility are likely the result of illusory biases. An early and prominent critique by Jensen (1965) gives a flavor of the sharp tone that has characterized some of the criticisms. Jensen asserted that the Rorschach "is a very poor test and has no practical worth for any of the purposes for which it is recommended" (p. 501) and "scientific progress in clinical psychology might well be measured by the speed and thoroughness with which it gets over the Rorschach" (p. 509). Although Exner's (1974, 2003) work with the CS quelled many of these earlier criticisms, over the past decade there has been a renewed and vigorous series of critiques led by James Wood, Howard Garb, and Scott Lilienfeld, including arguments that psychology departments and organizations should discontinue Rorschach training and practice (see e.g., Garb, 1999; Grove, Barden, Garb, & Lilienfeld, 2002; Lilienfeld, Wood, & Garb, 2000). Counterarguments and rejoinders also have been published and at least seven journals have published a special series of articles concerning the Rorschach.[10]

The most thorough of these special series was an 11-article series published in *Psychological Assessment* (Meyer, 1999b; 2001c). Authors participated in a structured, sequential, evidence based debate that focused on the strengths and limitations of using the Rorschach for applied purposes. The debate took place over four iterations, with each containing contributions from authors who tended to be either favorable or critical of the Rorschach's evidence base. At each step, authors read the articles that were prepared in the previous iteration(s) to ensure the debate was focused and cumulative. As noted earlier, Robert Rosenthal was commissioned for this special series to undertake an independent evidence based review of the research literature through a comparative meta-analysis of Rorschach and MMPI-2 validity. In addition, the final summary paper in the series was written by authors with different views on the Rorschach's merits (Meyer & Archer, 2001). They attempted to synthesize what was known, what had been learned, and what issues still needed to be addressed in future research. We strongly encourage any student or psychologist interested in gaining a full appreciation for the evidence and issues associated with the applied use of the Rorschach to read the full series of articles (Dawes, 1999; Garb, Wood, Nezworski, Grove, & Stejskal, 2001; Hiller et al., 1999; Hunsley & Bailey, 1999, 2001; Meyer, 1999a, 2001b; Meyer & Archer, 2001; Rosenthal et al., 2001; Stricker & Gold, 1999; Viglione, 1999; Viglione & Hilsenroth, 2001; Weiner, 2001).

More recently, the Board of Trustees for the Society for Personality Assessment (2005) addressed the debate about the Rorschach. Drawing on

the recent literature, their official statement concluded that the Rorschach produces evidence of reliability and validity that is similar to the evidence obtained for other personality tests. Given this, they concluded that its responsible use in applied practice was justified.

Nonetheless, as we indicated in previous sections, there are still unresolved issues associated with the Rorschach's evidence base and applied use. Some of the most important issues concern recently recognized variability in the way the CS can be administered and scored when examiners are trying to follow Exner's (2003) current guidelines, the related need to treat normative reference values more tentatively, the impact of response-complexity on the scores obtained in a structural summary, and the need for more research into the stability of scores over time.

Another issue that we have not previously discussed concerns the evidence base for specific scores. The meta-analytic evidence provides a systematic review for several individual variables in relation to particular criteria (e.g., the ROD and observed dependent behavior; the Prognostic Rating Scale and outcome from treatment), but much of the systematically gathered literature speaks to the global validity of the test, which is obtained by aggregating evidence across a wide range of Rorschach scores and a wide range of criterion variables. It would be most helpful to have systematically organized evidence concerning the construct validity of each score that is considered interpretively important. Accomplishing this is a daunting task that initially requires cataloging the scores and criterion variables that have been examined in every study over time. Subsequently, researchers would have to reliably evaluate the methodological quality of each article so greater weight could be afforded to more sturdy findings. Finally, researchers would have to reliably classify the extent to which every criterion variable provides an appropriate match to the construct thought to be assessed by each Rorschach score so that one could meaningfully examine convergent and discriminant validity. Although conducting this kind of research would be highly desirable, we also note how no cognitive or personality test in use today has this kind of focused meta-analytic evidence attesting to the validity of each of its scales in relation to specific and appropriate criterion variables. We say this not as an excuse or a deterrent, but simply as an observation. Because of the criticisms leveled against the Rorschach having this kind of organized meta-analytic evidence is more urgent for it than for other tests.

Clinical Dilemma

Dr. A is a 30-year-old unmarried Asian man who has been in the United States for 5 years and is employed as a university math professor. Two months before being referred for psychological assessment, he was evaluated psychiatrically

for the first time in his life and diagnosed with major depression, for which he was receiving antidepressants by a psychiatrist and weekly cognitive-behavioral psychotherapy by an outpatient psychotherapist. His depression has been present for 2 years, with symptoms of weakness, low energy, sadness, hopelessness, and an inability to concentrate that fluctuated in severity. At the time of assessment, he taught and conducted research for about 40 hours per week and spent almost all of his remaining time in bed. He denied any previous or current hypomanic symptoms, had normal thyroid functions, and reported no other health problems. In his home country, his father had been hospitalized for depression, his brother diagnosed with schizophrenia, and his sister was reported to have "problems" but had not received psychiatric care. His father was physically abusive to his mother, his siblings, and him. Dr. A reported that his father hit him in the face or head on an almost weekly basis while growing up. He is the only one in his family in the United States and he has no history of intimate relationships, though sees several friends for dinner approximately every other week.

