A REVIEW OF LITERATURE REGARDING SCIENTIFIC CONTROVERSIES SURROUNDING THE PSYCHOMETRIC PROPERTIES OF THE RORSCHACH INKBLOT TEST

by

Kevin N. Park

APPROVED:

_______________________________________ Date _______________
David M. Cimbora, PhD

_______________________________________ Date _______________
Joan W. Jones, PsyD

APPROVED:

____
Paul L. Poelstra, PhD, Interim Dean

____
Date

A REVIEW OF LITERATURE REGARDING SCIENTIFIC CONTROVERSIES SURROUNDING THE PSYCHOMETRIC PROPERTIES OF THE RORSCHACH INKBLOT TEST

A Doctoral Research Paper

Presented to

the Faculty of the Rosemead School of Psychology

Biola University

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Psychology

by

Kevin N. Park

May 2009

ABSTRACT

A REVIEW OF LITERATURE REGARDING SCIENTIFIC CONTROVERSIES SURROUNDING THE PSYCHOMETRIC PROPERTIES OF THE RORSCHACH INKBLOT TEST

by

Kevin N. Park

The Rorschach Inkblot Test has been the focus of intense controversy, significantly impacting clinicians who currently rely on Exner's Comprehensive System (CS; Exner, 2003) in clinical and forensic settings. This paper evaluates recent empirical CS research to determine whether or not it reveals lack of scientific merit as some skeptics have claimed. Relevant psychometric properties of the Rorschach are identified and evaluated to determine whether they meet accepted standards applicable to other psychological instruments in current use. Past reviews and recent empirical research published from 1998 to 2008 are critically reviewed.

Specific properties under investigation in this controversy are validity (incremental, convergent, and construct), reliability (interrater, test-retest), standardized procedures, and normative data. This literature review critiques a sample of recent empirical research related to each of these categories, and the findings generally reveal average to excellent psychometric properties. Trained and experienced examiners can use

the Rorschach with confidence in its scientific legitimacy, and charges that call for a moratorium on the Rorschach due to insufficient empirical backing are rejected.

The Rorschach should continue to be scrutinized in peer-reviewed journals using criteria similar to those applied to other psychological instruments. The legitimacy and utility of the Rorschach should be based on unbiased empirical evidence, and results should be communicated clearly to all consumers of Rorschach data in clinical, educational, and forensic settings. Areas of future research are discussed under the assumption that weaknesses in certain aspects of the Rorschach do not equate to illegitimacy of the CS as a whole.

TABLE OF CONTENTS

PAGE

LIST OF TABLES

A REVIEW OF LITERATURE REGARDING SCIENTIFIC CONTROVERSIES SURROUNDING THE PSYCHOMETRIC PROPERTIES OF THE RORSCHACH INKBLOT TEST

Introduction

Hermann Rorschach first introduced the Rorschach Inkblot Test to the psychological assessment community in 1921 with the publication of *Psychodiagnostik* (Rorschach, 1921). The Rorschach rapidly increased in popularity, and leading figures in the field were inspired by the Rorschach's potential as a unique approach to understanding personality. Early enthusiasm for its potential utility was eventually met by a demand for empirical validation, inspiring efforts to harmonize several disjointed Rorschach interpretive systems (Exner, 2003).

The Rorschach rose in popularity due in part to its association with psychoanalysis. However, as its popularity grew, so did fundamental disagreements on how the data should be interpreted. Some saw Rorschach responses as products of perception, which were indicators of how a person structures and organizes the external world. Others, following a more psychoanalytic approach, saw the responses as "projective productions or symbolic manifestations of a person's internal world" (Plake & Impara, 2001, p. 1034). Many leading researchers were increasingly convinced that the former approach held the most potential for empirical validation and acceptance in the scientific community.

By 1936, five different Rorschach interpretive systems were in use, and each system competed for validation for the next 32 years until 1968. After years of intense dispute between proponents of the various systems, the Rorschach Research Foundation was established to develop one system that could be subjected to empirical scrutiny. For about the next six years, members of the foundation focused on integrating the best aspects of all the systems, resulting in the introduction of John Exner's Comprehensive System in 1974 (Exner, 2003). Acceptance of the CS in the scientific community gradually widened based on the growing body of validation data and promising directions for further research (Exner).

Despite these advances, the association between the Rorschach and psychoanalysis persisted, and those who objected to "Freudian" theoretical underpinnings were also highly skeptical of any attempts at scientific validation of a projective test (Plake & Impara, 2001, p. 1034). A "culture clash," which is still evident in the current controversy, escalated in the scientific community with the introduction of the Minnesota Multiphasic Personality Inventory (MMPI; p. 1034). While the MMPI grew to be representative of the psychometric approach to psychology, the Rorschach entered the public's imagination through television and movies as the Freudian approach. The Rorschach's rise in popularity appeared to overshadow its scientific credentials which some still consider severely lacking. Nevertheless, since 1974 the CS has become the most widely researched and most commonly taught system for administering, scoring, and interpreting Rorschach responses (Guarnaccia, Dill, Sabatino, & Southwick, 2001).

