CHAPTER

2

An Overview of Rorschach Psychometrics for Forensic Practice

Donald J. Viglione
Alliant International University

Gregory J. Meyer
University of Toledo

This chapter addresses current evidence concerning the Rorschach Inkblot Test relevant to forensic practice. We present a selective overview of research findings and some new data to help explicate the scientific and empirical foundations of the test. The focus is primarily on psychometric issues of reliability, validity, normative reference values, and utility. Even when limiting ourselves to these topics, we are selective because it is not possible to address them comprehensively within a single chapter. We focus on topics of most interest in the forensic arena and that have attracted the most research and controversy lately.[1] There is no attempt to select research that supports or does not support the test, but rather a bias for selecting recent versus older and well-known and established evidence.

This review emphasizes Rorschach variables from the Comprehensive System (CS; Exner, 2003), but non-CS variables are included where relevant. In response to pressing concerns of most forensic psychologists when using the Rorschach, we address the recent criticisms of the Rorschach by synthesizing research findings. In doing so, we identify legitimate and spurious criticisms and describe and illuminate related limitations of the Rorschach. This entails our using the existing research literature and theory about the Rorschach to recommend certain alterations to interpretive practices and to identify important research needs.

CRITICISMS OF THE RORSCHACH FROM A HISTORICAL PERSPECTIVE

Before addressing psychometric issues, we present a brief historical perspective. Exner (1974) published the first edition of the Comprehensive System (CS), which was eventually recognized as being largely successful in meeting historical psychometric chal-

[1]For coverage of issues not included in the chapter, see Meyer and Archer (2001) and Viglione and Hilsenroth (2001).

lenges of reliability and validity. In the 1980s into the 1990s, the CS became the dominant system in teaching and practice (Hilsenroth & Handler, 1995; Mihura & Weinle, 2002) and it has become to be used extensively on an international basis (e.g., in Argentina, Belgium, Brazil, Denmark, Finland, France, Holland, Japan, Israel, Italy, Peru, Portugal, Sweden, and Spain). Exner's works are contained in three volumes with eight editions and in five editions of his workbook.

Since 1995, the Rorschach has once again been subjected to a series of repetitive critical reviews from a group of coauthors (e.g., Garb, 1999; Grove, Barden, Garb, & Lilienfeld, 2002; Hunsley & Bailey, 1999, 2001; Lilienfeld, Wood, & Garb, 2000; Nezworski & Wood, 1995; Wood & Lilienfeld, 1999; Wood, Nezworski, Garb, & Lilienfeld, 2001a; Wood, Nezworski, Garb, & Lilienfeld, 2001b; Wood, Nezworski, & Stejskal, 1996), although controversy has existed since its origin (e.g., Hirt, 1962; Murstein, 1965; Rabin, 1981; Viglione & Rivera, 2003). Some of these criticisms are written to challenge the Rorschach in court (e.g., Dawes, 1999; Grove & Barden, 1999; Grove et al, 2002; Lilienfeld et al., 2000; Wood et al., 1996). Criticisms and controversies have waxed and waned in the literature. A regular tension has emerged between practitioners using the Rorschach, many of whom find the Rorschach to be indispensable in their applied work, and some academic researchers who consider the Rorschach and its evidentiary foundation to be fundamentally flawed.

Atkinson, Quarrington, Alp, and Cyr (1986), after presenting results from one of the earliest meta-analytic reviews on Rorschach validity, questioned why its validity is continuously challenged despite the evidence. They asserted bluntly, "The oft-cited explanation is that deprecation of the Rorschach is a sociocultural, rather than scientific, phenomenon" (p. 244). Others have asked whether the debate about the utility of the Rorschach is more philosophical and political, rather than academic and scientific (Viglione & Rivera, 2003).

To a degree, these recent challenges of the Rorschach and the CS prompted the current book on forensic issues. Although the controversy is part political and philosophical debate and part scientific and rational debate, one goal is to focus on the latter. Nevertheless, because it is probably impossible to step outside of the former, we note that we consider ourselves political centrists when it comes to the Rorschach. That is, we believe the evidence supports its use in clinical practice, but we also believe that, like all tests, it has its limitations. Continued research is needed to specify the applications and limitation for many interpretive postulates. Like all tests, it needs to be used cautiously and conscientiously.

RELIABILITY: DO WE MEASURE CONSISTENTLY?

Reliability can be globally defined as the extent to which a construct is assessed consistently. Once we are measuring something consistently, it is necessary to establish that what is being measured is actually what we want to measure (validity) and that the measured information is helpful in some applied manner (utility). We focus on reliability first.

There are four main types of reliability—internal consistency, stability, alternate forms, and interrater. *Internal consistency reliability* refers to the consistency or homogeneity of content over items, that is, whether the items of a scale or test measure the same

construct. In the Rorschach, the notion of an item would have two meanings. First, responses or cards could be considered items. This form of internal consistency reliability entails an assumption that each card or response provides an equal opportunity to measure the same construct (Exner, Armbruster, & Viglione, 1978). However, it is readily recognized that each card does not allow an equal opportunity for all scores (e.g., cards vary greatly in their pull for color or texture determinants), so that internal consistency reliability is infrequently evaluated or reported and it is considered largely inapplicable to the test.

In terms of internal consistency reliability, an item also translates to the individual subcomponents or criteria of composite indices (e.g., the subcomponents of the *DEPI* or Ego Impairment Index, *EII*). As an example of this version of internal consistency research, Hilsenroth, Fowler, and Padawer (1998) and Stokes, Pogge, Grosso, and Zaccario (2001) examined the internal consistency of the six criteria forming the Schizophrenia Index (*SCZI*), whereas Dao and Prevatt (2006) examined the five criteria of its successor, the Perceptual Thinking Index (*PTI*). Although they found evidence for a reasonable degree of homogeneity (*KR—20* = .79, .70, and .75, respectively), these analyses are difficult to interpret because the six *SCZI* and five *PTI* criteria draw on just two types of scores, form quality and the cognitive special scores. As such, there should be a certain degree of artificial correlation among the criteria, although the precise magnitude would be hard to determine.

More substantively, the *SCZI* or *PTI* and all the other CS Constellation Indices were created as composites that draw on the full range of information available in a protocol to maximize validity; they were not developed as scales designed to measure a single homogeneous construct. As Streiner (2003) has pointed out, internal consistency reliability is important for scales assessing a homogenous construct but immaterial for a composite index. Indeed, efficiency in measurement is achieved through low rather than high intercorrelations among subcomponents or items. Accordingly, weak internal consistency reliability can accompany strong validity and utility.

Another type of reliability that has been largely considered inapplicable to the Rorschach is *alternate forms reliability*, which assesses the consistency of scores across parallel versions of an instrument. Although Holtzman specifically developed his set of inkblots to have two parallel forms and Behn-Eschenberg made an early effort at developing a set of inkblots to parallel Rorschach's inkblots (see, e.g., Exner, 2003, p. 12), at present a good parallel set of the 10 standard Rorschach inkblots does not exist.

Stability reliability, also known as temporal consistency or test–retest reliability, is essentially the consistency of scores over time. It has been applied to the Rorschach and the results generally have been acceptable to good (Grønnerød, 2003; Meyer & Archer, 2001; Viglione & Hilsenroth, 2001). Comprehensive System scores thought to measure traitlike aspects of personality have produced relatively high retest coefficients, even over extended time periods. Also, scores thought to reflect statelike emotional process have produced relatively low retest coefficients even over short time intervals.

However, the most recent large-scale and well-designed study of CS stability found lower than anticipated consistency over a 3-month retest period (Sultan, Andronikof, Réveilòre, & Lemmel, 2006). For instance, stability coefficients for *R* and *Lambda*,

which index the overall richness or complexity of a protocol, were .75 and .72, respectively. Because these scores are related to the frequency of other scores in the protocol, when they are unstable most other scores will be unstable as well. Indeed, in this study the median level of stability reliability across a core set of 47 scores was .53 and the median across 87 ratios, percentages, and derivations in the lower portion of the Structural Summary was .55. Number of responses (*R*) and *Lambda*, as markers of task engagement, moderated stability. Stability reliability was greater among those individuals whose *R* and *Lambda* did not change much over time, as compared to the stability among those individuals whose *R* and *Lambda* differed at the two testings.

Conducted in France, the Sultan et al. study was a carefully executed investigation with a sound methodology and adequate controls. It also used the most sophisticated statistical analyses to date to examine potential moderators of stability, and several were identified that would increase stability if they were controlled (e.g., engagement with the Rorschach task, situational distress/emotional status). Variation over time due to situational distress or emotional status is not related to the true stability reliability of the test, so that test–retest statistics underestimate the Rorschach's true reliability. Nevertheless, even taking this situational variation into consideration, the stability for the majority of the Rorschach CS variables in this study was limited.

More investigation of Rorschach stability reliability is needed (Meyer & Archer, 2001; Viglione & Hilsenroth, 2001), and Sultan et al.'s (2006) findings should be replicated. However, given the care that went into designing and executing this study, forensic examiners should be aware of the challenges to the CS that might emerge in the courtroom from these data. The Sultan data indicate that nonpatient volunteers for a study can provide notably different protocols when tested by one reasonably trained examiner and again 3 months later by a different reasonably trained examiner. This finding will remain even if it is subsequently discovered that certain methodological factors account for the lower than expected stability or if the majority of future studies find superior stability.

Putting these results in context might be illuminating. Forensic examiners should recognize that the global stability of Rorschach scores might, under some circumstances, be more similar to the stability of memory tests than the stability of intelligence tests. For instance, although the manual for the third edition of the Wechsler Memory scale (WMS: Psychological Corporation, 1997) does not report data for all subscales, the 1-month stability for 13 of its subscores is .71 ($N = 297$). Over a 7 ½-month retest interval, the average stability coefficient for 5 of its subscores was .66 (Dikmen, Heaton, Grant, & Temkin, 1999) and over a 9-month interval the average stability for 10 scores was .68 (Martin et al., 2002). Although these coefficients are higher than those observed in Sultan et al., more similar stability values are found for tests like the California Verbal Learning Test (CVLT) and the Hopkins Verbal Learning Test (HVLT). Over a retest interval of 1 to 2 months, the average stability of HVLT scores was about .50 (Barr, 2003; Benedict, Schretlen, Groninger, & Brandt, 1998). Average stability for CVLT scores also has been about .50 over a 1-year retest interval (Paolo, Tröster, & Ryan, 1997). Finally, as another example, the average stability of scores on the Extended Complex Figure Test was .46 over the course of a 1-week interval ($N = 55$; Woodrome & Fastenau, 2005). It should be

pointed out that memory ability is thought to be a stable trait similar to many personality and information-processing variables accessed by the Rorschach and, as such, should possess stability reliability.

Forensic examiners addressing work-related issues might also note that the Sultan et al. (2006) findings are similar to the stability of job performance measures. In a recent meta-analysis, Sturman, Cheramie, and Cashen (2005) found that over a 6-month retest interval, the temporal consistency of objective job performance measures was .45. For both objective and subjective measures of job performance, consistency was .56.