Dr. A's outpatient therapist requested the evaluation to assess the severity of Dr. A's depression and to understand his broader personality characteristics. In particular, the therapist wondered about potential paranoid characteristics. Dr. A was primarily interested in whether he had qualities similar to his father or brother and, if so, what he could do to prevent similar conditions from becoming full blown in him. The assessment involved an interview, several self-report inventories (including the MMPI-2, BDI, and a personality disorder questionnaire), and the Rorschach.

Dr. A produced a very complex Rorschach protocol with 42 responses, of which only 8 were determined by straightforward form features (i.e., the percent of pure form responses [Form%] was .19 and the proportion of pure form to non-pure form responses [Lambda] was .24). As a result, his protocol was an outlier relative to the CS norms. The complexity of his record appeared to be a function of his intelligence, his desire to be thorough in the assessment, and also some difficulty stepping back from the task with a consequent propensity to become overly engaged with the stimuli (particularly to the last three brightly colored cards, to which he produced almost half of his responses [20 of 42]). After adjusting for the length and complexity of his protocol, Dr. A exhibited some notable features. First, his thought processes were characterized by implausible and illogical relationships, with the weighted sum of cognitive special scores (see Table 8.1) several standard deviations above what is typically seen in nonpatient or even outpatient samples. Importantly, however, this occurred in the context of perceptions that had typical and conventional form features (XA%, which is the percent of all responses with adequate form quality, was .79 and WDA%, which is the percent of responses to the whole card or to common detail locations

with adequate form quality, was .92). In addition, even though he would be considered to have extensive assets for coping with life demands (M = 18, Weighted Color = 14.5, Zf = 33, DQ+ = 22), he saw an unexpectedly large number of inanimate objects in motion (m = 7), suggesting he was experiencing a considerable degree of uncontrollable environmental stress, internal tension, and agitated cognitive activity. Finally, he had a marked propensity to perceive objects engaged in aggressive activity (AG = 8) and to identify percepts where objects were damaged, decaying, or dying (MOR = 10). This combination of scores suggested he had an implicit depressive perceptual filter in which he experienced himself as deficient, vulnerable, and incapable of contending with a dangerous, menacing, and combative environment.

Although this chapter does not provide the actual inkblot images, we include his responses from a number of the cards to give a flavor of the characteristics described above. As a general principle, response verbalizations should be considered after examining the previously presented quantitative data so as to minimize the prospect for erroneous speculations.

At the bottom of the second card, Dr. A saw, "Blood. Yeah, I don't really want to say—it's dirty words—but it looks like an asshole with blood coming out of it . . . spilling over, all over the place." A bit later using the entire card he saw, "the face of a human being . . . looks like its weeping. It may be partly vomiting… The eyes look like they're teary… this is what it's vomiting." To the third card Dr. A saw "two people meeting and bowing to each other, but they're kind of hating each other…this red thing signifies the hatred between the two people." In his next response he saw "two ugly waitresses—actually they look like birds—who are bringing some strange plate or dish… I mean gruesome stuff like snakes, spiders, something like that." On the next card he saw "a gruesome monster… as tall as a tower…it's about to come and crush me out. He looks very angry at me… these look like his hands but also like a weapon and it's very, very dangerous…the whole posture makes me feel like it's angry. I don't see any specific… maybe the only thing that makes me feel that way is the hidden expressions." The final response to this card consisted of "a small animal… which has been killed on a street by a car—flattened out… sometimes you can see small animals dying on the road." On the fifth card he returned to the same themes, seeing "a butterfly which is kind of dying—injured and dying" and "a witch with two horns… trying to approach me and catch me… some massive, dark object." On the ninth card he saw "a knife thrust into a body and blood is coming out as a result," which was followed by the perception of "two monsters… who are maybe shaking hands," and then a new response of "three people… sitting in a row… controlling from behind… the red person controlling the green one and the green one is controlling the yellow one." On the final card, Dr. A saw "an abdomen of organs which are not functioning because of the various poisons. The organs

are poisoned, as you see from the colors… weak and not functioning… very bad condition." In another response to the whole inkblot he saw "an island as you see it from the skies. Island where there is a military secret. So it's very secret. And they are hiding the ships and weapons in the very center of the island. So they make use of the very complicated coastline. And they made a lot of traps so that you can't very easily approach the center of the island… traps to capture the enemies." This response was followed by "interior walls of some organ, like stomach or heart… these look like ulcers… this portion looks deteriorated, somehow damaged." Next he saw "a flying monster which is about to attack—attack something with its chisel-like mouth." As his final response to the task, Dr. A saw "two people fighting with weapons… they don't have heads somehow."

Although this is incomplete information, the curious reader could stop here and ponder several questions. To what extent do the scores and the images or themes in his responses suggest that Dr. A is depressed? Dr. A's outpatient therapist was concerned about paranoid characteristics. Do the data suggest that concerns in this regard are warranted? Also, do the results suggest that Dr. A might have other personality characteristics or personality struggles that were not part of the initial referral question but that will be important to consider? Dr. A was concerned about the possibility that he was like his brother who had a schizophrenic disorder. What features of the data would be consistent with a psychotic disturbance? Alternatively, are there features of the data that would contradict a disorder on the psychotic spectrum? These are important questions to address and how they are addressed will have significant consequences for Dr A. Thus, although we focus in this chapter on just the Rorschach data, in actual practice the assessment clinician would need to carefully consider each question while taking into account the full array of available information from testing and from history.