Exner’s scientific advances on the CS were commended in 1998 by the American Psychological Association (APA) Board of Professional Affairs (BPA).[1]

As is expected in the realm of normal scientific debate, researchers, clinicians, and educators have critiqued the Rorschach as they would any other instrument. Such scrutiny is intended to reveal strengths and weaknesses and encourage improvement in future research design. Rorschach research had generally progressed in this manner until it was openly challenged in 1995. At this time, articles opposing the Rorschach started appearing in prominent assessment journals (*Journal of Personality Assessment*, *Assessment,* and *Psychological Assessment*) as well as several specialty publications (*Journal of Clinical Psychology, Journal of Forensic Psychology,* and *Psychology, Public Policy, and Law*). The growing controversy drew the attention of researchers, practitioners, and academics, which created renewed skepticism over the Rorschach’s legitimacy. Controversy and scientific scrutiny are nothing new to the Rorschach, and indeed such scrutiny provided the impetus to develop the CS in the first place. However, despite attempts to advance the CS based on empirical research, those associated with the Rorschach have been defending the legitimacy of the test and even their professional qualifications ever since the current controversy surfaced (Hilsenroth & Stricker, 2004).

A small cadre of opponents has been quite open about their opinion of the Rorschach as a “shoddy” test with an empirical foundation based on “junk science” (Hunsley & Bailey, 2001; Lilienfeld, Wood, & Garb, 2000). They have

[1] The APA BPA bestowed on John Exner its Award for Distinguished Professional Contributions to Knowledge (Weiner, Speilberger, & Abeles, 2002) for his work on the CS.

described the research methods used by CS researchers as “pseudoscience” (Lohr, Fowler, & Lilienfeld, 2002, p. 8), and they have concluded that Rorschach proponents are simply refusing to face the facts (Wood, Lilienfeld, Nezworski, & Garb, 2001) that clearly reveal the Rorschach’s lack of scientific credibility. Lohr et al. assert that Rorschach researchers have avoided the scientific credibility issue by relying on the Rorschach’s popularity and making “extravagant claims” that the Rorschach “…possesses special, even remarkable capacities” (p. 6). Rorschach proponents deny this claim and point out that these comments were taken directly from the text of the citation by the APA BPA when it honored John Exner’s work on the CS in 1998 (Weiner, Speilberger, & Abeles, 2002).

In response to these charges, Rorschach proponents have attempted to conduct empirical research under the scrutiny of peers, opponents, and unbiased sources. Additionally, they have conducted multiple meta-analyses (Meyer, 2000, 2001; Meyer et al., 2001; Rosenthal, Hiller, Bornstein, Berry, & Brunell-Neuleib, 2001; Viglione & Hilsenroth, 2001) on the psychometric properties of the Rorschach in an attempt to respond to attacks with scientific evidence. This concerted effort by Rorschach proponents to exemplify the scientific method appears to have had little or no effect on opponents (Weiner et al., 2002), who have concluded that there is a “…half-century of largely negative scientific evidence” against the Rorschach and a “…wealth of scientific evidence that the test is of questionable utility for real-world decision making” (Wood, Nezworski, Lilienfeld, & Garb, 2003, p. 1). Lohr et al. (2002), speaking on behalf of the opposition to the Rorschach, have appealed to the APA and other professional organizations to "…impose stiff sanctions, including expulsion if necessary, on practitioners who routinely use therapeutic and assessment practices that are devoid of scientific support" (p. 8).

In addition to dismissing the research supporting the Rorschach, opponents declare that it is not a legitimate psychological instrument and should no longer be taught or administered. They characterize it as an unscientific mind-reading technique in their recent book, *What's Wrong with the Rorschach? Science Confronts the Controversial Inkblot Test* (Wood et al., 2003). In the introduction, they state:

> [M]ost psychologists in clinical practice have treasured the test as one of their most precious tools. And . . . many of their respected scientific colleagues have been trying to persuade them that the test is well-nigh worthless, a pseudoscientific modern variant on tea leaf reading and Tarot cards. (p.1)

Proponents of the Rorschach assert that only those with adequate training and clinical experience with the CS understand its true nature as a complex and nuanced behavioral observation methodology. They staunchly defend the psychometric soundness of the CS; however, they readily admit that it is difficult to administer, score, and interpret accurately (Archer, 1999; Exner, 2003). Highlighting these Rorschach weaknesses, opponents characterize Rorschach scoring issues as complex, vague, and unscientific. Wood et al. state:

> [T]he test bears a charming resemblance to a party game. A person is shown ten inkblots and asked, "What might this be?" Like swirling images in a crystal ball, the ambiguous blots tell a different story to every person who looks upon them. There are butterflies and bats, diaphanous dresses and bow ties, monkeys, monsters, and mountain-climbing bears. When scored and interpreted by an expert, people's responses to the blots

> are said to provide a full and penetrating portrait of their personalities. (p. 1)

This sardonic portrayal of the process of Rorschach administration, scoring, and interpretation appears consistent with the perception of the Rorschach in popular culture. In an attempt to separate popular perception from scientific discourse, this literature review attempts to focus primarily on CS empirical data and avoid philosophical arguments about the Rorschach or projective assessment in general.

Regardless of politics and ideology of those involved in this controversy, the fact remains that the Rorschach is currently being utilized extensively in graduate programs, clinical practices, psychiatric hospitals, and courtrooms (Bornstein, 2001). It is important to emphasize that the controversy reveals markedly opposing views of the same extant literature. For well over a decade, proponents have dedicated considerable resources in producing empirical data as a rejoinder to ongoing attacks. Opponents also have invested considerable resources, and it appears now that the conflict is at an impasse. Even so, there does not appear to be any argument on either side over the need for expedient resolution in the best interest of patients, clients, clinicians, and educators.

The relevancy of the current controversy extends well beyond the academic realm to those who use the Rorschach and depend on reliable data to back up their decision-making. Consumers of Rorschach data are left with a poignant real-world question raised by this controversy (Meyer et al., 2001): Does the Rorschach demonstrate acceptable psychometric properties according to the same scientific standards applied to other instruments? This review hypothesizes that this question can be addressed by analysis of a broad sample of recent empirical CS research. Past research and meta-analyses are relevant in this review insofar as they have contributed to the most recent research on the CS.

This paper reviews a sample of recent empirical articles relevant to the following psychometric properties commonly discussed by both sides of the controversy: (a) validity (incremental, convergent, and construct) (b) reliability (interrater and test-retest) (c) standardized procedures, and (d) normative data (Weiner et al., 2002). This rather broad approach is in response to the broad nature of the attacks on the CS as having little or no scientific merit (Wood et al., 2001). These psychometric properties are evaluated to determine whether they fall within currently accepted standards for psychological instruments. Failure of empirical research to meet minimum standards in these categories would warrant concern over the fundamental psychometric soundness of the Rorschach. Past reviews and recent empirical research published from 1998 to 2008 are critically reviewed. Although CS variables are the main focus of critique, several experimental non-CS variables are discussed only as they relate to the psychometric properties examined in this report. Suggestions for further research are also explored.

Chapter One: The Rorschach Inkblot Test

In order to determine the standard to which the Rorschach should be held, it must first be determined if psychological assessment instruments in general meet standards universally recognized throughout all fields of scientific research. The Rorschach should be at least as comparable to other widely used psychological tests, and these same psychological tests should be comparable to tests used in other fields of study. The American Psychological Association's Board of Professional Affairs appointed a Psychological Assessment Work Group (PAWG) to objectively evaluate efficacy of psychological assessment compared to instruments used in other fields of research. The group conducted an exhaustive meta-analysis of 125 other meta-analyses on test validity, and they reached four conclusions: (a) psychological tests have acceptable validity, (b) psychological tests have validity similar to medical tests, (c) psychological tests have empirically proven clinical utility, and (d) psychological tests offer significant improvement over interview alone (Meyer et al., 2001). They concluded, "…psychological test validity is strong and compelling" and "…comparable to medical test validity" (p. 128). A few specific correlations they listed for comparison purposes are shown in Table 1.

Table 1

A Sample of Psychological and Medical Test Validity:

(a) Weight and height for U.S. adults (.44)

(b) MMPI validity scales and detection of malingered psychopathology (.44)

(c) WAIS IQ and obtained level of education (.44)

(d) Rorschach PRS scores and subsequent psychotherapy outcome (.44)

(e) Viagra and improved male sexual functioning (.38)

(f) Rorschach dependency scores and dependent behavior (.37)

(g) MMPI scale scores and ability to detect depressive or psychotic disorders (.37)

(h) Screening mammogram results and detection of breast cancer within 1 year (.32)

(i) Sleeping pills and short-term improvement in chronic insomnia (.30)

(Meyer et al., 2001, p. 128).

These results have been supported by similar studies by Viglione (1999), Viglione and Hilsenroth (2001), and Weiner (2001b). The wide scope of these studies suggests that as a psychological test, the Rorschach should be able to meet minimally acceptable standards as compared to instruments reviewed in the PAWG study.