In a summary of the research data available at the time, Viglione and Hilsenroth (2001) reported that CS stability was adequate or better in all respects, especially in the context of comparing Rorschach findings to other personality tests. Revisiting the data about other tests leads to the conclusion that the level of stability reported by Sultan is similar to that reported for the MMPI in a meta-analysis over a 1-year period (Mauger, 1972; Stone, 1965; Sines, Silver, & Lucero, 1961; all as cited in Dahlstrom, Welsh, & Dahlstrom, 1975; Milott, Lira, & Miller, 1977; Ryan, Dunn, & Paolo, 1995). The Sultan stability reliability coefficients are also similar to those reported in a comprehensive meta-analysis of self-report, observer, and performance tests of personality (Roberts & Del Vecchio, 2000), but less than that reported in a more limited and less definitive meta-analysis of eight self-report tests over a 1-year period (Schuerger, Zarrella, & Holtz, 1989). At this point, forensic examiners should be alert to the possibility, based on this one study, that CS scores can be more changeable and responsive to statelike influences than previously thought. In forensic cases, when making dispositional attributions, examiners might consider repeating a Rorschach and other personality assessment measures to more definitively differentiate state and trait influences.

The type of reliability that has received the most attention recently—and one that may be most relevant to forensic practice—is *interrater reliability*, or the consistency of judgments across raters. For the Rorschach, this type of reliability concerns coding (scoring) reliability as well as the reliability of interpretation across test users. We address research in coding reliability because it has received most of the recent research attention. For issues involving interpretive reliability, we refer to Meyer, Mihura, and Smith (2005).

Exner (2003) has primarily presented percentage agreement (*%A*) between coders as a means of addressing interrater reliability and coding accuracy. Percentage agreement is the proportion of responses in which two raters agree on a code, that is, code a given response parameter the same way. He had required that any code have a *%A* of 80% to be included in the CS. Weiner (1991) also required that studies submitted to the *Journal of Personality Assessment* meet this *%A* benchmark for 20 records. For example, for *Human Movement* (*M*), if two raters independently code 50 responses and agree 45 times on the presence or absence of *M*, then *%A* = 90%. However, *M* only occurs in about one fifth of responses from adults, so that two raters are expected to agree, by chance, about 70% of the time. This high incidence of chance agreement occurs largely because raters with knowledge of base rates could agree that *M* is absent even if they randomly scored *M*. Accordingly, *%A* does not consider base rates and chance agreement, and it overestimates reliability for single scores, so that it has been subjected to criticism (Wood et al., 1996).

Although true in some respects, criticism of *%A* has been greatly overextended to all types of coding and response combinations. It is not nearly as problematic for response segments that have multiple choices for codes. The term *response segment* refers to a coding category, for example, determinants or content. To achieve agreement for determinants, one would have to agree on all determinants in a given response (e.g., *FT.CF* and *FT.CF* represents an agreement, whereas *FT.CF* and *FT.FC* do not.) Obviously, chance agreement for response segments (e.g., determinants or content) is much lower than it is for individual codes. For determinants and content chance *%A* is about 20%; for all special scores chance %A is about 40%; for location, DQ, and FQ chance %A is about 30%–50% (Meyer, 1997a, 1997c). Thus, *it is mathematically impossible to discount 80%* agreement for response segments among 20 records as being due to chance.

Nevertheless, there are preferred statistics that do take base rate into consideration, namely, kappa for response level data and the intraclass correlation (ICC) for protocol level data. Kappa is appropriate for nominal or categorical variables, as represented by individual Rorschach scores or codes. Accordingly, if one wanted to evaluate how reliably two raters or two teams of raters scored *Texture* (*T*) on a response by response basis, one could use kappa. This statistic could, for example, estimate reliability for the presence or absence of any form of *T*. Alternatively, it could detect whether or not raters reliably distinguished between *FT*, *TF*, *T* and no Texture.

Whereas kappa is applied to response-level variables, the ICC is applied to dimensional variables at the protocol level. In other words, if one wanted to evaluate the reliability of the sum of all *T* responses, *X* – %, or the Suicide Constellation across records, ICC is ideal. Score levels and interpretation of ICC are equivalent to kappa, and it is an excellent statistic for Rorschach summary scores (i.e., those types of scores that are found on the CS Structural Summary). Given that the preponderance of interpretive inferences emerges from the Structural Summary, the ICC is more related to the foundation of interpretation and how the test is used in practice. Kappa, however, may be more useful in training raters and evaluating the ease to which a new score can be coded.

Janson (Janson & Olsson, 2001, 2004) has introduced a new statistic called iota. As a more general statistic, it can be used instead of kappa or ICC. It is a multivariable extension of kappa and can be applied to response level variables (e.g., individual codes), response segments (e.g., determinants or contents in a given response), or even all the codes of a response or protocol in its entirety. Like the ICC, it also can also be applied to dimensional or protocol level variables. Accordingly, it has considerable flexibility and is recommended for research and training. For training or forensic applications, one could measure the reliability or agreement of two raters for a single record across all scores.

Given that kappa, ICC, and iota are more demanding types of reliability statistics, the benchmarks for interpreting their magnitude differ from those associated with Pearson *r* and *%A*. Kappa, ICC, and iota at .75 or above is considered excellent, .60 and above good, and .40 and above fair (Cicchetti, 1994; Shrout & Fliess, 1979).

There are four meta-analyses addressing Rorschach interrater reliability. Two related studies address CS reliability (Meyer, 1997a, 1997c; Meyer et al., 2002) and the others address two non-CS scales, the Rorschach Prognostic Rating scale and the Rorschach Oral Dependency scale (see Meyer, 2004). Meyer (2004) compared these interrater reli-

ability data to all
Comparisons with
deed an attorney,
Rorschach fares.

These interrate
presented separate
of judgment, the a
ratings summarize
mary scores, wher
ual responses. A
because random er
scales. The overall
are excellent with
the range between
high, with r = .84 f

Thus, one must
and compares favo
cine. Attorneys, juc
agree much more th
agnoses of breast a
care from record re
sistency across rate
have the same degr
medicine. For exam
nal canal and spina
missing teeth in ear
that Rorschach cod
process, one in whi

Clearly, the ans
and motivated rate
ported that standar
their reliability esti
2000; Meyer, 1997
rate variables, for e
be loosely defined

[2]Meyer (2004) compar
both types of statistics, the
the findings for those 16
Meyer's (2004) in that it p
separately conducted inte

[3]Weiner (2001) descri
M–, *WS*+, or *M*^{a} with *Pure*
fined variables are more li
variables, presumably bec

ability data to all the other meta-analyses of interrater reliability available at the time. Comparisons with these other types of judgments allow forensic psychologists—or indeed an attorney, judge, or jury—to derive a "gut feel" sense of how the reliability of the Rorschach fares.

These interrater reliability comparisons are presented in Table 2–1.[2] Reliabilities are presented separately for scale-level judgments and item-level judgments. With each type of judgment, the average reliability coefficient is listed along with the number of pairs of ratings summarized. For the Rorschach, scale data corresponds to protocol level summary scores, whereas item data corresponds to coding determinations made on individual responses. A consistent pattern is that scale reliabilities exceed item reliabilities because random errors tend to cancel each other out when items are aggregated to form scales. The overall reliability of the Rorschach CS and Rorschach Oral Dependent scale are excellent with summary score coefficients about .90 and response-level judgments in the range between .80 and .85. The Rorschach Prognostic Rating scale reliability is not as high, with $r = .84$ for summary scores, but still more than adequate.

Thus, one must conclude that the Rorschach interrater reliability is good to excellent and compares favorably to a wide range of determinations made in psychology and medicine. Attorneys, judges, or juries may be very interested to know that the Rorschach raters agree much more than do superiors' evaluations of job performance, surgeons/nurses' diagnoses of breast abnormalities, and physicians' estimations of the quality of medical care from record review, all of which are subject to considerable disagreement and inconsistency across raters. Rorschach CS and Oral Dependent scale coding determinations have the same degree of agreement or reliability as do simple, physical measurements in medicine. For example, Rorschach coding is as reliable as estimating the size of the spinal canal and spinal cord from MRI, CT, or x-ray scans, or counts of decayed, filled, or missing teeth in early childhood. These comparisons are consistent with the conclusion that Rorschach coding for the trained examiner is typically a relatively straightforward process, one in which consistency and agreement are attainable across raters.

Clearly, the answer to the question, "Do we code reliably?" is yes, as well-trained and motivated raters code reliably. However, there are limitations. Several studies reported that standard errors of reliabilities of low base rate variables are large so that their reliability estimates are erratic (Acklin, 1999; Acklin, McDowell, & Verschell, 2000; Meyer, 1997a, 1997c; Meyer et al., 2002; Viglione & Taylor, 2001). Low base rate variables, for example, sex, reflections, color projection, or refined variables[3] can be loosely defined as occurring on the average once or less often per record. This is a

[2]Meyer (2004) compared types of statistics, contrasting *r* with kappa or the ICC. Across 16 topics that provided both types of statistics, the mean kappa/ICC was .70 and the mean *r* was .74. Because these differences are not large, the findings for those 16 topics were combined in our version of the table. Our table also differs slightly from Meyer's (2004) in that it presents two coefficients for job selection interviews (one for joint interviews and one for separately conducted interviews), rather than just a single undifferentiated coefficient.

[3]Weiner (2001) described refined variables as coding combinations that encompass multiple categories, so that *M*–, *WS*+, or *M*^{a} with *Pure H* are refined variables. In contrast, *M*, *W*, and *H* are unrefined variables. He stated that refined variables are more likely to demonstrate validity in research. There is not a great deal of research with refined variables, presumably because large samples are needed.

TABLE 2–1

Meta-Analyses of Interrater Reliability in the Psychological and Medical Literature

		n(k–1) = independent pairs of judgments		*Reliability r/κ/ICC*	
Target reliability construct		*Scale*	*Item*	*Scale*	*Item*
1. Measured bladder volume by real-time ultrasound			360		.92[h]
2. Measured size of spinal canal and spinal cord on MRI, CT, or X-ray		200	86	.90[a]	.88[a]
3. Count of decayed, filled, or missing teeth (or surfaces) in young children		113	237	.97[a]	.79[i]
4. Rorschach Oral Dependency Scale scoring		974	6,430	.91[b]	.84[c]
5. Scoring the Rorschach Comprehensive System:	Summary scores	784		.91[b]	
	Response segments		11,518		.86[c]
	Scores per response		11,572		.83[c]
6. Neuropsychologists' test-based judgments of cognitive impairment			901		.80[d]
7. Hamilton Depression Rating Scale scoring from joint interviews[d]		3,847	495	.86[h]	.71[b]
8. Level of drug sedation by ICU physicians or nurses		1,116	165	.86[b]	.71[c]
9. Functional independence measure scoring (joint and separate interviews)		1,365	1,345	.91[c]	.62[c]
10. TAT Personal Problem-Solving Scale scoring		385		.85[b]	
11. Rorschach Prognostic Rating Scale scoring		472		.84[j]	
12. TAT Social Cognition and Object Relations Scale scoring		934		.82[b]	
13. TAT Defense Mechanism Manual scoring		743		.80[b]	
14. Hamilton Anxiety Rating Scale scoring from joint interviews[d]		752	214	.80[b]	.72[c]
15. Borderline personality disorder (joint and separate interviews)	Diagnosis	402		.82[c]	
	Specific symptoms		198		.64[c]
16. Signs and symptoms of temporomandibular disorder (separate exams)		192	562	.86[c]	.56[c]
17. Hamilton Depression Rating Scale scoring from separate interviews		1,012	597	.82[b]	.52[b]
18. Therapist or observer ratings of therapeutic alliance in treatment		(*S* = 31)		.78[a]	
19. Job selection ratings by joint interviews		9,364		.77[a]	
20. Hamilton Anxiety Rating Scale scoring from separate interviews		268	208	.76[b]	.58[c]
21. Axis I psychiatric diagnosis by SCID in joint interviews		216		.75[c]	
22. Type A behavior pattern by structured interview		(*S* = 3)		.74[a]	
23. Axis II psychiatric diagnosis by semistructured joint interviews		740		.73[c]	
24. Personality or temperament of mammals (variable observations)		151	637	.71[a]	.49[a]