With respect to the Rorschach data, Dr. A's vivid images provide idiographic insight into his particular way of experiencing the qualities suggested by the relatively impersonal quantitative structural summary variables. We learn and come to understand his deep fears, fragile vulnerabilities, and powerful preoccupation with aggression and hostility. As suggested in his last response, identification with aggression is likely to leave him feeling "headless" and out of control. Although generally it is not possible to determine whether clients positively identify with aggressive images or fear them as dangers emanating from the environment, the extensive morbid imagery of damaged, decaying, dying, pierced, and poisoned objects all suggest the latter (as did his denial of anger and aggressiveness on self-report inventories). Depression, at least for some people, can be understood as aggression turned toward the self rather than directed outward at its intended target. Given the pervasiveness of aggressive imagery in his Rorschach protocol,

Dr. A's therapist could pursue this hypothesis in her work with him after he stabilized at a more functional level.

Paranoid themes were also evident in Dr. A's responses (e.g., people bowing in respect but internally hating each other, "bird" waitresses serving snakes or spiders, creatures with weapons for appendages, hidden expressions, secretive traps guarding weapons, external control by others). In combination with the disrupted formal thought processes seen on his Rorschach and results

Key Points to Remember:

- The Rorschach provides a sample of behavior obtained under standardized conditions in response to artistically elaborated visual stimuli in which problem solving operations are elicited by the prompt "What might this be?"
- The term "projective" is not a good label to describe the type of information obtained by the Rorschach (and the term "objective" is not a good label to describe the type of information obtained from self-report inventories).
- Rorschach responses can be reliably scored on a wide number of variables that characterize structural, perceptual, or thematic features of the response.
- The Rorschach Comprehensive System (CS) is the approach to administration and scoring that is most commonly taught, used in clinical practice, and researched. When the CS was developed, it integrated the most reliable and valid features of five previous systems used in the United States.
- At the present time, some scores that fall outside the CS have a larger body of psychometric evidence supporting their use than some scores within the CS.
- Meta-analytic summaries support Rorschach reliability for scoring and the stability of its scores over time.
- Meta-analytic summaries support the general validity of the Rorschach across scales that have been subjected to research. Globally, it is as valid as other personality tests.
- Meta-analytic summaries support the focused validity of the Rorschach for predicting dependent behavior, assessing disordered thinking and psychotic disorders, predicting response to therapy, and quantifying change as a result of therapy. However, the CS Depression Index does not validly identify patients with a diagnosed depressive disorder.
- Recent evidence suggests some of the seeming differences between normative samples collected in the United States and internationally are likely due to unexpected differences in local benchmarks used for administration and scoring.
- The Rorschach is considered a valuable asset in clinical practice because it is an office based procedure that provides a unique method for observing personality characteristics.
- Characteristics assessed by Rorschach scores are not necessarily represented in conscious awareness and they reflect perceptual, schematic, or processing propensities rather than focused, overt, and conscious symptoms. To understand how these propensities are experienced and expressed, Rorschach data needs to be integrated with other sources of information.

from the other tests he completed, Dr. A was considered to be experiencing a severe agitated depressive episode with psychotic features. This was considered a conservative diagnosis because psychological assessment provides a snapshot of current functioning so it was not possible to determine whether a major depressive disorder was co-occurring with an independent and longer standing delusional disorder. However, the latter seemed less likely, given the pervasiveness of his affective turmoil and the fact that the form quality of his perceptions remained healthy and conventional despite such a lengthy and complex protocol. In feedback to Dr. A, his therapist, and his psychiatrist, it was recommended that Dr. A begin antipsychotic medication on at least a trial basis and that therapy be ego-supportive rather than uncovering, with an emphasis on cognitive interventions to evaluate suspicions and correct his propensity to misattribute aggressive intentions onto others in the environment.

Chapter Summary

It is not possible to learn Rorschach administration, scoring, and interpretation from a chapter like this. Consequently, our goal was to provide readers with an overview of the Rorschach as a task that aids in assessing personality. We described the instrument and the approaches that have been used to develop test scores. We then focused on the psychometric evidence for reliability, showing that its scores can be reliably assigned, are reasonably stable over time, and can be reliably interpreted by different clinicians. We also focused on evidence related to its validity and utility, showing that it is a generally valid method of assessment that provides unique and meaningful information for clinical practice. In the process, we pointed out the kinds of information the test generally can and cannot provide and provided psychometrically based guidelines to aid with interpretation. Next, we reviewed current evidence associated with its multicultural and cross-national use and noted a need for tighter guidelines governing administration and scoring to ensure consistency in the data that is collected across sites around the world. Finally, we provided a case vignette that illustrated how a person's perceptions could be meaningfully interpreted in idiographic clinical practice even in the absence of the inkblot stimuli themselves.

Although additional research and refinement are needed on numerous fronts, the systematically gathered data indicate there is solid evidence supporting the Rorschach's basic reliability and validity. Overall, we advocate for an evidence-based, behavioral- representation approach to conceptualizing the test that attempts to focus on concrete and experience near test-based inferences at the expense of more elusive abstract ones. We hope readers will pursue some of the additional readings we have suggested and other studies

we have cited. Also, we urge readers to seek out high quality training from qualified supervisors so they can experience the Rorschach's strengths and limitations first hand. Doing so will provide important experiential data about the test's utility that will help when considering the evidence presented here and the recurrent controversy about this unique instrument.

We close with a final caution to keep in mind when considering some of the controversy associated with the Rorschach. Consistent with evidence based principles, we urge readers to attend to the systematically generated evidence and to be wary of partial reviews or selective citations. On average, personality and cognitive tests produce heteromethod validity coefficients that are about equal to a correlation of .30 (Meyer et al., 2001). This means that about half of the research literature will produce validity coefficients that are lower than this and about half will produce coefficients that are higher. Authors who selectively cite the literature or focus on just a subset of individual studies can (inadvertently or intentionally) make the literature seem more or less supportive than is actually warranted.