The Rorschach is essentially a behavior observation methodology that relies on the application of a complex coding system to samples of verbal behavior. It encompasses nomothetic trait and idiographic behavior approaches to personality assessment, yielding a plethora of behavioral observation data. Individual codes are applied to target behaviors (i.e., key words) within each response, and then they are tabulated, summed, and combined to form interpretive indexes. Organization of response-level verbal behaviors (e.g., location, developmental quality, determinants) facilitates the process of coding and interpretation (Acklin, McDowell, Verschell, & Chan, 2000). The Rorschach is not a diagnostic test. It is intended to identify personality characteristics of the individual, and its utility derives from the relevance of these characteristics to "decision-making in clinical, forensic, health care, educational, and organizational settings" (Weiner et al., 2002, p. 6).

For the respondent, the Rorschach is an unstructured task of perception and verbalization of perception, characterized by the following: (a) test stimuli have no face validity, (b) the testing administration is verbally interactive, (c) test results are not vulnerable to impression management due to lack of face validity, (d) stimuli are novel, unless the examinee has taken it before, (e) there are no clear expectations of the examinee and examiner, (f) the task elicits subconscious thought processes, and (g) ideographic and nomothetic interpretations are used (Exner, 2003). The Rorschach response process is a complex phenomenon that occurs according to the respondent's own perceptions, classifications, decision-making processes, and psychological traits and states (Dean, Viglione, Perry, & Meyer, 2007). Respondents generate many more perceptions to the stimulus than they report during the response phase.

As a projective test, the Rorschach has significant differences from self-report measures such as the MMPI, which is one of the most widely used and extensively researched self-report measures. The relevancy of the MMPI to this literature review is observed in the significant number of studies that include both the MMPI and Rorschach. Additionally, both instruments are frequently used in clinical settings as part of a multi-modal assessment approach that is considered a standard of assessment practice (Meyer, 1999b).

The Rorschach is often compared to the MMPI in the literature to establish a comparative baseline of acceptability in the field of psychological assessment. Evaluation of the Rorschach in conjunction with the MMPI as part of an assessment battery is a vital part of determining real world utility. Therefore, this literature review attempts to evaluate the Rorschach alone and in conjunction with the MMPI (Meloy, Acklin, Gacono, Murray, & Peterson, 1997). The rationale for this approach is based on an extensive meta-analysis by Rosenthal et al. (2001), who found an unweighted mean validity coefficient of .29 for Rorschach variables in 2,276 protocols. In this study, similar methods were applied to 5,007 MMPI protocols in research published during the same time period that yielded an unweighted mean validity coefficient of .30. These values are considered well within the acceptable range for both tests.

Chapter Two: Review of the Literature

The specific empirical research studies reviewed in this paper are representative of the current literature from 1998 to 2008 on the psychometric properties of the CS. Empirical articles reviewed are organized according to their relevance to the following: Validity (incremental, convergent, and construct) (b) reliability (interrater and test-retest) (c) standardized procedures, and (d) normative data. Some articles cover more than one psychometric property; therefore, only the primary property under investigation is reviewed for each article. Utility of the Rorschach in psychotherapy is not included as part of this review.

Psychometric Properties of the CS

Validity

Validity is the extent to which a test measures what it is supposed to measure. Without adequate validity, Rorschach results have no practical value in real-world decision-making. It is important to note that validity is not determined by a single statistic, but by a body of research that demonstrates the relationship between the test and the behavior it is intended to measure. The articles reviewed in this section address three types of validity commonly discussed in CS literature: incremental, convergent, and construct.

Incremental validity. Incremental validity is widely recognized as a property necessary for a clinical diagnostic instrument to possess (Blais, Hilsenroth, Castlebury, Fowler, & Baity, 2001). It is determined by comparing the rates of correct diagnostic identification for psychological tests individually and in combination with one another. Incremental validity is important for clinicians in that a battery of tests is assumed to be more accurate in identifying or predicting pathology than a single measure (Ganellen, 1996). This involves determining whether the rate of correct identification of diagnoses increases when multiple psychological instruments are administered. In the current healthcare environment, incremental validity can be used to determine whether it is worth the additional time, effort, and cost to administer, score, and interpret multiple instruments. The Rorschach is a labor-intensive instrument compared to others, which highlights the importance of incremental validity in the current controversy (Meyer et al., 2001).

Opponents accurately cite extant literature that ostensibly reveals mediocre empirical support for combining the Rorschach and MMPI-2 to establish incremental validity. Analysis of 37 prior studies on the relationship of the two tests reveals mostly weak or insignificant overall correlation between conceptually similar MMPI-2 variables (Wood, Nezworski, & Stejskal, 1996). The analysis revealed that 51% of the studies reported non-significant associations and 22% yielded only weak associations across the two tests overall, including personality disorder (PD) criteria. Only 27% of the studies reported significant findings, but these correlations tended to be modest (*rs* = .24 to .34), suggesting that conceptually similar MMPI-2 scales and Rorschach variables are essentially unrelated. Wood et al. (1996) interpreted these results as evidence that the Rorschach is an illegitimate test based on an inadequate scientific foundation compared to the MMPI-2.