Target reliability construct		*n(k–1) = independent pairs of judgments*		*Reliability r/κ/ICC*	
		Scale	*Item*	*Scale*	*Item*
25. Visual analysis of plotted behavior change in single-case research			1,277		.57[b]
26. Editors' ratings of the quality of manuscript reviews or reviewers			3,721		.54[b]
27. Presence of clubbing in fingers or toes[e]			630		.52[c]
28. Stroke classification by neurologists			1,362		.51[c]
29. Child or adolescent problems:	Teacher ratings	2,100		.64[a]	
	Parent ratings	4,666		.59[a]	
	Externalizing	7,710		.60[a]	
	Internalizing	5,178		.54[a]	
	Direct observers	231		.57[a]	
	Clinicians	729		.54[a]	
30. Job performance ratings by supervisors		1,603	10,119	.57[a]	.48[a]
31. Axis I psychiatric diagnosis by SCID in separate interviews		693		.56[c][f]	
32. Job selection ratings by separate interviews		3,185		.53[a]	
33. Axis II Psychiatric diagnosis by semistructured separate interviews		358		.52[c]	
34. Self and partner ratings of conflict:	Men's aggression	616		.55[a]	
	Women's aggression	616		.51[a]	
35. Determination of systolic heart murmur by cardiologists			500		.45[c]
36. Abnormalities on clinical breast examination by surgeons or nurses			1,720		.42[c]
37. Mean quality scores from two grant panels:	Dimensional ratings		2,467		.43[b]
	Yes/No decision		398		.39[c]
38. Job performance ratings by peers		1,215	6,049	.43[a]	.37[a]
39. Number of factors in a correlation matrix by scree plots[f]			2,300		.35[c]
40. Medical quality of care as determined by physician peers			9,841		.31[c]
41. Job performance ratings by subordinates		533	4,500	.29[a]	.31[a]
42. Definitions of invasive fungal infection in the research literature			21,653		.25[c]
43. Research quality by peer-reviewers:	Dimensional ratings		31,068	.25[b]	
	Yes/No decision		4,807		.21[c]

Note. Adapted from Meyer (2004), which provides a complete description of the meta-analytic data sources contributing to this table. ICC = intraclass correlation, ICU = intensive care unit, S = number of studies contributing data, SCID = Structured Clinical Interview for the *Diagnostic and Statistical Manual of Mental Disorders* (DSM), and TAT = Thematic Apperception Test.

[a]Pearson's *r*. [b]Combination of *r* and κ or agreement ICC. [c]κ or agreement ICC. [d]Category includes videotaped interviews and instances when the patient's report fully determined both sets of ratings (e.g., identical questions in written and oral format). [e]One study produced outlier results (κ = .90) relative to the others (κ range from .36–.45) so the results should be considered tentative. [f]Finding should be treated cautiously because agreement varied widely across studies, with values below .10 in several samples but above .70 in several others.

statistical issue and one would need large samples to accurately estimate reliability for low base rate variables.

In addition, there are some codes for which reliabilities are lower so that they are presumably more of a challenge to code accurately. Table 2–2 identifies these CS codes associated with lower reliabilities in multiple research reports. Forensic examiners should pay special care to code these variables accurately, consistent with CS principles. Some examiners have protocols in high-stakes cases blindly rescored by a colleague. Viglione wrote *Rorschach Coding Solutions* (2002) to address these and other coding challenges. Along with the workbook (Exner et al., 2001) and volume I text (Exner, 2003), it is a good resource to consult to eliminate rater drift from CS standards. Indeed, interrater reliability is not a fixed property of the score or instrument. In forensic practice, this means that what counts is the reliability of the person who coded the protocol, not the general reliability found in the literature. As such, it would behoove forensic examiners to document that they have achieved good interrater reliability with another expert rater.

In the forensic arena, the single most problematic implication of the data on variables with lower reliabilities might be the possibility of over coding *ALOG*, *DR*, and *FQ–* so as to overestimate pathology, thought disorder, and the likelihood of a psychotic or schizo-

TABLE 2–2

CS Codes Decisions with Lower Reliabilities in Some Studies

Developmental Quality
DQv and *DQv/+*
Form Dominance
FC vs. *CF* vs. *C*
Form Shading vs. *Shading Form* vs. *Shading*
Shading Subtypes
Y vs. *T* vs. *C'* vs. *V*
Form Quality
Occasionally *FQ* subcategories, especially *FQu*
Failure to code or neglect of *FQ+*
Contents
Art, Ay, Sc, Bt vs. *Na* vs. *Ls, Id*
Special Scores
DV vs. *INC*
ALOG vs. no special score, coding too many *ALOG*s
CONTAM vs. *INC*
PER or *DR* vs. task comment, coding too many *DR*s
Level 1 vs. *Level 2*

phrenic diagnosis. In forensic assessment, such an error might translate to underestimates of, for example, sanity, capacity, culpability, or parenting ability. Some comfort can be drawn by the fact that the research indicates that the summary scores for cognitive special scores, *WSum6* and *Sum6*, generally demonstrate better reliability than do the individual scores (e.g., *DV2* or *ALOG* individually). This superior reliability is important because interpretation is primarily based on these summary scores rather than on individual cognitive special score codes.

Research reports from around the world (Erdberg, 2005; Viglione, 1999; Viglione & Hilsenroth, 2001) also reveal that the CS is transportable to other languages and cultures and that coding reliability is very similar to the results from the meta-analyses. For the most part, those codes that achieve lower or more variable reliabilities in U.S. samples are the same as codes that are more variable in the international samples (Exner et al., 1999).

Another issue or complication is that most reliability research studies generally use raters who work or train in the same setting. If 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. Thus, existing reliability research may then give an overly optimistic view of reliability across sites or across forensic examiners working independently.

In a preliminary presentation, Meyer, Viglione, Erdberg, Exner, and Shaffer (2004) examined this across site interreliability issue by having 40 randomly selected protocols from Exner's new CS nonpatient reference group sample and 40 protocols from Shaffer, Erdberg, and Haroian's (1999) from a California (CA) sample recoded by a third group of trained raters. This third group, advanced graduate students supervised in Viglione's lab, were blind to the original coding, the origin of the samples, and the nature and purpose of the study. The coding assigned by the original sites was compared to the coding assigned by this single additional site and yielded an across site median ICC of .72, an acceptable level of reliability in the good range.

These across site results can be contrasted with within site data sets, that is, samples coded by raters working in the same setting. We have three such relevant within site research reports available to us: (a) the meta-analysis data in Table 2–1, (b) a large international sample (Erdberg, 2005), and (c) a smaller sample from Viglione's lab. All report greater reliabilities than our across site median ICC of .72. As noted earlier, the Table 2–1 meta-analysis yields a reliability estimate for summary scores of .91. Erdberg (2005) compiled 467 protocols from 17 internationally collected nonpatient reference samples. The initial median within site ICC from the international sample was .82, a reliability estimate in the excellent range. Although the pool of protocols was collected from many different countries, all the scoring for each protocol took place locally by examiners who trained together. Thus, these data provide a reasonable sample of within site scoring reliability across the world and attest to the cultural adaptability of the test and its administration procedures. The third within site reliability estimate is pertinent because it is from the same lab that provided the across site coding. Viglione and Taylor (2001) reported a median within site reliability of .92 for 84 protocols.

Although the across site reliability estimates are preliminary, these findings suggest that there are complexities in the coding process that are not fully clarified in the standard

CS training materials (Exner 2003; Exner et al., 2001). As a result, training sites (e.g., specific graduate programs) may develop guidelines for coding that help resolve these residual complexities but they may not generalize well to other training sites. Forensic examiners may find it helpful to consult an advanced coding text (Viglione, 2003) or to practice coding with colleagues trained in a different setting.

NORMATIVE DATA: HOW ADEQUATE ARE CS NORMS?

Rorschach normative reference group data have been criticized for pathologizing examinees. Wood et al. (2001b) compared CS reference values on 14 selected variables to the values reported in 8 to 19 comparison samples from the literature. They reported small to very large differences (Cohen's *d* from .18 to 1.67)[4] for the 13 variables where mean differences could be computed.[5] All differences were in the more pathological or problematic direction for the comparison samples. There were nine variables for which these differences were at least medium size: (a) lower values for *X+%*, *Afr*, *FC*, *P*, *WSumC*, and *Pure H*; and (b) higher values for reflections, *X–%* and *Y*. Variability of these scores (i.e., the *SD*) was greater than in the original CS sample—a worrisome finding because it might suggest that current confidence intervals and normative interpretive ranges are too narrow.

The samples in the Wood et al. report were portrayed as nonpatient or normative reference samples but had serious problems and were not fully representative of nonpatients (Meyer, 2001). From a total pool of 32 studies, 22 samples (69%) did not have a procedure to exclude patients or low functioning or disturbed individuals; 16 (50%) samples were college students or the elderly; one had a mean *R* of 15, whereas another had a mean *R* of 39, suggesting atypical administration; respondents in one sample were held motionless with electrodes on their head; and just two samples had data for all 14 scores. Obviously, these samples are not representative of nonpatients and are not a good source for comparisons. Nevertheless, it is hard to dismiss these findings totally, as others (Viglione & Hilsenroth, 2001) have examined similar data and found that the distributions for form quality and *R* appeared to diverge to some degree from CS expectations.

To investigate these normative issues with a better comparison sample, Meyer (2001) contrasted Exner et al.'s (1993) original CS adult normative reference sample to a composite of 2,125 protocols from nine adult samples presented in Erdberg and Shaffer's (1999) symposium on international CS reference data. These samples (which include the Shaffer et al., 1999, sample from the United States) provided data on all CS variables and encompassed great variability and thus generalizability across subject selection procedures, examiner training, examination context, language, culture, and national boundaries. Across 69 composite scores from the lower portion of the Structural Summary, distributions for 49 variables were similar in the original CS sample and international

[4]Cohen's *d* is an effect size measure for comparing two groups. It basically is the difference between the means of the groups in standard deviation units, i.e. the z-score for the differences. For example, a difference of 10 IQ points should result in a Cohen's *d* of 0.67.

[5]The 14th variable was EB style, a categorical variable for which means could not be computed.

data, a finding consistent with the conclusion that the original CS norms are generally adequate. These data, in addition to the similarities between U.S. and international findings for interrater reliability, again indicate considerable cultural and international adaptability of the Rorschach. One can adapt it to different cultures, languages, and regions, and the test behaves largely as it does in the United States.