Notes

1. The authors would like to thank Joni L. Mihura and Aaron D. Upton for their helpful comments and suggestions.
2. Historically, the Rorschach was classified as a "projective" rather than "objective" test. However, these archaic terms are global and misleading descriptors that should be avoided because they do not adequately describe instruments or help our field develop a more advanced and differentiated understanding of personality assessment methods (see Meyer & Kurtz, 2006).
3. There are other inkblot stimuli that have been developed and researched over the years, including a complete system by Holtzman, a series by Behn-Eschenberg that was initially hoped to parallel Rorschach's blots, a short 3-card series by Zulliger, an infrequently researched set by Roemer, and the Somatic Inkblots, which are a set of stimuli that were deliberately created to elicit responses containing somatic content or themes.
4. For ICC or kappa values, findings above .74 are considered excellent, above .59 are considered good, and above .39 are considered fair (Cicchetti, 1994; Shrout & Fliess, 1979).
5. At the same time, data clearly show that Rorschach scales validly identify psychotic diagnoses and validly measure psychotic symptoms (Lilienfeld, Wood, & Garb, 2000; Meyer & Archer, 2001; Perry, Minassian, Cadenhead, & Braff, 2003; Viglione, 1999, Viglione & Hilsenroth, 2001; Wood, Lilienfeld, Garb, & Nezworski, 2000). Unlike most other disorders, which are heavily dependent on the patient's self-reported symptoms, psychotic conditions are often diagnosed based more on the patient's observed behavior than on their specific reported complaints.
6. At present, one or more national Rorschach societies exist in the following countries: Argentina, Brazil, Canada, Cuba, Czech Republic, Finland, France, Israel, Italy, Japan, The Netherlands, Peru, Portugal, South Africa, Spain, Sweden, Switzerland, Turkey, United States, and Venezuela.
7. Fully structured interviews can be differentiated from semistructured interviews. To some degree, semistructured interviews allow a clinician's inferences to influence the final scores or determinations from the assessment. However, the inferences and determinations remain fundamentally grounded in the client's self-reported characteristics. Fully structured interviews are wholly dependent on this source of information.
8. The Rorschach's first factor is a dimension of complexity. The first factor of a test indicates the primary feature it measures. The Rorschach's first factor typically accounts for about 25% of the total variance in Rorschach scores. For self-report scales like the MMPI-2 or MCMI,

the first factor, which is a dimension of willingness versus reluctance to report problematic symptoms, typically accounts for more than 50% of the total variance in scores (see Meyer et al., 2000).

9. There was evidence suggesting that CS psychosis indicators may underpredict pathology in AAs, a finding that also has been observed with MMPI-2 psychosis indicators (Arbisi, Ben-Porath, & McNulty, 2002), though it was not possible to fully evaluate this finding.
10. These journals include *Assessment*; *Clinical Psychology: Science and Practice*; *Journal of Clinical Psychology*; *Journal of Forensic Psychology Practice*; *Journal of Personality Assessment*; *Psychology, Public Policy, and Law*; and *Psychological Assessment*.

References

Acklin, M. W. (2000). Rorschach Interpretive Assistance Program: Version 4 for Windows [Software Review]. *Journal of Personality Assessment, 75*, 519–521

Acklin, M. W., McDowell, C. J., & Verschell, M. S. (2000). Interobserver agreement, intraobserver reliability, and the Rorschach Comprehensive System. *Journal of Personality Assessment, 74,* 15–47.

Allen, J., & Dana, R. H. (2004). Methodological issues in cross-cultural and multicultural Rorschach research. *Journal of Personality Assessment, 82*, 189–206.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). New York: Macmillan.

Arbisi, P. A., Ben-Porath, Y. S., & McNulty, J. (2002). A comparison of MMPI–2 validity in African American and Caucasian psychiatric inpatients. *Psychological Assessment, 14,* 3–15.

Aronow, E., Reznikoff, M. & Moreland, K. L. (1995). The Rorschach: Projective technique or psychometric test? *Journal of Personality Assessment, 64*, 213–228.

Atkinson, L., Quarrington, B., Alp, I. E., & Cyr, J. J. (1986). Rorschach validity: An empirical approach to the literature. *Journal of Clinical Psychology, 42,* 360–362.

Balcetis, E., & Dunning, D. (2006). See what you want to see: Motivational influences on visual perception. *Journal of Personality and Social Psychology, 91*, 612–625.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory – II*. San Antonio, TX: Psychological Corporation.

Bihlar, B., & Carlsson, A. M. (2001). Planned and actual goals in psychodynamic psychotherapies: Do patients' personality characteristics relate to agreement? *Psychotherapy Research, 11*, 383–400.

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*(4), 364–373.

Bornstein, R. F. (1996). Construct validity of the Rorschach Oral Dependency Scale: 1967-1995. *Psychological Assessment, 8*, 200–505.

Bornstein, R. F. (1998). Implicit and self-attributed dependency strivings: Differential relationships to laboratory and field measures of help-seeking. *Journal of Personality and Social Psychology, 75,* 779–787.

Bornstein, R. F. (1999). Criterion validity of objective and projective dependency tests: A meta-analytic assessment of behavioral prediction. *Psychological Assessment, 11*, 48–57.

Bornstein, R. F. (2001). Clinical utility of the Rorschach Inkblot Method: Reframing the debate. *Journal of Personality Assessment, 77*, 39–47.

Bornstein, R. F. (2002). A process dissociation approach to objective-projective test score interrelationships. *Journal of Personality Assessment*, *78*, 47–68.