Contrary to this interpretation, Blais et al. (2001) contend that the same 37 past studies have revealed weak correlation due to inadequate methodology that does not take into account the variation in response styles for the Rorschach and MMPI-2. Support for this hypothesis is based on Meyer's (1999b) findings regarding the issue of response style. Meyer found that cross-method correlations in general are typically modest in psychometric research when comparing any two instruments. To address this problem, he developed a more sophisticated methodology that could be applied to Rorschach and MMPI-2 data. According to Meyer, when Rorschach and MMPI-2 data are grouped according to similar response style (e.g., expanded versus dilated), strong cross-test associations are found. Therefore, Blais et al. (2001) used Meyer's methodology for their study by taking into account response style on the Rorschach and MMPI-2 to obtain the most valid measure of incremental validity.

Blais et al. (2001) investigated incremental validity by measuring the ability of the Rorschach and the MMPI-2 to predict *DSM-IV* Cluster B PD criteria. The authors used multiple-regression analyses to explore the ability of select Rorschach variables and the MMPI-2 PD scales to predict *DSM-IV* Cluster B PD criteria in a sample of treatment-seeking outpatients diagnosed with the following PDs: antisocial personality disorder (ANPD), borderline personality disorder (BPD), histrionic personality disorder (HPD), and narcissistic personality disorder (NPD). The importance of studying the incremental validity of the Rorschach and MMPI-2 is seen in the fact that they are by far the most utilized instruments by researchers and clinicians, who have produced over 16,000 publications at the time this article was written. However, there is a need for more research about the interrelationship between the two instruments and their combined utility in predicting *DSM-IV* diagnoses.

The data used in this study were drawn from an archival search of patients' files at a university-based outpatient psychology clinic. This was a reanalysis of existing data used by Baity and Hilsenroth (1999), who focused on the conceptual and empirical relation to one of the Cluster B PDs. All the multiple-regression analyses and the majority of the correlations reported in this article were original to the Blais et al. (2001) study.

This article included approximately 800 case files covering a 7-year period. The three phases used to select cases were similar to the Baity and Hilsenroth (1999) methods and procedures for data collection. Mark Hilsenroth, who was blind to previous scores and patient diagnoses, re-scored all Rorschach protocols. Interrater reliability was obtained by having J. Christopher Fowler, a prominent Rorschach researcher and clinician, score 20 randomly selected protocols, and he was blind to Hilsenroth's scores and diagnoses. The resulting interrater agreement for the Structural Summary and the Rorschach content scales were above 80%, which is in the acceptable range. All

Rorschach variables used in this study were selected a priori based on either prior theoretical or empirical links to *DSM-IV* Cluster B PD criteria.

The MMPI-2 was administered and reviewed for validity, and the MMPI-2 Cluster B PD (ANPD, BPD, HPD, NPD) scales were scored and included as variables in this study. These scales have been frequently used in personality research. Stepwise multiple regression analysis was used to explore the interrelationship between select MMPI-2 scales and CS variables. Results indicated a limited relationship, with only 5 of 30 correlations reaching significance ($p < .05$). This was considered a weak interrelation between the two instruments, which was expected. However, multiple and hierarchical regression analysis showed that the combined MMPI-2 and Rorschach data added incrementally to the prediction of some *DSM-IV* Cluster B PD criteria total scores. Specifically, combined data accounted for 33% of the variance for NPD, and 48% of the variance for BPD, which the authors considered to be an "impressive" finding (Blais et al., p. 163). This data suggests that two different types of assessment instruments, projective and self-report, can accurately predict some *DSM-IV* diagnoses when the data are combined.

In conclusion, it appears that the Blais et al. (2001) study improves the CS incremental validity literature. The authors have provided adequate empirical support for combining Rorschach and MMPI-2 data to predict clinically relevant, non-test, and real-world behaviors according to *DSM-IV* criteria. They conclude that their findings support the tradition of combining Rorschach and MMPI-2 data in clinical practice. Future studies should take into account the examinee's response style on the Rorschach and MMPI-2 when comparing conceptually similar constructs.