Nevertheless, some differences between the CS sample and the composite of international samples persist, so that we need to adjust our normative expectations. International samples have higher scores for *Dd*, *S*, *FQu*, *FQ–*, *Hd*, *(Hd)*, and *Sum6*, and lower scores for *WSumC*, *EA*, *FQo*, *P*, *COP*, *AG*, and *Afr*. In all cases, the CS norms come across as "healthier." In other words comparison to the CS norms would lead to more pathological interpretations than would comparisons to the international norms. Accordingly, normative expectations for these and for variables that subsume them (e.g., *X–%* for *FQ–*) need to be adjusted. More specific recommendations are given here.

A reasonable question becomes, "Why do the original CS norms look healthier than other normative approximation samples?" The CS respondents were recruited largely through work, unions, or social organizations. Compensation was in the altruistic form of contributions to charity in name of the place of business or organizations, so that respondents were not paid themselves as volunteers. Thus, differences could be due to situational differences or examination context. The CS respondents may feel that their responses matter more than do volunteers in other studies, so that they may "tidy-up" their answers a bit more through filtering in the response process (Exner, 2003). One might speculate that making the examination matter to the respondent is a better approximation of the use of the test in the real world, and thus a better contrast sample. Alternatively, these recruiting practices involving employment and social involvement might lead to a selection bias in terms of attracting healthier and better adapted individuals to volunteer. Indeed, the literature indicates that the garden variety volunteers tend to possess problematic characteristics and are less well-adapted (Berman, Fallon, & Coccaro, 1998; Rosenthal & Rosnow, 1975).

Other explanations of the observed health in the CS norms include differences in administration or coding. There are considerable differences between the initial CS form quality tables first published in 1974 (Exner, 1974) and the current version (Exner et al., 2001), with most of these differences resulting in more *FQ–* and fewer *FQo* responses (Meyer & Richardson, 2001; Viglione, 1989). In addition, criteria and examples for other coding distinctions have changed or been elaborated on over time in ways that alter the benchmarks for assigning a score (Meyer, 2001). Another explanation is simple aging of the norms and increasing mental health difficulties over time.

To address these normative issues, Exner started collecting a new adult normative reference group in 1999 (Exner, 2002; Exner & Erdberg, 2005). This new sample, which is now approaching 500 respondents, was collected largely in the same way as the original CS sample, but there are some differences. The new sample involves the workplace or organizations less formally, so that individuals may feel that they represent themselves instead of an organization. For example, charity donations are made in a respondent's name rather than the organization's name. In the original CS sample, a manager acted as the liaison between examiners and data collection sites and actually solicited respondents. In

the new sample, examiners recruit participants on their own. Respondents are now excluded due to "prolonged or significant history" of psychotropic medications or illegal drug use.

Exner and Erdberg (2005) provided data for 450 of the individuals in this sample. The more important differences in terms of mean differences and interpretive cutoffs between the two groups are summarized in Table 2–3. The selected frequencies differ by 5% in the two samples. As can be seen, form quality is less optimal in the new reference sample. There are fewer *Populars*, more special scores, and more of the serious *Level 2 Cognitive Special Scores*. There is less color overall and more color-dominated relative to form-dominated color responses. The *Afr* is lower, there is a notable increase in space responses, and there is a lower frequency of both cooperative and aggressive movement scores. In addition, it is more common for passive movement to exceed active and for the *Depression Index* (*DEPI*) to be elevated. Although the frequencies remain low, it is worth pointing out that the *SCON* did not exceed 7 in any of the old 600 records, but it does for 11 of the current 450 records. These changes incorporate many of the same variables discussed earlier as divergences between the old CS samples and the international composite pool of references samples collected by other researchers.

Another notable finding is that the standard deviation for *R* is 5.68, as compared to 4.40 in the original CS sample of 600. This change may be problematic because this increased variability of *R* should be associated with more variability for all other scores. Indeed, the great majority of *SD*s is larger in the new sample as compared to the original. This greater variability means that interpretive postulates need to have wider confidence intervals (i.e., the range of expected scores is broader).

Although the new CS reference sample reduces some of the differences with the composite of international reference samples, it does not eliminate them. For example, the new CS sample still has means for *Dd* and *X–%* that are lower and means for *X+%* and *EA* that are higher than other reference samples.

The study that initiated the concerns about the original CS normative reference sample is mentioned in the previous reliability discussion and was published by Shaffer, Erdberg, and Haroian (1999). Its respondents had MMPI–2 *T* score means at approximately 50 and WAIS–R IQs of about 100, thus at normative values. Most Rorschach values were consistent with the original CS normative reference group, but values for the variables already identified as diverging from normative expectations also demonstrated such divergence in this sample. The Shaffer et al. California (CA) sample also differed from both the original and new CS samples in terms of overall complexity. The mean for *R* in the Shaffer et al. sample is only 20.8 versus 23.36 for the new CS sample, and the *Lambda* is 1.22 (median = .75) versus .58 (median = .47) in the new CS sample, with 41% of the Shaffer et al. sample having a *Lambda* greater than .99 versus 14% in the new CS sample. These findings indicate that the Shaffer et al. sample was not very productive and they produced relatively simplistic records in comparison to the CS and other samples included in the international group.

Along with our interrater reliability investigations with these samples (Meyer et al., 2004), we have conducted some initial investigations into the differences between the CA normative reference sample and the new CS reference sample. In this research, we ex-

Illustrative C

Domain/Score
Quality of Percep
X+%
Xu%
X–%
X+% < .55
X% > .20
X–% > .20
XA% > .89
WDA% < .85
P > 7
Sum6
WSum6
Lvl2 SS > 0
Color
FC > CF +C
FC > CF +C
CF + C > FC
CF + C > FC
Extratensive
Miscellaneous
S > 2
DQv > 2
T > 1
Ego < .33
Ego > .44
Afr < .40
Afr < .50
Zd < 3.0
Intell > 5
COP = 0
AG = 0
Hd
(Hd)
DEPI > 4
p > a + 1
Mp > Ma

TABLE 2–3

Illustrative Changes in the New Target Reliability Construct Versus Original CS Normative Reference Samples

Domain/Score	*Original 600*	*New 450*
Quality of Perception and Thinking		
X+%	.77	.68
Xu%	.15	.20
X–%	.07	.11
X+% < .55	2%	12%
X% > .20	22%	45%
X–% > .20	3%	10%
XA% > .89	74%	45%
WDA% < .85	5%	16%
P > 7	31%	18%
Sum6	1.91	2.54
WSum6	4.48	7.12
Lvl2 SS > 0	6%	13%
Color		
FC > CF +C + 2	25%	15%
FC > CF +C + 1	41%	26%
CF + C > FC + 1	12%	26%
CF + C > FC + 2	4%	14%
Extratensive	38%	31%
Miscellaneous		
S > 2	14%	38%
DQv > 2	12%	2%
T > 1	11%	17%
Ego < .33	13%	20%
Ego > .44	23%	30%
Afr < .40	3%	9%
Afr < .50	11%	24%
Zd < 3.0	7%	14%
Intell > 5	2%	8%
COP = 0	17%	11%
AG = 0	37%	44%
Hd	.84	1.14
(Hd)	.21	.62
DEPI > 4	5%	14%
p > a + 1	2%	10%
Mp > Ma	14%	23%

amined whether coding conventions might contribute to the differences between the data sets. More specifically, we wondered if CS–CA differences would be reduced when records from both samples were recoded at a third site. If the Shaffer et al. records were coded according to somewhat different benchmarks than Exner's protocols, the differences between the two samples would be reduced if records from both samples were coded by a third group.

To address this question and as described earlier, we obtained 80 protocols from both the CA and CS samples. These 80 protocols were then recoded by a new group of examiners who were trained together in one setting. We then computed two sets of difference scores, using Cohen's *d* as the effect size index. The first difference score compared mean scores for the CS and CA samples using the original coding from the two sites. The second difference score compared the means for the CS and CA samples based on the new coding. Because the new coding was done by raters who trained together within one site, it eliminates the potential influence of site-specific differences in coding conventions. We anticipated that the initial differences would decrease with the revised coding; that is, the second set of differences from single site scores would be smaller than the first set of differences generated from separate sites.

Initially, with the original CS and CA scoring, across 129 structural summary variables the differences for 36 scores (28%) were moderate to large, with *d* values greater than .40 or less than –.40. Thus, the normative expectations differed for 36 of the 129 variables in our randomly selected protocols from both samples. However, with the new single site coding, there were only three means (2%) that remained different at this magnitude. Thus, almost all the seemingly important differences between the new CS sample and the CA sample disappeared when the protocols were rescored by a different group. In general, for most variables, our new coding split the difference between the CS sample and Shaffer et al. sample. By and large, the groups now were much more similar: Relative to the original scores, with the new coding, the CS sample looked less healthy than before and the Shaffer et al. sample looked healthier than before.

However, there were instances when the new scores were more similar to one of the reference samples than the other. For complexity variables (*Lambda*, *DQ+*, *Blends*, etc.) and for *Dd*, the values from the rescored protocols more closely resembled the CS reference sample than the CA sample. Furthermore, with the possible exception of *Dd*, the CS reference sample is more similar than the CA sample to the internationally collected reference samples for these particular complexity scores. In contrast, form quality values from the rescored protocols were more in line with the Shaffer et al. CA sample than the CS sample. Equally important, the CA reference sample is more similar than Exner's CS sample to the form quality values observed in other U.S. and international reference samples.

The overall findings suggest that site-specific coding practices may contribute in important and previously unappreciated ways to some of the seeming differences across normative approximation samples. In addition, these initial data suggest a convergence between the CS and CA sample, with the international normative sample. These suggestions are hypotheses that need to be tested with additional samples and coding sites.

There is less research into the suitability of the CS normative reference samples for children. In a study similar to the Shaffer et al. (1999) study and from the same group of researchers, Hamel, Shaffer, and Erdberg (2000) reported on 100 6- to 12-year-old children. This research has also attracted a lot of attention. To establish this group as a normative reference sample, their parents identified them as average to psychologically healthy on a commonly used multidimensional rating scale. However, once again, the Rorschach data diverged from the CS normative reference groups in some respects. In many ways, the differences are similar to those found in the adult normative reference samples. Like the adult samples, Hamel et al. found more distorted form quality values, less color, more use of unusual blot locations, elevated rates of dysfunction on the constellation indices, and less complexity. However, unlike the adult CA versus CS sample differences, the reference values observed by Hamel et al. tended to be more extreme. For instance, the average *Dd* was 8.3, the average *X–%* was .41, 62% of the sample had an elevated *SCZI* (value of 4 or more), and the median *Lambda* value was 1.14 (mean = 1.91).[6]

Although Hamel et al. (2000) took a careful and conscientious approach to their study, several characteristics of the sample suggest it is idiosyncratic and challenge its trustworthiness as a contemporary CS reference sample for children. First, all administration and coding was done by a single examiner, so that generalizability may be limited. Second, for interrater reliability, *%A* was reported in an unusual way.[7] This method would lead to the undetected possibility of coding inaccuracies for determinants, contents, and special scores. Also, in comparison to most research reports, *%A* was low for location and form quality. Third, the authors strongly emphasized the necessity for precision in documenting blot areas on the location sheet that appear to drift from CS standards:

> Students should be clearly taught to very carefully and accurately encircle the precise portion of the blot utilized by the examinee ... to enable any other clinician to precisely replicate the coding for location. The precision of location cannot be overemphasized; not only does the location code clearly depend upon an accurate location sheet, but so do other segments of the coding. Form quality and Popular are heavily dependent upon location. A Form Quality of ordinary can easily be altered to unusual or minus on the basis of location alone. (Hamel et al., 2000, p. 291)

If carried through in administration, this emphasis on precision may distort the interaction between the examiner and respondent in the inquiry and also influence the documentation of response areas on the location sheet. Moreover, along with the slack in interrater reliability, it may be related to the extraordinary *Dd* elevation. Excessive *Dd* locations, in turn, could negatively affect form quality codes and *Popular* responses, as well as *SCZI* scores. Accordingly, we do not recommend using the Hamel et al. (2000) sample as a normative approximation sample.