Bornstein, R. F., & Masling, J. M. (Eds.) (2005). *Scoring the Rorschach: Seven validated systems*. Mahwah, NJ: Erlbaum.

Bornstein, R. F., Hill, E. L., Robinson, K J., Calabreses, C., & Bowers, K. S. (1996). Internal reliability of Rorschach Oral Dependency Scale scores. *Educational and Psychological Measurement, 56*, 130–138.

Butcher, J. N., & Rouse, S. (1996). Clinical personality assessment. *Annual Review of Psychology,* 47, 87–111.

Butcher, J. N., Dahlstrom,W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *MMPI-2: Minnesota Multiphasic Personality Inventory-2: Manual for administration and scoring*. Minneapolis: University of Minnesota Press.

Butcher, J. N., Nezami, E., & Exner, J. (1998). Psychological assessment of people in diverse cultures. In S. S. Kazarian & D. R. Evans (Eds.), *Cultural clinical psychology: Theory, research, and practice* (pp. 61–105). New York: Oxford University Press.

Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice, 31,* 141–154.

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

Childs, R. A., & Eyde, L. D. (2002). Assessment training in clinical psychology doctoral programs: What should we teach? What do we teach? *Journal of Personality Assessment, 78,* 130–144.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284–290.

Clemence, A. J., & Handler, L. (2001). Psychological assessment on internship: A survey of training directors and their expectations for students. *Journal of Personality Assessment, 76*, 18–47.

Dao, T. K., & Prevatt, F. (2006). A psychometric evaluation of the Rorschach Comprehensive System's Perceptual Thinking Index. *Journal of Personality Assessment, 86*, 180–189.

Dao, T. K., Prevatt, F., & Horne, H. L. (in press). Differentiating psychotic patients from non-psychotic patients with the MMPI-2 and Rorschach. *Journal of Personality Assessment*.

Dawes, R. M. (1999). Two methods for studying the incremental validity of a Rorschach variable. *Psychological Assessment, 11*, 297–302.

Dean, K. L., Viglione, D. J., Perry, W., & Meyer, G. J. (in press). A method to increase Rorschach response productivity while maintaining Comprehensive System validity. *Journal of Personality Assessment*.

Elfhag, K., Barkeling, B., Carlsson, A. M., & Rössner, S. (2003). Microstructure of eating behavior associated with Rorschach characteristics in obesity. *Journal of Personality Assessment, 81*, 40–50.

Elfhag, K., Barkeling, B., Carlsson, A. M., Lindgren, T., & Rössner, S. (2004). Food intake with an antiobesity drug (sibutramine) versus placebo and Rorschach data: A crossover within-subjects study. *Journal of Personality Assessment, 82*, 158–168.

Elfhag, K., Carlsson, A. M. & Rössner, S. (2003). Subgrouping in obesity based on Rorschach personality characteristics. *Scandinavian Journal of Psychology, 44*, 399–407.

Elfhag, K., Rössner, S., Carlsson, A. M., & Barkeling, B. (2003). Sibutramine treatment in obesity: Predictors of weight loss including Rorschach personality data. *Obesity Research, 11*, 1391–1399.

Elfhag, K., Rössner, S., Lindgren, T., Andersson, I., & Carlsson, A. M. (2004). Rorschach personality predictors of weight loss with behavior modification in obesity treatment. *Journal of Personality Assessment, 83*, 293–305.

Erdberg, P. (2005, July). *Intercoder Agreement as a Measure of Ambiguity of Coding Guidelines.* Paper presented at the XVIII International Congress of the Rorschach and Projective Methods, Barcelona, Spain.

Erdberg, P., & Shaffer, T. W. (1999, July). *International symposium on Rorschach nonpatient data: Findings from around the world.* Paper presented at the International Congress of Rorschach and Projective Methods, Amsterdam, The Netherlands.

Exner, J. E. (1974). *The Rorschach: A comprehensive system, Vol. 1*. New York: Wiley.

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

Exner, J. E. (1996). Critical bits and the Rorschach response process. *Journal of Personality Assessment*, *67*, 464–477.

Exner, J. E. (2003). *The Rorschach: A comprehensive system, Volume 1* (4th ed.). New York: Wiley.

Exner, J. E. (with Colligan, S. C., Hillman, L. B., Metts, A. S., Ritzler, B., Rogers, K. T., Sciara, A., D., & Viglione, D. J.) (2001). *A Rorschach workbook for the Comprehensive System* (5th ed.). Asheville, NC: Rorschach Workshops.

Exner, J. E., & Erdberg, P. (2005). *The Rorschach: A Comprehensive System, Volume 2: Advanced Interpretation* (3rd ed.). Oxford: Wiley.

Exner, J. E., Jr. (2001). *A Rorschach Workbook for the Comprehensive System* (5th ed.). Asheville, NC: Rorschach Workshops.

Fischer, C. T. (1994). Rorschach scoring questions as access to dynamics. *Journal of Personality Assessment, 62*, 515–524.

Foster, S. L., & Cone, J. D. (1995). Validity issues in clinical assessment. *Psychological Assessment, 7,* 248–260.

Fowler, J. C., Piers, C., Hilsenroth, M. J., Holdwick, D. J., & Padawer, J. R. (2001). The Rorschach suicide constellation: Assessing various degrees of lethality. *Journal of Personality Assessment, 76,* 333–351.

Garb, H. N. (1999). Call for a moratorium on the use of the Rorschach Inkblot Test in clinical and forensic settings. *Assessment, 6*, 313–317.