These conclusions should be interpreted carefully, as only two PD scales revealed significance (NPD and BPD) after combining Rorschach and MMPI-2 data. Therefore, it is necessary to conduct future research that can replicate this study with more diagnostic categories before the results can be generalized. Nevertheless, the value of research in this area appears promising. No single instrument claims to measure all aspects of personality, and combining self-report and projective data to measure the same construct may yield a more in-depth and multifaceted understanding of personality functioning. This sentiment is echoed in the most recent edition of *The Fourteenth Mental Measurements Yearbook*, which states "Rorschach data are not similar to any other psychological data we are currently gathering, and we will most likely tap a rich vein of material otherwise overlooked" (Plake & Impara, 2001, p. 1037).

Convergent validity. Convergent validity (Ganellen, 1996) has to do with the rate of agreement among the different methods of assessment that are supposed to measure the same thing, such as the Rorschach and MMPI-2. Agreement between measures may increase the probability that a particular diagnosis is correct (true positive), while disagreement between measures may lower the probability that a diagnosis is accurate. This is an important issue for clinicians, as they can have more confidence in making a correct diagnosis with a strong convergent validity.

Convergence of Rorschach and MMPI-2 constructs is partially a function of how patients interact with the tests. When patients approach each test in a similar manner, conceptually similar constructs tend to correlate. When patients approach each test in an opposing manner, conceptually similar constructs tend to be negatively correlated. When the manner of test interaction is ignored, MMPI-2 and Rorschach constructs tend to be uncorrelated.

Meyer, Riethmiller, Brooks, Benoit, and Handler (2000) studied convergent validity by taking into account test interaction approaches for the Rorschach and MMPI-2. They attempted to expand upon a previous study (Meyer, 1999b) suggesting that test interaction approaches were important moderators of convergent validity. In the prior research, when test interaction approaches were ignored, there was virtually no association between Rorschach and MMPI-2 scales that shared similar names. When analyses were restricted to the subset of patients who interacted with both tasks in a similar manner, the same Rorschach and MMPI-2 scales were substantially correlated.

Based on this prior research, Meyer et al. (2000) tested several hypotheses in a new sample of patients. First, they expected conceptually related Rorschach and MMPI-2 scales to be uncorrelated when response approaches were ignored. Clinically, this reflects the assumption that each test would generally provide distinct information that could not be obtained directly from the other. Second, they expected conceptually related Rorschach and MMPI-2 scales to be positively correlated when analyses were limited to those patients who had similar test interaction approaches on both methods.
Third, they expected conceptually related MMPI-2 and Rorschach scales to be negatively correlated when analyses were limited to patients who had an opposing response

approaches on each method. Clinically, this is based on the assumption that opposing response-styles would be influenced in part by deliberate efforts to manipulate the tests.

The authors attempted to address prior criticism of weak CS convergent validity by developing new criteria to identify the following test-taking approaches: Constricted on both tests, dilated on both tests, constricted on the Rorschach and dilated on the MMPI-2, and constricted on the MMPI-2 and dilated on the Rorschach. Dilated and restricted protocols are defined by the amount of information obtained from the tests in terms of number of responses (*R)* and their complexity. Determinants, Developmental Quality, Form Quality, and Special Scores require coding decisions of varying complexity. Dilated refers to a higher than optimal amount of information, and constricted refers to a lower than optimal amount of information.

Meyer et al. (2000) defined the specific parameters of dilated and constricted protocols by using the same cutoff values in the newer study as in the previous Meyer (1999b) study. Patients were classified by scales that are well-established indicators of test-taking style for the Rorschach and MMPI-2: *R* and *Lambda* for the Rorschach, and *F* and *K* for the MMPI-2. The authors used the following cutoff values to define dilated and constricted protocols for the Rorschach and MMPI-2. For the Rorschach, protocols were constricted if *R* was less than 21 and *Lambda* was greater than .55, and they were dilated if *R* was greater than 21 and *Lambda* was less than .55. For the MMPI-2 scales, protocols were constricted if *F* was less than 58 and *K* was greater than 50, and they were dilated if *F* was greater than 58 and *K* was less than 50.

Meyer et al. (2000) obtained the sample for the new study from an outpatient clinic in Tennessee that serves university students and community residents. The sample contained 327 patients who completed the Rorschach and the MMPI-2. It is important to note that MMPI-2 protocols reflecting random responding were excluded. Rorschach protocols with 13 or less responses and Lambda > .50 were also excluded. Convergent validity of the Rorschach and MMPI-2 was assessed using variables related to three psychological constructs: affective distress, psychotic processes, and interpersonal suspiciousness or wariness.

The hypotheses of this study mentioned earlier are reviewed in light of the following results.[2] In the first hypothesis, the authors expected that conceptually related Rorschach and MMPI-2 scales would be uncorrelated when response styles were ignored. The average correlation among the 17 variable pairs tested was .055, which was expected. This generally parallels prior research reported by Meyer (1997), who found an average correlation of .03 for the same 17 variable pairs.