Nevertheless, other samples suggest clinicians should be cautious about using the existing CS reference values for children. Besides Hamel et al. (2000), other child and ado-

[6]Because of the skew inherent with *Lambda*, we recommend that median *Lambda* values be reported and that *Pure F%* (*Pure F/R*) be used (Meyer, Viglione, & Exner, 2001).

[7]It should be pointed out that the Hamel reliability data was derived using an across site coding procedure where the comparison scoring was done by a person trained in the same lab that did the rescoring for the Meyer et al. (2004) across site reliability study.

lescent reference samples have been collected in the United States and abroad (Erdberg, 2005; Erdberg & Shaffer, 1999), including France, Italy, Japan, and Portugal. These samples show some notable variability, particularly for *Dd*, *Lambda*, and form quality scores. It is too early to determine whether these differences reflect genuine cultural differences in personality and/or childrearing practices or if they are artifacts due to differences in administration, inquiry, or scoring conventions. However, the composite of data suggest that the adjustments offered earlier for adults should be made for children. In addition, for children, forensic examiners need to factor in developmental trends. The available international data suggest trends consistent with those for Exner's CS reference data (Wenar & Curtis, 1991) across the ages from 5 to 16. These include developmental increases in complexity markers like *DQ+*, *Blends*, and *Zf*, as well as increases in *M* and *P*. In addition, there is a decrease in *WSum6* and to a lesser extent in *DQv*. Unlike Exner's CS reference samples, the composite of alternative reference samples suggests clinicians should anticipate a decrease in *Lambda* as children age and an improvement in form quality scores. Ultimately, the same reasons that instigated the collection of a new adult CS normative reference group also apply to children, so that a new carefully collected age-stratified children's sample is desirable.

Based on the available evidence from the new CS adult reference sample and the other reference samples collected in the United States and internationally, we offer the following recommendations regarding normative expectations and use of nonpatient reference samples with the CS. For adults, we recommend that examiners use the new CS sample as their primary benchmark, but adjust for those variables that have consistently looked different in normative approximation and international samples. Examiners should consider the Shaffer et al. sample as an outer boundary for what might be expected from reasonably functioning nonpatients, because it shows what can be observed for nonpatients within the limits of current administration, inquiry, and scoring guidelines. Table 2–4 summarizes our current recommendations for modifications in normative expectations for crucial variables that have consistently diverged from CS norms. These include adjustments to form quality, color, texture, and human representations.

For children, we would recommend using the available age norms and make similar recommendations or adjustments to the same variables. Although we would not recommend the Hamel et al. sample as an outer boundary for what could be expected for younger U.S. children, its data illustrate how ambiguity or flexibility in current administration and scoring guidelines can result in obtaining some unhealthy looking data from apparently normal functioning children.

POTENTIAL MODERATORS FOR NORMATIVE EXPECTATIONS

Recent CS texts (Exner, 2003; Exner & Erdberg, 2005; Exner et al., 2001) have presented normative reference sample data broken out by *Lambda* and *EB* style. There are three *EB* styles formed by the ratio of *M* to *WSumC*: *Ambitent* ($M \approx WSumC$), *Extratensive* ($M < WSumC$), and *Introversive* ($M > WSumC$), all with *Lambda* less than 1. The fourth style is the *Avoidant* type, with *Lambda* greater than or equal to 1. Thus, the CS position is that style acts as a moderator variable. In other words, the association between a given Ror-

schach variable and an outcome or construct differs according to style. For example, one might interpret an Affective Ratio of .40 differently according to *EB/Lambda* style, that is, for a person with an *Ambitent* style versus an *Avoidant* style. If the *EB* and *Lambda* styles were such a moderator, then one would need to use different normative tables for each of the four styles, as recommended by the CS.

However, research support for *Lambda/EB* style as a moderator is lacking. The only such published empirical support known to us is greater validity for an old version of the *DEPI* among extratensives (Viglione, Brager, & Haller, 1988). Subsequently, in a study with adolescents, Krishnamurthy and Archer (2001) failed to find support for *EB* as a moderator for the current *DEPI*. Most of the support in CS texts for these four styles (Exner, 2003) relies solely on the fact that the norms differ for the four groups. For the most part, the differences in mean values across groups are redundant with *EB* and *Lambda*. For example, Extratensives produce a higher *Affective Ratio* and more *Blood* responses. These responses involve color or color cards, so that they are redundant with the Extratensive style, because they are concomitant to the *WSumC* elevation, $M < WSumC$.

There is considerable research support for *EB* as a measure of coping characteristics and for *Lambda* with simplification and coping limitations. There is no systematic or comprehensive data demonstrating incremental validity for other variables when taking *EB/Lambda* style into consideration. Or, put another way, there is no body of evidence to suggest that interpretation of other variables routinely varies by *EB/Lambda* style. Accordingly, we recommend that the forensic psychologists not rely on the normative reference group tables broken out by *EB/Lambda* style. Instead, as already recommended, use the tables that encompass all *EB/Lambda* styles, the $N = 450$ sample found in the new volume II (Exner & Erdberg, 2005), along with the interpretive adjustments recommended in Table 2–4.

This recommendation does not mean that there are no variables that might act as moderators for the interpretation of other variables. The most likely candidate is *R*, the number of responses. The relationship between *R* and other variables, as well as whether *R* is a moderator that should be controlled, has been argued in the Rorschach literature for a long time (Cronbach, 1949; Exner, 1974, 1992; Fiske & Baughman, 1953; Holtzman, 1958; Kinder, 1992; Lipgar, 1992; Meyer, 1992a). Research findings suggest that just about every score is associated with *R* and every other score when *R* is not controlled (Exner, Viglione, & Gillespie, 1984). Number of responses is closely related to the first factor on the Rorschach, characterized by Meyer (1992b) as "task engagement." This factor accounts for approximately 25% of Rorschach variance. Number of responses is already often controlled to some extent in percentages and ratios. The percentages (*X+%*, *Affective Ratio*, *Egocentricity Index*) are the variables with the most normal distributions, a desirable quality for applying psychometrics to refine interpretations and for efficiently evaluating validity through research.

Exner (1974) originally decided to let *R* vary and to use the less directive Klopfer (Klopfer & Kelley, 1942) response phase administration. This decision was partially based on the idea that the variation of *R* in the initial normative reference sample was considerably less compared to other research (e.g., Fiske & Baughman, 1953). However, as noted earlier, the new CS normative reference group sample ($SD = 5.68$; Exner &

TABLE 2–4

Recommendations for Adjustments to CS Adult Normative Expectations

Variable	*Research Adjustments to Expectations*	*vs. New CS Sample*[a]
Location + Form Quality		
Dd	3–4	1
X–%	.15–.25	.09–.14
X+%	.50–.60	.65–.70
XA%/WDA%	.70- .85	.80–.95
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
Ratio of *GHR* to *PHR* (*HRV*)	Between 3:2 and 1:1	2 to 1 ratio
AG	1 in 2 records	1 every record
Color + Related 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 in 5 records	1 in 3 records
EA	6–7	9
Texture		
T = 0	1 in 3 records	1 in 6
T = 1	1 in 3	2 in 3
T = 2	1 in 3	1 in 6
Miscellaneous		
Ambitent	1 in 2 or 3 records	1 in 5 records

[a]Exner & Erdberg, 2005, *N* = 450

Erdberg, 2005) is more variable than the previous normative reference group (*SD* = 4.40; Exner, 2001). Given that other scores are associated with *R*, this increased variation makes most other scores more variable. Thus, CS interpretive bands may be too narrow or may need to be modified by *R*. Eventually, research will need to supply the specifics: which variables and which criteria or interpretations are most affected by levels of *R*. We already know that the variables in the percentages, especially the *FQ* percentage scores, remain valid when *R* levels are considered. Also, the *Ego Impairment Index* contains an *R* correction that apparently contributes to interpretive accuracy.

Overall, the CS variables' association with *R* suggests that *R* should be at least considered in interpretation. To provide some data about which variables are most sensitive to levels of *R*, Table 2–5 classifies the correlations between CS variables and *R* by their magnitude. For those variables in the categories of very strong ($r > .6$) and strong ($r > .5$), and to some extent moderate ($r > .4$), it is probably advisable and perspicacious to take *R* into consideration when interpreting a protocol. A number of these variables, particularly the *FQ* variables and pairs (2), already have corrections for *R* in the form of percentages of *R* calculations. Correlations are quite abstract, so it is difficult to get a gut feel for the score differences that correspond to these correlations. Table 2–6 is provided to give the forensic examiners a feel for the various levels of correlations in Table 2–5. It is very clear that normative expectations for short and long records diverge considerably for variables that have moderate to strong correlations with *R*.

Looking at individual variables, the *FQ* percentage variables, for example *X–%* or *XA%*, and the *Egocentricity Index*, are typically relatively unaffected by *R*, except in very long records. Interpreting these variables in ratios to other variables partially corrects for *R*, although not completely. For other variables, notably the *HVI* and *OBS*, longer records are expected given the overproductive and detail-oriented coping styles exhibited by hypervigilant and obsessive individuals.

For the remaining variables, it is probably wise to consider *R*. These would include *Dd* and *D* locations; individual Developmental Quality and *Z-scores*; *FM*, *m*, shading, *es*,

TABLE 2–5

Score Correlations with R in a Mixed Patient, Offender, and Nonpatient sample (*N* = 1,342)

Correlation with **R**	*Location,* **DQ** *and* **Z**	*Determinants*	**FQ**	*Contents*	*Special Scores*	*Actuarial*
Very strong (> .6)	*D, Dd, DQo*	*F*	*FQo, FQu, FQ–*	*All A, Hd+(Hd)+Ad +(Ad)*		
Strong (> .5)	*ZF*	*(2), es*		*A, All H,* Non *Pure H*	*PHR*	
Moderate (> .4)	*S, DQ+, ZSum*	*FM + m, SumSh, EA*		*Hd, Ad*		*HVI*-tot, *OBS*-tot
Weak (> .3)		*M, FM, m, C', Y, a, p, Blends, WSumC, D-Score*[a]		*An + Xy, Cg, Sc*	*GHR, Lvl 2*	
Minimal (> .15)	*W, DQv, DQv/+, Afr*	*FC, CF+C, T, V, FD, Mp, Ma, MQ–, C-Sh-Blend*	*X–%, X+%, Xu%, WDA%, F + %*	*H, (H), (Hd), (A), (Ad), Art, Bt, Fd, Ge, Ls, Na, Sx, Iso/R, P*	*MOR, PER, HRV*[a]*, Individual Cog. SS, Sum6, WSum6*	*DEPI, SCZI*-tot,
Virtually none (≈ .0)	*W:D+Dd, Zd, W:M*		*FQ+, XA%, S–%,*	*H:*Non *pure H*	*AG, COP, CP*	*PTI, SCZI*-pos, *CDI, SCON, HVI*-pos, *EII*

[a]The correlations between R and both D-Score and HRV are negative.