Garb, H. N., Wood, J. M., Nezworski, M. T., Grove, W. M., & Stejskal, W. J. (2001). Towards a resolution of the Rorschach controversy. *Psychological Assessment, 13*, 433–438.

Grønnerød, C. (2003). Temporal stability in the Rorschach method: A meta-analytic review. *Journal of Personality Assessment, 80*(3), 272–293.

Grønnerød, C. (2004). Rorschach assessment of changes following psychotherapy: A meta-analytic review. *Journal of Personality Assessment, 83,* 256–276.

Grove, W. M., Barden, R. C., Garb, H. N., & Lilienfeld, S. O. (2002). Failure of Rorschach-Comprehensive-System-based testimony to be admissible under the Daubert-Joiner-Kumho standard. *Psychology, Public Policy, & Law, 8*, 216–234.

Hamel, M., Shaffer, T. W., & Erdberg, P. (2000). A study of nonpatient preadolescent Rorschach protocols. *Journal of Personality Assessment, 75*, 280–294.

Hartmann, E., Nørbech, P. B., & Grønnerød, C. (2006). Psychopathic and nonpsychopathic violent offenders on the Rorschach: Discriminative features and comparisons with schizophrenic inpatient and university student samples. *Journal of Personality Assessment, 86*, 291–305.

Hartmann, E., Sunde, T., Kristensen, W., & Martinussen, M. (2003). Psychological measures as predictors of military training performance. *Journal of Personality Assessment, 80*, 88–99.

Hartmann, E., Wang, C., Berg, M., & Sæther, L. (2003). Depression and vulnerability as assessed by the Rorschach method. *Journal of Personality Assessment, 81*, 243–256.

Hiller, J. B., Rosenthal, R., Bornstein, R. F., Berry, D. T. R., & Brunell-Neuleib, S. (1999). A comparative meta-analysis of Rorschach and MMPI validity. *Psychological Assessment, 11*, 278–296.

Hilsenroth, M. J., & Handler, L. (1995). A survey of graduate students' experiences, interests, and attitudes about learning the Rorschach. *Journal of Personality Assessment, 64*, 243–257.

Hunsley, J., & Bailey, J. M. (1999). The clinical utility of the Rorschach: Unfulfilled promises and an uncertain future. *Psychological Assessment, 11*, 266–277.

Hunsley, J., & Bailey, J. M. (2001). Wither the Rorschach? An analysis of the evidence. *Psychological Assessment, 13*, 472–485.

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment, 15*, 446–455.

Janson, H., & Stattin, H. (2003). Prediction of adolescent and adult antisociality from childhood Rorschach ratings. *Journal of Personality Assessment, 81*, 51–63.

Jensen, A. R. (1965). Review of the Rorschach Inkblot Test. In O. K. Buros (Ed.), *The sixth mental measurements yearbook* (pp. 501–509). Highland Park, NJ: Gryphon Press.

Krishnamurthy, R., Archer, R. P., & House, J. J. (1996). The MMPI-A and Rorschach: A failure to establish convergent validity. *Assessment, 3*, 179–191.

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest, 1*, 27–66.

Lis, A., Parolin, L., Calvo, V., Zennaro, A., & Meyer, G. J. (in press). The impact of administration and inquiry on Rorschach Comprehensive System protocols in a national reference sample. *Journal of Personality Assessment*.

Long, G. M., & Toppino, T. C. (2004). Enduring interest in perceptual ambiguity: Alternating views of reversible figures. *Psychological Bulletin, 130*, 748–768.

Lundbäck, E., Forslund, K., Rylander, G., Jokinen, J., Nordström, P., Nordström, A.-L., et al. (2006). CSF 5-HIAA and the Rorschach test in patients who have attempted suicide. *Archives of Suicide Research, 10*, 339–345.

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.

McAdams, D. P., & Constantian, C. A. (1983). Intimacy and affiliation motives in daily living: An experience sampling analysis. *Journal of Personality and Social Psychology, 45*, 851–861.

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, CA: Sage.

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

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

McGrath, R. E., Pogge, D. L., Stokes, J. M., Cragnolino, A., Zaccario, M., Hayman, J., Piacentini, T., & Wayland-Smith, D. (2005). Field reliability of comprehensive system scoring in an adolescent inpatient sample. *Assessment*, 12, 199–209.

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

Meyer, G. J. (1996). The Rorschach and MMPI: Toward a more scientifically differentiated understanding of cross-method assessment. *Journal of Personality Assessment, 67*, 558–578.

Meyer, G. J. (1997). Assessing reliability: Critical corrections for a critical examination of the Rorschach Comprehensive System. *Psychological Assessment, 9*, 480–489.

Meyer, G. J. (1999a). Introduction to the Special Series on the utility of the Rorschach for clinical assessment. *Psychological Assessment, 11*, 235–239.

Meyer, G. J. (2000a). Incremental validity of the Rorschach Prognostic Rating Scale over the MMPI Ego Strength scale and IQ. *Journal of Personality Assessment, 74*, 356–370.

Meyer, G. J. (2000b). On the science of Rorschach research. *Journal of Personality Assessment, 75*(1), 46–81.

Meyer, G. J. (2001a). Evidence to correct misperceptions about Rorschach norms. *Clinical Psychology: Science & Practice, 8*, 389–396.

Meyer, G. J. (2001b). Introduction to the final Special Section in the Special Series on the utility of the Rorschach for clinical assessment. *Psychological Assessment, 13*, 419–422.

Meyer, G. J. (2002a). Exploring possible ethnic differences and bias in the Rorschach Comprehensive System. *Journal of Personality Assessment, 78*, 104–129.