In the second hypothesis, they expected conceptually related Rorschach and MMPI-2 scales to be positively correlated when analyses were limited to those patients who had similar test interaction approaches on both methods. The authors found that conceptually aligned constructs on the Rorschach and MMPI-2 were highly correlated when approaches to testing were similar. For each of the three construct domains (emotional distress, psychotic process, and interpersonal wariness), results revealed

[2] The authors conclude that this data closely replicates previous research (Meyer, 1999b).

substantial correlations between similarly named Rorschach and MMPI-2 scales. The average correlation across all the results was .42, which was considered strong.

In the third hypothesis, the authors expected conceptually related MMPI-2 and Rorschach scales to be negatively correlated when analyses were limited to patients who had an opposing response style on each method. This hypothesis was not strongly supported by the results, which revealed that the average correlation across the results was -.26. The authors concluded that these results indicate a greater degree of instability associated with discordant styles than with similar test interaction styles. There is relatively little research on convergent validity for this response style, and the authors admit that the negative correlation is relatively weak.

In summary, Meyer et al. (2000) attempted to demonstrate that two distinct and independent measures, the Rorschach and MMPI-2, could produce statistically significant convergent validity when taking into account test interaction styles. They conclude that clinicians can anticipate stronger convergent validity with Rorschach and MMPI-2 findings related to affective distress, psychosis, and interpersonal wariness when patients display a similar style of interacting with each test. Conversely, clinicians can expect the same three Rorschach and MMPI-2 constructs will show a greater degree of disagreement when patients approach each task in a different manner (e.g., constricted on one measure and dilated on another). The authors suggest that future Rorschach research should explore in more detail the characteristics of patients who produce certain cross method interaction styles.[3] This would make convergent validity data more applicable in clinical situations.

One possible weakness of this study was the relative lack of examiner experience. The authors used student examiners to administer, score, and interpret results. Each student had a minimum of 2 years of training, including completion of a one-semester course. The examiners were enrolled in an advance personality assessment course, and they received 3 hours of supervision per week. The possibility of scoring inaccuracy was not addressed in this study, even though this is a major issue in Rorschach research.

Construct validity. In psychometrics, construct validity refers to whether a scale measures or correlates with a theorized psychological construct. Fowler, Piers, Hilsenroth, Holdwick, and Padawer (2001) investigated the construct validity of the CS Suicide Constellation (*S-CON*; Exner, 1993). They attempted to measure its ability to distinguish near-lethal suicide attempts from non-suicidal and parasuicidal inpatients. The sample consisted of 97 women and 7 men admitted to an inpatient treatment center in Massachusetts. They were classified as nonsuicidal (*n* = 37), parasuicidal (*n* = 37), and near-lethal (*n* = 30) according to extensive record review and categorization derived from the Lethality of Suicide Attempt Rating Scale (LSARS; Smith, Conroy, & Ehler, 1991).

In construct validity studies, one must first ensure adequate interrater reliability, which the authors attempted to ensure in response to pointed criticism of *S-CON* reliability by Wood, Nezworski, and Stejskal (1996). The first author, who was blind to

[3] Constricted on both tests, dilated on both tests, constricted on the Rorschach and dilated on the MMPI-2, and constricted on the MMPI-2 and dilated on the Rorschach.

all identifying information and group assignments, re-scored all 104 CS protocols. Twenty protocols were then randomly selected and scored independently by the second author, producing percentage of agreement and kappa coefficients (κ) for all major scoring categories comprising the *S-CON*. Percentage agreement ranged from 97% to 100%, and values of κ ranged from 0.96 to 1.00, which demonstrates excellent interrater reliability.

Analysis of data, according to ANOVAs of total *S-CON* scores, revealed significant differences among the three groups under investigation.[4] Post hoc analyses using Tukey's HSD confirmed that the near-lethal group mean scores were significantly higher than parasuicidal and nonsuicidal groups, $F(2, 100) = 14.3, p < .00001$.[5] The *S-CON* demonstrated strong validity when the score was equal to or greater than 7, revealing 81% true positive rate for predicting near-lethal versus parasuicidal activity and nonsuicidal patients, which is an impressive finding. Comparison with the non-clinical control group yielded a large effect size ($\eta = .84$).

The authors concluded that the *S-CON* is significantly more useful than self-report measures in distinguishing between suicidal and parasuicidal behavior for inpatients. Results revealed that an *S-CON* score of 7 or more was the sole predictor of near-lethal suicide attempts among 9 psychiatric and demographic variables. This

[4] $F(2, 100) = 14.3, p < .00001$.

[5] The relative magnitude of correlations are equivalent to small, medium, and large effect sizes, $r = .20$, .50, and .80 respectively in the psychological sciences.

finding supports the construct validity of the *S-CON* as a unique and valuable tool in the assessment of suicide risk.