TABLE 2–6

Examples of Changes in Mean Scores for Selected Variables Corresponding to Levels of Correlation With *R* in a Mixed Patient, Offender, and Nonpatient Sample (*N* = 1,342)

Correlation With R	*Example Variable*	*Low* R n = *493*	*Optimal* R n = *619*	*High* R n = *230*
		R = *14–17* *Mean* R = *15.4*	R = *18–27* *Mean* R = *21.7*	R > *27* *Mean* R = *35.1*
Very strong (> .6)	*Dd*	1.6	2.8	7.1
Strong (> .5)	*es*	6.1	8.9	14.5
Moderate (> .4)	*S*	1.7	2.7	4.7
Weak (> .3)	*Y*	1.1	1.6	2.7
Minimal (> .15)	*HRV*	0.6	0.4	1.3
Virtually none (≈ .0)	*CDI*	3.2	2.9	2.8

and to a lesser degree *EA*; all summary scores encompassing multiple human and animal contents and *A*, and to a lesser degree *Hd* and *Ad*; and *PHR*. For these variables within records containing less than 18 responses, the CS reference tables probably overestimate their expected frequency.

On the other hand, CS tables probably underestimate frequencies for long records with 28 responses or more. When possible, these "*R*-sensitive" variables should be interpreted with ratios or other arrays. For example, interpreting the pattern and interrelationships of the four *DQ* scores, or *HRV* rather than *PHR* or *GHR* individually, will reduce distortions due to *R*.

For records with 18 to 27 responses, the normative reference group tables probably provide excellent estimates of normative expectations, when modified by our recommendations in Table 2–4. Most likely, the Rorschach has the most validity for records of this length. For records outside of that range, interpretations *might* be more tentative and interpretations based on incorporating *R* adjustments should be considered.

VALIDITY: DOES THE RORSCHACH MEASURE WHAT WE THINK IT MEASURES?

We focus on construct validity, whether or not the test scale is measuring what we intend it to measure. Do the Rorschach variables, as a whole, show a pattern of convergent and discriminant validity? In other words, what is the evidence that given Rorschach variables are associated with appropriate and relevant criteria and not correlated with irrelevant or conceptually independent criteria? Oftentimes, distinctions in construct validity are made in terms of timing: An empirical association demonstrated at the same time the test is administered is referred to as concurrent validity, whereas an association with a criterion collected sometime in the future is referred to as predictive validity. Incremental

validity concerns whether we are deriving information that is not attainable elsewhere, a type of validity that we consider under utility, which concerns the usefulness of the test.

Of course, validity is ultimately demonstrated between a specific Rorschach variable and a specific construct or criterion relevant to that particular variable. However, organizing the vast literature by all the variables is a nearly insurmountable task, so that we address the global validity of the test. Does the evidence suggest that the Rorschach as a test produces valid measures of appropriated and relevant outcomes and constructs?

There have been thousands of studies addressing Rorschach validity from around the world (e.g., see summaries in Exner & Erdberg, 2005; Viglione, 1999), demonstrating considerable support for its validity and cultural adaptability. Based on these studies, Meyer and Archer (2001; also see Meyer, 2004) summarized the available evidence from Rorschach meta-analyses, including those that examined the global validity of the test and those that examined the validity of specific scales in relation to particular criteria. The scales included CS and non-CS variables. They then considered the evidence for the Rorschach in the context of evidence from meta-analyses on other psychological and medical tests (Meyer, Finn et al., 2001).

A number of factors make it challenging to compare findings across meta-analyses. Coefficients were not corrected for unreliability, range restriction, or the imperfect construct validity of criterion measures. Moreover, results emerged from different types of research designs and types of validation tasks. These differences cause effect sizes to fluctuate and make definitive comparisons of effect sizes difficult.

Nonetheless, the results of these meta-analyses indicate that psychological and medical tests have varying degrees of validity, ranging from tests that were essentially unrelated to a criterion, to tests that were strongly associated with relevant criteria. Contrary to some opinions, it was difficult to distinguish between medical tests and psychological tests in terms of their effects size patterns. At the same time, it was clear that test validity was a function of the criteria used to evaluate the instrument: Validity for a particular test was greater with some criteria and weaker with others. Within these findings, validity for the Rorschach was much the same as it was for other instruments. Thus, Meyer and Archer concluded that the systematically collected data showed the Rorschach produced validity coefficients that were on par with other personality tests, with meta-analytic effect sizes that supported its overall validity and usefulness. More specifically, they concluded that the results demonstrated that "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).

Consistent with Atkinson et al.'s 1986 comment that criticism of the Rorschach might be as much political as it is scientific, Meyer and Archer also express some puzzlement as to why the Rorschach might be singled out for intense scrutiny and criticism when its broadband validity is equal to other psychological tests.

Some individual meta-analyses have identified moderators of Rorschach validity, that is, factors or conditions that influence the validity of the test. Bornstein (1999) found considerable support for the validity of the Rorschach Oral Dependency scale as a predictor of

observed dependent behavior. Although his moderator analyses examined inkblot data combined with Thematic Apperception Test (TAT) data, he found that validity was consistent across criteria derived from lab, field, or classroom settings. It is true that in the single study from a hospital-clinic setting validity was nonsignificant. Bornstein also found that validity was consistent across ratings made by researchers or other observers and regardless of whether behavior was classified dichotomously or measured dimensionally. Thus, on the whole, the findings were generalizable across settings and methodology.

Hiller and Rosenthal and their coworkers (Hiller, Rosenthal, Bornstein, Berry, & Brunell-Neuleib, 1999) produced a comparative meta-analysis of Rorschach and MMPI research. They found that the Rorschach demonstrated greater association with what they called "objective" criteria. In contrast, the MMPI was more closely associated with psychiatric diagnostic classification and other self-reported measurements. A wide variety of events or outcomes was encompassed under the objective modifier. These criteria were largely behavioral events, medical conditions, behavioral interactions with the environment, or classifications that required minimal to no judgment from others, for example, dropping out of treatment, history of abuse or not, number of driving accidents, history of criminal offenses, medical disorder versus control, cognitive test performance, behavioral test of ability to delay gratification, or response to medication. Such characteristics and events were also identified as valid Rorschach criteria in a descriptive review of the same literature (Viglione, 1999). Many are behavioral events and life outcomes involving interactions between the individual and the environment that emerge over time. From a concrete perspective, these criteria for which the Rorschach is most valid might also be identified by an exclusionary definition as not self-report and not diagnostic classification. On the other hand, the data are clear that the Rorschach does identify psychotic diagnoses and measure psychotic symptoms well (Meyer & Archer, 2001; Perry, Minassian, Cadenhead, & Braff, 2003; Viglione, 1999, Viglione & Hilsenroth, 2001). Unlike many other disorders, these diagnoses are often based more on patients' observed behavior than on their self-reported presenting complaints.

In a recent meta-analysis, Grønnerød (2004) reviewed the literature examining the extent to which Rorschach variables changed as a function of psychological treatment. The Rorschach produced a level of validity that was equivalent to alternative instruments, so that it was as sensitive and able to measure change as self-report and clinician rating scales. Like Bornstein (1999) and Hiller et al. (1999), Grønnerød examined moderators to Rorschach validity. He found that Rorschach scores changed more with longer treatment, presumably because of more personality change over time. He also addressed some methodological issues. Suggesting some potential bias, when coders were blind to whether the protocol was obtained before or after treatment, there was less change. Another methodological note was that those studies that paid more attention to coding reliability, and how coding was accomplished, yielded greater validity coefficients. This is one of the few demonstrations of reliability constraining validity with real-world assessment applications.

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*, *RPRS*, and *SCZI/PTI*, although not for the *DEPI* when used as a diagnostic indicator. As is true for other commonly used tests, such as the MMPI–2, PAI, MCMI, or Wechsler scales, additional focused meta-analytic reviews that systematically catalog the validity of particular Rorschach variables relative to specific types of criteria will continue to refine and enhance clinical and forensic practice.

UTILITY: IS THE RORSCHACH USEFUL?

Utility of an assessment instrument can be globally defined as the practical value of the information it provides. It may be further specified as a function of the beneficial influence of a test on information, decisions, and outcomes relative to its costs (Viglione, 1999). Taking into consideration cost–benefit issues and the time necessary for examination and interpretation, the Rorschach "should provide information that is not routinely available through less time-consuming self-report, interview, or observational methods" (Viglione, 1999, p. 251). As an example, Viglione and Hilsenroth's (2001) review of the research on the CS Suicide Constellation revealed that it provided information about self-harm risk that was not easily attainable from the client through interview or from direct observation. This cost–benefit approach is typically translated statistically, even if it oversimplifies the issue, into an evaluation of incremental validity. In other words, the Rorschach and a more readily available or less time-intensive method are compared statistically. The requirement for incremental validity then would be the Rorschach accounts for variance in the outcome beyond that accounted for by the simpler method. Such a finding demonstrates statistically that the Rorschach provides unique information. Equating utility with statistical demonstrations of incremental validity is certainly reductionistic and research reports frequently lack adequate sample sizes to test it sufficiently. Nevertheless, much of the literature referring to utility uses this statistical method.

In addition to incremental validity, utility involves the prediction of real-world behavior and life outcomes, as demonstrated by the Hiller et al. (1999) meta-analysis. Research demonstrating validity within clinical or forensic practice, referred to as ecological validity, is especially important because it demonstrates the usefulness of the test in that applied context (i.e., utility). Having information about what is going to happen in the future and about patterns over time also provides great benefit. In this way, predictive validity, as contrasted to concurrent validity, also supports utility.

The empirical literature demonstrates that the Rorschach possesses utility in all of these forms. Research reviews (Viglione, 1999; Viglione & Hilsenroth, 2001; Weiner, 2001) contain empirical data consistent with the conclusion that Rorschach variables possess incremental validity over other tests, including self-report scales, intelligence test scores, demographic data, and other types of information. Meta-analyses (Hiller et al., 1999; Meyer, 2000a; Meyer & Archer, 2001) have reached the same conclusion. Moreover, these reviews and meta-analyses have demonstrated that the test is especially relevant for real-world behaviors, characteristics manifested over time, and life outcomes.

It is beyond the scope of this chapter to review individual studies, but a sampling of recent utility findings, many of them quite impressive, are presented here. This Rorschach research has continued to support the validity of the test through demonstrations of incremental validity with real-life outcome criteria. Thus, this sampling of studies from the United States and Europe continue to support the conclusion that the Rorschach yields important information that is not attainable through simpler, less time-consuming methods. Among the outcomes included in these studies are future success in naval special forces training in Norway (Hartmann, Sunde, Kristensen, & Martinussen, 2003), future adolescent and adult delinquency from clinician ratings of ego strength from Rorschach protocols taken at ages 4 to 8 in Sweden (Janson & Stattin, 2003), future psychiatric relapse among previously hospitalized children (Stokes et al., 2003), previous glucose stability levels among diabetic children in France (Sultan, Jebrane, & Heurtier-Hartemann, 2002), and future emergency medical transfers and drug overdoses in inpatients during a 60-day posttest period (Fowler, Hilsenroth, & Piers, 2001). In these studies, the Rorschach has demonstrated incremental validity over, for example, various self-report scales, collateral reports, *DSM* diagnoses, and intelligence tests.