Meyer, G. J. (2002b). Implications of information-gathering methods for a refined taxonomy of psychopathology. In L. E. Beutler & M. Malik (Eds.). *Rethinking the DSM: Psychological perspectives* (pp. 69–105). Washington, DC: American Psychological Association.

Meyer, G. J. (2004). The reliability and validity of the Rorschach and TAT compared to other psychological and medical procedures: An analysis of systematically gathered evidence. In M. Hilsenroth & D. Segal (Eds.), Personality assessment. Vol. 2 in M. Hersen (Ed.-in-Chief), *Comprehensive handbook of psychological assessment* (pp. 315–342). Hoboken, NJ: Wiley.

Meyer, G. J. (Ed.). (1999b). Special Section I: The utility of the Rorschach for clinical assessment [Special Section]. *Psychological Assessment, 11*, 235–302.

Meyer, G. J. (Ed.). (2001c). Special Section II: The utility of the Rorschach for clinical assessment [Special Section]. *Psychological Assessment, 13*, 419–502.

Meyer, G. J., & Archer, R. P. (2001). The hard science of Rorschach research: What do we know and where do we go? *Psychological Assessment, 13*, 486–502.

Meyer, G. J., & Handler, L. (1997). The ability of the Rorschach to predict subsequent outcome: A meta-analysis of the Rorschach prognostic rating scale. *Journal of Personality Assessment, 69*, 1–38.

Meyer, G. J., & Kurtz, J. E. (2006). Guidelines editorial—Advancing personality assessment terminology: Time to retire "objective" and "projective" as personality test descriptors. *Journal of Personality Assessment, 87*, 1–4.

Meyer, G. J., Finn, S. E., Eyde, L., Kay, G. G., Moreland, K. L., Dies, R. R., et al. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56*, 128–165.

Meyer, G. J., Hilsenroth, M. J., Baxter, D., Exner, J. E., Jr., Fowler, J. C., Piers, C. C., et al. (2002). An examination of interrater reliability for scoring the Rorschach Comprehensive System in eight data sets. *Journal of Personality Assessment, 78*, 219–274.

Meyer, G. J., Riethmiller, R. J., Brooks, R. D., Benoit, W. A., & Handler, L. (2000). A replication of Rorschach and MMPI-2 convergent validity. *Journal of Personality Assessment, 74*(2), 175–215.

Meyer, G. J., Viglione, D. J., Erdberg, P., Exner, J. E., Jr., & Shaffer, T. (2004, March). *CS scoring differences in the Rorschach Workshop and Fresno nonpatient samples*. Paper presented at the annual meeting of the Society for Personality Assessment, Miami, FL, March 11.

Mihura, J. L., & Weinle, C. A. (2002). Rorschach training: Doctoral students' experiences and preferences. *Journal of Personality Assessment, 79*, 39–52.

Millon, T. (1994). *Manual for the MCMI-III.* Minneapolis, MN: National Computer Systems.

Morey, L. C. (1991). *Personality Assessment Inventory: Professional manual*. Odessa, FL: Psychological Assessment Resources.

Muniz, J., Prieto, G., Almeida, L., & Bartram, D. (1999). Test use in Spain, Portugal, and Latin American countries. *European Journal of Psychological Assessment, 15*, 151–157.

Nygren, M. (2004). Rorschach Comprehensive System variables in relation to assessing dynamic capacity and ego strength for psychodynamic psychotherapy. *Journal of Personality Assessment, 83*, 277–292.

Paulhus, D. L., Lysy, D. C., & Yik, M. S. M. (1998). Self-report measures of intelligence: Are they useful as proxy IQ tests? *Journal of Personality, 66,* 525–554.

Payne, B. K., Cheng, C. M., Govorun, O. & Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology, 89*, 277–293.

Peebles-Kleiger, M. J. (2002). Elaboration of some sequence analysis strategies: Examples and guidelines for level of confidence. *Journal of Personality Assessment, 79*, 19–38.

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.

Perry, W., Minassian, A., Cadenhead, K., Sprock, J., & Braff, D. (2003). The use of the Ego Impairment Index across the schizophrenia spectrum. *Journal of Personality Assessment*. 80, 50–57.

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

Presley, G., Smith, C., Hilsenroth, M., & Exner, J. (2001). Clinical utility of the Rorschach with African Americans. *Journal of Personality Assessment, 77*(3), 491–507.

Ritsher, J. B. (2004). Association of Rorschach and MMPI psychosis indicators and schizophrenia spectrum diagnoses in a Russian clinical sample. *Journal of Personality Assessment, 83*, 46–63.

Rorschach, H. (1921/1942). *Psychodiagnostics* (5th ed.). Berne, Switzerland: Verlag Hans Huber. (Original work published 1921).

Rorschach, H. (1969). *Psychodiagnostics: A diagnostic test based on perception* (7th ed.) (P. Lemkau & B. Kronenberg, Trans.). Bern, Switzerland: Hans Huber. (Original work published in 1921)

Rosenthal, R., Hiller, J. B., Bornstein, R. F., Berry, D. T. R., & Brunell-Neuleib, S. (2001). Meta-analytic methods, the Rorschach, and the MMPI. *Psychological Assessment, 13*, 449–451.

Rubin, N. J., & Arceneaux, M. (2001). Intractable depression or psychosis. *Acta Psychiatrica Scandinavica, 104*, 402–405.

Schafer, R. (1954). *Psychoanalytic interpretation in Rorschach testing*. New York: Grune & Stratton.

Shaffer, T. W., Erdberg, P., & Haroian, J. (1999). Current nonpatient data for the Rorschach, WAIS-R, and MMPI-2. *Journal of Personality Assessment, 73*(2), 305–316.

Shrout, P.E. & Fliess, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–425.

Smith, S. R., & Hilsenroth, M. J. (2003). ROR–SCAN 6: Rorschach Scoring for the 21st Century [Software review]. *Journal of Personality Assessment, 80*, 108–110.

Society for Personality Assessment (2005). The status of the Rorschach in clinical and forensic practice: An official statement by the Board of Trustees of the Society for Personality Assessment. *Journal of Personality Assessment, 85*, 219–237.

Stokes, J. M., Pogge, D. L., Powell-Lunder, J., Ward, A. W., Bilginer, L., DeLuca, V. A. (2003). The Rorschach Ego Impairment Index: Prediction of treatment outcome in a child psychiatric population. *Journal of Personality Assessment, 81*, 11–19.

Streiner, D. L. (2003a). Being inconsistent about consistency: When coefficient alpha does and doesn't matter. *Journal of Personality Assessment, 80*, 217–222.

Streiner, D. L. (2003b). Starting at the beginning: An introduction to Coefficient Alpha and internal consistency. *Journal of Personality Assessment, 80*, 99–103.

Stricker, G., & Gold, J. R. (1999). The Rorschach: Toward a nomothetically based, idiographically applicable configurational model. *Psychological Assessment, 11*, 240–250.

Sultan, S. (2006). *Is productivity a moderator of the stability of Rorschach scores?* Manuscript submitted for publication.

Sultan, S., Andronikof, A., Réveillère, C., & Lemmel, G. (2006). A Rorschach stability study in a nonpatient adult sample. *Journal of Personality Assessment, 87*.113–119

Sultan, S., Jebrane, A., & Heurtier-Hartemann, A. (2002). Rorschach variables related to blood glucose control in insulin-dependent diabetes patients. *Journal of Personality Assessment, 79*, 122–141.

Urist, J. (1977). The Rorschach test and the assessment of object relations. *Journal of Personality Assessment, 41*, 3–9.
Viglione, D. J. (1999). A review of recent research addressing the utility of the Rorschach. *Psychological Assessment, 11*, 251–265.
Viglione, D. J. (2002). *Rorschach coding solutions: A reference guide for the Comprehensive System*. San Diego, CA: Donald J. Viglione.
Viglione, D. J., & Hilsenroth, M. J. (2001). The Rorschach: Facts, fictions, and future. *Psychological Assessment, 13*(4), 452–471.
Viglione, D. J., & Meyer, G. J. (2007). An overview of Rorschach psychometrics for forensic practice. In C. B. Gacono & F. B. Evans with N. Kaser-Boyd & L. A. Gacono (Eds.) *Handbook of forensic Rorschach psychology* (pp. 21–53). Mahwah, NJ: Erlbaum.
Viglione, D. J., & Rivera, B. (2003). Assessing personality and psychopathology with projective tests. In J. R. Graham & J. A. Naglieri (Eds.), *Comprehensive handbook of psychology: Assessment psychology* (Vol. 10, pp. 531–553). New York: Wiley.
Viglione, D. J., & Taylor, N. (2003). Empirical support for interrater reliability of the Rorschach Comprehensive System coding. *Journal of Clinical Psychology, 59*, 111–121.
Viglione, D. J., Perry, W., & Meyer, G. (2003). Refinements in the Rorschach Ego Impairment Index incorporating the Human Representational Variable. *Journal of Personality Assessment, 81*, 149–156.
Viglione, D. J., Perry, W., Jansak, D., Meyer, G. J., Exner, J. E., Jr. (2003). Modifying the Rorschach Human Experience Variable to create the Human Representational Variable. *Journal of Personality Assessment, 81*, 64–73.
Viglione, D.J. (1996). Data and issues to consider in reconciling self report and the Rorschach. *Journal of Personality Assessment, 67*, 579–587.
Watkins, C. E., Campbell, V. L., Nieberding, R., & Hallmark, R. (1995). Contemporary practice of psychological assessment by clinical psychologists. *Professional Psychology: Research and Practice, 26*, 54–60.
Wechsler, D. (1997). *WAIS–III manual: Wechsler Adult Intelligence Scale* (3rd ed.). San Antonio, TX: Psychological Corporation.
Weiner, I. B. (1977). Approaches to Rorschach validation. In M. A. Rickers-Ovsiankina (Ed.), *Rorschach psychology* (pp. 575–608). Hungtington, NY: Krieger.
Weiner, I. B. (2001). Advancing the science of psychological assessment: The Rorschach Inkblot Method as exemplar. *Psychological Assessment, 13*, 423–434.
Weiner, I. B. (2003). *Principles of Rorschach interpretation* (2nd ed.). Mahwah, NJ: Erlbaum.
Wenar & Curtis (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., Lilienfeld, S. O., Garb, H. N., & Nezworski, M. T. (2000). The Rorschach test in clinical diagnosis: A critical review, with a backward look at Garfield (1947). *Journal of Clinical Psychology, 56*, 395–430.
Wood, J. M., Nezworski, M. T., Garb, H. N., & Lilienfeld, S. O. (2001a). The misperception of psychopathology: Problems with norms of the Comprehensive System for the Rorschach. *Clinical Psychology: Science & Practice, 8*(3), 350–373.
Wood, J. M., Nezworski, M. T., Garb, H. N., & Lilienfeld, S. O. (2001b). Problems with the norms of the Comprehensive System for the Rorschach: Methodological and conceptual considerations. *Clinical Psychology: Science & Practice, 8*(3), 397–402.