Other studies demonstrate utility by using real-life behavioral and life outcome criteria. Several different research projects conducted in Sweden illustrate this nicely, using criteria such as eating behavior in an experimental setting, eventual weight loss, and positive response to obesity medication in an obesity treatment program (Elfhag, Barkeling, Carlsson, Lindgren, & Rossner, 2004; Elfhag, Barkeling, Carlsson, & Rossner, 2003; Elfhag, Carlsson, & Rossner, 2003; Elfhag, Rossner, & Carlsson, 2004); agreement between therapist's planned goals for treatment and what they actually focused on (Bihlar, 2001; Bihlar & Carlsson, 2001); and selection for intensive, long-term psychoanalytic therapy (Nygren, 2004a, 2004b). Many of these studies demonstrated predictive validity, which is another way of demonstrating utility because such information is not easily attainable.

This summary of recent utility studies is limited in a number of ways. Largely, the studies support the overall or broadband utility of the Rorschach. In other words, they support the test as a useful instrument. This summary does not address the utility of specific variables for specific applications. Most importantly, the findings for specific variables need to be replicated. Also, a strength shared by all of these studies was that the researchers articulated thoughtful hypothesized associations for specific Rorschach variables. Although the results were largely supportive, there also were negative findings, where results did not support the hypothesized variables. For instance, Elfhag et al. did not find support for the *ROD* in relation to eating behavior and Nygren did not find support for *m*, *X–%*, or *FD* as predictors of who would be selected for intensive psychotherapy.

As with reliability and validity research reports, most of these utility studies have used CS variables, but considerable incremental validity utility has also been demonstrated for some non-CS scales (Garb, 1999). These include the Rorschach Prognostic Rating scale (Meyer, 2000a; Meyer & Handler, 1997), the Rorschach Oral Dependency scale (Bornstein & O'Neill, 1997), and the *Ego Impairment Index* (Perry & Viglione, 1991; Viglione, Perry, & Meyer, 2003). The *Ego Impairment Index* is derived from standard CS variables, including *HRV*, *FQ–*, *WSum6*, *M–*, and certain critical contents—*An*, *Bl*, *Ex*,

Fd, *Fi*, *Sex*, *X-Ray*, *MOR*, and *AG*, in addition to *R* as a control variable. It has a great deal of empirical validity and utility support in the literature (Dawes, 1999; Perry & Viglione, 1991; Perry et al., 2003; Stokes et al., 2003, Viglione, Perry, & Meyer, 2003).

It has been demonstrated and reported many times in the literature that like-named Rorschach and self-report scales that purportedly measure similar constructs are weakly associated with one another, if at all (see, e.g., Archer & Krishnamurthy, 1993a, 1993b, 1997; Krishnamurthy, Archer, & House, 1996; Meyer, 1996, 1999; Meyer & Archer, 2001; Meyer, Riethmiller, Brooks, Benoit, & Handler, 2000; Viglione, 1996). Most of this work has used the MMPI as the self-report measure. These data suggest that the Rorschach should display incremental validity over self-report scales. From a logical and mathematical point of view, if both the Rorschach and a given self-report test are related with a given real-life outcome, and the Rorschach and self-report measure are not related to each other, both should be uniquely related to that outcome and both should provide incremental validity over the other. Nevertheless, the lack of association between the two methods and by implication the amount of method variance involved in assessment techniques, forces the forensic psychologist to employ a multimethod strategy (see Erdberg, chap. 27, this volume). Findings suggest that CS and self-reports are more highly correlated when patients take similar open/ guarded approach to both tests, and may be negatively correlated when they adopt opposing styles (Meyer, 1997b, 1999). Research also indicates that self-report is more easily manipulated (e.g., Meyer & Archer, 2001; Viglione, 1999). Accordingly, the Rorschach may be more useful in forensic assessment contexts when the respondent is motivated to exaggerate or minimize certain features. However, it has been demonstrated that many individuals—but not all—can influence obvious or dramatic Rorschach content and, to lesser extent, actuarial indices (Exner & Erdberg, 2005; Ganellen, chap. 5, this vol.; Meisner, 1988; Morgan & Viglione, 1992; Netter, 1991; Perry & Kinder, 1990).

CONCLUSIONS

Overall, the empirical evidence is consistent with the conclusion that the Rorschach can be reliably scored, is valid, and provides unique information. Generalizability of administrative procedures and global reliability, validity, and utility findings has been demonstrated in many countries internationally so that applicability to domestic subcultural groups is not problematic. However, there is much more to learn and document. The Rorschach is a complex instrument and, like any complex assessment tool, it poses challenges for reliable and accurate administration, scoring, and interpretation. We have highlighted some of the issues that we think are most important for forensic examiners to consider and have offered guidelines for revised interpretation based on the literature. The test will continue to be challenged in forensic practice because it is considered controversial by some and a symbol of problems with clinical practice and judgment by others. However, because it provides utility in the form of information that cannot necessarily be obtained easily from other sources, it will continue to be used in forensic contexts. We hope what we provided here assists forensic practitioners in accurately describing litigants and clients in an empirically defensible fashion, while being cognizant of the strengths and limitations of the test so that the legal system is served well.

REFERENCES

Acklin, M. W. (1999). Behavioral science foundations of the Rorschach test: Research and clinical applications. *Assessment, 6*, 319–324.

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.

Archer, R. P., & Krishnamurthy, R. (1993a). Combining the Rorschach and the MMPI in the assessment of adolescents. *Journal of Personality Assessment, 60*(1), 132–140.

Archer, R. P., & Krishnamurthy, R. (1993b). A review of MMPI and Rorschach interrelationships in adult samples. *Journal of Personality Assessment, 61*(2), 277–293.

Archer, R. P., & Krishnamurthy, R. (1997). MMPI–A and Rorschach indices related to depression and conduct disorder: An evaluation of the incremental validity hypothesis. *Journal of Personality Assessment, 69(3)*, 517–533.

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.

Barr, W. B. (2003). Neuropsychological testing of high school athletes: Preliminary norms and test–retest indices. *Archives of Clinical Neuropsychology, 18*, 91–101.

Benedict, R. H. B., Schretlen, D., Groninger, L., & Brandt, J. (1998). Hopkins Verbal Learning Test–Revised: Normative data and analysis of inter-form and test–retest reliability. *The Clinical Neuropsychologist, 12*, 43–55.

Berman, M. E., Fallon, A., & Coccaro, E. F. (1998). The relationship between personality psychopathology and aggressive behavior in research volunteers. *Journal of Abnormal Psychology, 107*, 651–658.

Bihlar, B., & Carlsson, A. M. (2000). An exploratory study of agreement between therapists' goals and patients' problems revealed by the Rorschach. *Psychotherapy Research, 10*(2), 196–214.

bihlar, B., & Carlsson, A. M. (2001). Planned and actual goals in psychodynamic psychotherapies; Do patients' personality characteristics relate to agreement? *Psychotherapy Research, 11*(4), 383–400.

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., & O'Neill, R. M. (1997). Construct validity of the Rorschach Oral Dependency (ROD) scale: Relationship of ROD scores to WAIS–R scores in a psychiatric inpatient sample. *Journal of Clinical Psychology, 53*(2), 99–105.

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

Cronbach, L. J. (1949). Statistical methods applied to Rorschach scores: A review. *Psychological Bulletin, 46*, 393–429.

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1972). *A MMPI handbook: Vol. 1: Clinical interpretation*. Minneapolis: University of Minnesota Press.

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.

Dawes, R. M. (1999). Two methods for studying the incremental validity of a Rorschach variable. *Psychological Assessment, 11*(3), 297–302.

Dikmen, S. S., Heaton, R. K., Grant, I., & Temkin, N. R. (1999). Test–retest reliability and practice effects of Expanded Halstead–Reitan Neuropsychological Test Battery. *Journal of the International Neuropsychological Society, 5*, 346–356.

Elfhag, K., Barkeling, B., Carlsson, A. M., Lindgren, T., & Rossner, S. (2004). Food intake with an antiobesity drug (sibutramine) versus placebo and Rorschach data: A crossover within-subjects study. *Journal of Personality Assessment, 82*(2), 158–168.

Elfhag, K., Barkeling, B., Carlsson, A. M., & Rossner, S. (2003). Microstructure of eating behavior associated with Rorschach characteristics in obesity. *Journal of Personality Assessment, 81*(1), 40–50.

Elfhag, K., Carlsson, A. M., & Rossner, S. (2003). Subgrouping in obesity based on Rorschach personality characteristics. *Scandinavian Journal of Psychology, 44*(5), 399–407.

Elfhag, K., Rossner, S., & Carlsson, A. M. (2004). Degree of body weight in obesity and Rorschach personality aspects of mental distress. *Eating & Weight Disorders, 9*(1), 35–43.

Erdberg, P. (2005, July). *Intercoder agreement as a measure of ambiguity of coding guidelines* Paper presented at the 18th International Congress of Rorschach and Projective Methods, Barcelona.

Erdberg, P., & Schaffer, 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.* Oxford, England: Wiley.

Exner, J. E. (1992). R in Rorschach research: A ghost revisited. *Journal of Personality Assessment, 58,* 245–251.

Exner, J. E. (2002). A new nonpatient sample for the Rorschach Comprehensive System: A progress report. *Journal of Personality Assessment, 78,* 391–404.

Exner, J. E. (2003). *The Rorschach: A Comprehensive System* (4th ed.). New York: Wiley.

Exner, J. E., Armbruster, G. L., & Viglione, D. (1978). The temporal stability of some Rorschach features. *Journal of Personality Assessment, 42*(5), 474–482.

Exner, J. E., 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: Vol. 2. Interpretation* (3rd ed.). Oxford, England: Wiley.

Exner, J. E., Meyer, G. J., Renteria, L., Mattlar, C.-E., Tuset, A. M., Gonzalez, Y., Nakamura, N., & Nihashi, N. (1999, July). *A cross-national review of Rorschach interscorer reliability.* Paper presented at the 16th congress of the International Rorschach Society, Amsterdam, The Netherlands.

Exner, J. E., Viglione, D. J., & Gillespie, R. (1984). Relationships between Rorschach variables as relevant to the interpretation of structural data. *Journal of Personality Assessment, 48*(1), 65–70.

Fiske, D. W., & Baughman, E. E. (1953). Relationships between Rorschach scoring categories and the total number of responses. *Journal of Abnormal and Social Psychology, 48,* 25–32.

Fowler, J. C., Hilsenroth, M. J., & Piers, C. (2001). An empirical study of seriously disturbed suicidal patients. *Journal of the American Psychoanalytic Association, 49*(1), 161–186.

Garb, H. N. (1999). Call for a moratorium on the use of the Rorschach inkblot test in clinical and forensic settings. *Assessment, 6*(4), 313–317.

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. (1999). Protecting the integrity of the legal system: The admissibility of testimony from mental health experts under *Daubert/Kumho* analyses. *Psychology, Public Policy, & Law, 5*(1), 224–242.

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*(2), 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., Sunde, T., Kristensen, W., & Martinussen, M. (2003). Psychological measures as predictors of military training performance. *Journal of Personality Assessment, 80,* 87–98.

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*(3), 278–296.

Hilsenroth, M. J., Fowler, J. C., & Padawer, J. R. (1998). The Rorschach Schizophrenia Index (SCZI): An examination of reliability, validity, and diagnostic efficiency. *Journal of Personality Assessment, 10*, 514–534.

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.

Hirt, M. E. (1962). *Rorschach science: Readings in theory and method*. Oxford, England: Free Press Glencoe.

Holtzman, W. H. (1958). *Holtzman inkblot technique*. San Antonio, TX: Psychological Corporation.

Hunsley, J., & Bailey, J. M. (1999). The clinical utility of the Rorschach: Unfulfilled promises and an uncertain future. *Psychological Assessment, 11*(3), 266–277.

Hunsley, J., & Bailey, J. M. (2001). Whither the Rorschach? An analysis of the evidence. *Psychological Assessment, 13*(4), 472–485.

Janson, H., & Olsson, U. (2001). A measure of agreement for interval or nominal multivariate observations. *Educational and Psychological Measurement, 61*, 277–289.

Janson, H., & Olsson, U. (2004). A measure of agreement for interval or nominal multivariate observations by different sets of judges. *Educational and Psychological Measurement, 64*, 62–70.

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

Kinder, B. N. (1992). The problems of *R* in clinical settings and in research: Suggestions for the future. *Journal of Personality Assessment, 58*, 252–259.

Klopfer, B., & Kelley, D. M. (1942). *The Rorschach technique*. Oxford, England: World Book.

Krishnamurthy, R., & Archer, R. P. (2001). An evaluation of the effects of Rorschach *eb* style on the diagnostic utility of the depression index. *Assessment, 8*(1), 105–109.

Krishnamurthy, R., Archer, R. P., & House, J. J. (1996). The MMPI-A and Rorschach: A failure to establish convergent validity. *Assessment, 3*(2), 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*(2), 27–66.

Lipgar, R. M. (1992). The problem of *R* in the Rorschach: The value of varying responses. *Journal of Personality Assessment, 58*, 223–230.

Martin, R., Sawrie, S., Gilliam, F., Mackey, M., Faught, E., Knowlton, R., & Kuzniekey, R. (2002). Determining reliable cognitive change after epilepsy surgery: Development of reliable change indices and standardized regression-based change norms for the WMS-III and WAIS-III. *Epilepsia, 43*, 1551–1558.

Mauger, P. A. (1972). *The test–retest reliability of persons: An empirical investigation utilizing the MMPI and the Personality Research Form*. Unpublished doctoral dissertation, University of Minnesota.

Meisner, S. (1988). Susceptibility of Rorschach distress correlates to malingering. *Journal of Personality Assessment, 52(3)*, 564–571.

Meyer, G. J. (1992a). Response frequency problems in the Rorschach: Clinical and research implications with suggestions for the future. *Journal of Personality Assessment, 58*(2), 231–244.

Meyer, G. J. (1992b). The Rorschach's factor structure: A contemporary investigation and historical review. *Journal of Personality Assessment, 59*(1), 117–136.

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. (1999). The convergent validity of MMPI and Rorschach scales: An extension using profile scores to define response/character styles on both methods and a re-examination of simple Rorschach response frequency. *Journal of Personality Assessment, 72*, 1–35.

Meyer, G. J. (1997a). Assessing reliability: Critical corrections for a critical examination of the Rorschach Comprehensive System. *Psychological Assessment, 9*(4), 480–489.

Meyer, G. J. (1997b). On the integration of personality assessment methods: The Rorschach and MMPI. *Journal of Personality Assessment, 68*(2), 297–330.

Meyer, G. J. (1997c). Thinking clearly about reliability: More critical corrections regarding the Rorschach Comprehensive System. *Psychological Assessment, 9*(4), 495–498.

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*(3), 356–370.

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

Meyer, G. J. (2001). Evidence to correct misperceptions about Rorschach norms. *Clinical Psychology: Science & Practice, 8*(3), 389–396.

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. Comprehensive handbook of psychological assessment* (pp. 315–342). Hoboken, NJ: Wiley.

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., Finn, S. E., Eyde, L., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56*, 128–165.

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), 1–38.

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*(2), 219–274.

Meyer, G. J., Mihura, J. L., & Smith, B. L. (2005). The interclinician reliability of Rorschach interpretation in four data sets. *Journal of Personality Assessment, 84*(3), 296–314.

Meyer, G. J., & Richardson, C. (2001, March). *An examination of changes in form quality codes in the Rorschach Comprehensive System from 1974 to 1995*. Paper presented at the annual meeting of the Society for Personality Assessment, Philadelphia, PA.

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.

Meyer G. J., Viglione, D. J., & Exner, J. E., Jr. (2001). Superiority of *Form %* over *Lambda* for research on the Rorschach. *Journal of Personality Assessment, 76*, 68–75.

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

Milott, S. R., Lira, F. T., & Miller, W. C. (1977). Psychological assessment of the burned patient. *Journal of Clinical Psychology, 33*, 425–430.

Morgan, L., & Viglione, D. J. (1992). Sexual disturbances, Rorschach sexual responses, and mediating factors. *Psychological Assessment, 4*(4), 530–536.

Murstein, B. I. E. (1965). *Handbook of projective techniques*. Oxford, England: Basic Books.

Netter, B. C., & Viglione, D. J., Jr. (1994). An empirical study of malingering schizophrenia on the Rorschach. *Journal of Personality Assessment, 62*(1), 45–57.

Nezworski, M. T., & Wood, J. M. (1995). Narcissism in the Comprehensive System for the Rorschach. *Clinical Psychology: Science & Practice, 2*(2), 179–199.

Nygren, M. (2004a). Differences in Comprehensive System Rorschach variables between groups differing in therapy suitability. In A. Andronikof (Ed.), *Rorschachiana xxvi: Yearbook of the International Rorschach Society* (pp. 110–146). Ashland, OH, US: Hogrefe & Huber.

Nygren, M. (2004b). Rorschach Comprehensive System variables in relation to assessing dynamic capacity and ego strength for psychodynamic psychotherapy. *Journal of Personality Assessment, 83*(3), 277–292.

Paolo, A. M., Tröster, A. I., & Ryan, J. J. (1997). Test–retest stability of the California Verbal Learning Test in older persons. *Neuropsychology, 11*, 613–613.

Perry, G. G., & Kinder, B. N. (1990). The susceptibility of the Rorschach to malingering: A critical review. Journal of Personality Assessment, 54(1–2), 47–57.

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*(1), 50–57.

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(3)*, 487–501.

The Psychological Corporation. (1997). *WAIS–III—WMS–III technical manual*. San Antonio: Author.

Rabin, A. I. (1981). *Assessment with projective techniques: A concise introduction*. New York: Springer.

Roberts, B. W., & DelVecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old Age: A quantitative review of longitudinal studies. *Psychological Bulletin, 126*, 3–25.

Rosenthal, R., & Rosnow, R. L. (1975). *The volunteer subject*. New York: Wiley.

Ryan, J. J., Dunn, G. E., & Paolo, A. M. (1995). Temporal stability of the MMPI–2 in a substance abuse sample. *Psychotherapy in Private Practice, 14*, 33–41.

Schuerger, J. M., Zarrella, K. L., & Holtz, A. S. (1989). Factors that influence the temporal stability of personality by questionnaire. *Journal of Personality and Social Psychology, 56*, 777–783.

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*, 420–428.

Sines, L. K., Silver, R. J., & Lucero, R. J. (1961). The effect of therapeutic intervention by untrained "therapists." *Journal of Clinical Psychology, 17*, 394–396.

Stokes, J. M., Pogge, D. L., Grosso, C., & Zaccario, M. (2001). The relationship of the Rorschach Schizophrenia Index to psychotic features in a child psychiatric sample. *Journal of Personality Assessment, 76*, 209–228

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.

Stone, L. A. (1965). Test–retest stability of the MMPI scales. *Psychological Reports, 16*, 619–620.

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

Sturman, M. C., Cheramie, R. A., & Cashen, L. H. (2005). The impact of job complexity and performance measurement on the temporal consistency, stability, and test–retest reliability of employee job performance ratings. *Journal of Applied Psychology, 90*, 269–283.

Sultan, S., Andronikof, A., Réveillère, C., & Lemmel, G. (2006). A Rorschach stability study in a non-patient adult sample. *Journal of Personality Assessment, 87*, 330–348.

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*(1), 122–141.

Viglione, D. J. (1989). Rorschach science and art. *Journal of Personality Assessment, 53*, 195–197.

Viglione, D. J. (1996). Data and issues to consider in reconciling self report and the Rorschach. *Journal of Personality Assessment, 67*, 579–587.

Viglione, D. J. (1999). A review of recent research addressing the utility of the Rorschach. *Psychological Assessment, 11*(3), 251–265.

Viglione, D. J. (2003). *Rorschach coding solutions: A reference guide for the Comprehensive System*. San Diego: Author.

Viglione, D. J., & Hilsenroth, M. J. (2001). The Rorschach: Facts, fictions, and future. *Psychological Assessment, 13*(4), 452–471.

Viglione, D. J., & Rivera, B. (2003). Assessing personality and psychopathology with projective methods. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Assessment psychology* (Vol. 10, pp. 531–552). 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*(1), 111–121.

Viglione, D. J., Brager, R. C., & Haller, N. (1988). Usefulness of structural Rorschach data in identifying inpatients with depressive symptoms: A preliminary study. *Journal of Personality Assessment, 52*(3), 524–529.

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*(2), 149–156.

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

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

Weiner, I. B. (2001). Advancing the science of psychological assessment: The Rorschach inkblot method as exemplar. *Psychological Assessment, 13*, 423–434

Wood, J. M., & Lilienfeld, S. O. (1999). The Rorschach Inkblot Test: A case of overstatement? *Assessment, 6*(4), 341–351.

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.

Wood, J. M., Nezworski, M. T., & Stejskal, W. J. (1996). The Comprehensive System for the Rorschach: A critical examination. *Psychological Science, 7*(1), 3–10.

Wood, J. M., Nezworski, M. T., Stejskal, W. J., & McKinzey, R. K. (2001). Problems of the Comprehensive System for the Rorschach in forensic settings: Recent developments. *Journal of Forensic Psychology Practice, 1*(3), 89–103.

Woodrome, S. E. & Fastenau, P. S. (2005). Test–retest reliability of the Extended Complex Figure Test–Motor Independent administration (ECFT–MI). *Archives of Clinical Neuropsychology, 20*, 291–299.

The Handbook of
Forensic Rorschach Assessment

Edited by
Carl B. Gacono · F. Barton Evans
with Nancy Kaser-Boyd · Lynne A. Gacono

Routledge
Taylor & Francis Group
New York London

Cover design by Kathryn Houghtaling.

Lawrence Erlbaum Associates
Taylor & Francis Group
270 Madison Avenue
New York, NY 10016

Lawrence Erlbaum Associates
Taylor & Francis Group
2 Park Square
Milton Park, Abingdon
Oxon OX14 4RN

Lawrence Erlbaum Associates is an imprint of Taylor & Francis Group, an Informa business

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-0-8058-5823-5 (Hardcover)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com