In addition to predicting near-lethal suicidal behavior, the *S-CON* failed to predict parasuicidal behavior. Other CS indexes and behavioral indicators of pathology also failed to predict near-lethal or medically serious suicide attempts. These results strengthen four previous studies that have validated the ability of the *S-CON* to assess lethality of suicidal behavior. Fowler et al. (2001) cited previous studies indicating that 8 to 10 times as many failed suicides as completed suicides occur in the U.S. population, revealing prediction of near-lethal suicide attempts as a critical problem facing clinicians in short-term acute care facilities.[6] The overall results of this study appear to be promising for the construct validity of the *S-CON*. However, the *S-CON* is only one index on the CS, and other indexes and variables must be empirically validated in order to accurately assess the overall validity of the Rorschach. This study does not reveal any indication that the *S-CON*, a prominent CS index, lacks empirical backing. On the contrary, the Fowler et al. (2001) study appears to support ongoing improvement of construct validity with other Rorschach variables.

Kamphuis, Kugeares, and Finn (2000) also explored construct validity. They investigated several Rorschach variables to determine their utility in discriminating

[6] Fowler et al. (2001) also compared the predictive validity of the *S-CON* to several widely used self-report measures related to suicide risk. The diagnostic efficiency of the *S-CON* was evaluated according to five statistics: sensitivity (SN), specificity (SP), positive predictive power (PPP), negative predictive power (NPP), and overall correct classification (OCC). Although comparisons are preliminary and exploratory, results suggest significant sensitivity and specificity of the *S-CON* over self-report measures in predicting near-lethal suicidal behavior.

between nondissociative outpatients with histories of (a) definite sexual abuse (DSA; *n* = 22), (b) suspected but unconfirmed sexual abuse (SSA; *n* = 13), or (c) no sexual abuse (NSA; *n* = 43). Selected CS variables were hypothesized to be associated with sexual abuse. The Trauma Content Index (*TC/R*) incorporates several CS variables to form an index (Kamphuis et al., 2000). Although this is not a CS index, it was developed as a ratio of certain CS responses (e.g., blood, anatomy, sex, morbid, and aggressive movement) to the total number of responses.[7]

The second CS variable used by Kamphuis et al. (2000) was Gacono and Meloy's (1994) Aggressive Past variable (*AgPast*), which is a non-CS variable. It is included in this article review for comparison purposes to aggregate and individual CS variables. The authors hypothesized that *AgPast* would be associated with past sexual abuse based on the observation that themes of aggression are seen more frequently in Rorschach responses of trauma victims. The authors intended to explore the construct validity of the *TC/R* and *AgPast* as an indicator of past sexual trauma in forensic and clinical settings. In addition to construct validity, the authors also addressed discriminate validity by examining two Exner (1995) variables that were predicted to have no association with past sexual trauma: the science content score (*Sc*) and the special score for personalized responses (*PER*).

Kamphuis et al. (2000) drew participants from assessment and therapy case files from 1992 to 1996 at a treatment center in Austin, Texas. Diagnostic accuracy of the three

[7] This index has been associated with the diagnosis of PTSD in Vietnam combat veterans and traumatized civilians according to the *Diagnostic and Statistical Manual of Mental Disorders, 3rd Edition, Revised* (*DSM-III-R*; American Psychiatric Association, 1987).

experimental groups was controlled by asking nine staff members to submit only cases with valid Rorschach protocols and extensive clinical information that either confirmed or ruled out definite sexual abuse histories. To assure intercoder reliability, one of the authors re-scored all 79 protocols. Of these protocols, 28 (35%) were scored by another author, producing an excellent range of reliability (.94 to .99) for the variables under examination. Additionally, extensive detail was acquired about the definite sexual abuse group (DSA). This included age and frequency of abuse, total number of incidents, demographics of the perpetrator, and whether the abuse was violent or sadistic.

Multivariate analysis of covariance was performed comparing the three groups across two dependent variables (*TC/R* and *AgPast*), adjusting for gender. Results were statistically significant: $F(4, 146) = 3.73$, $p < .01$. As expected, DSA clients scored significantly higher on *TC/R* than did the NSA control group, $F(2, 74) = 7.0$, $p < .01$, with a large effect size (1.01). Contrary to the authors' expectations, *AgPast* showed no reliable difference between NSA and DSA. A second multivariate analysis of variance (MANOVA) compared the three experimental groups on the Exner variables (*PER* and *Sc*), which were not expected to be associated with sexual abuse. As suspected, this test of discriminate validity revealed no statistical significance.

Finally, exploratory analysis within the DSA group was performed to determine the relationship between *TC/R* and *AgPast* with the various characteristics of sexual abuse previously mentioned. A significant association was found between frequency of sexual abuse and the *TC/R* index ($r = .49$; $p < .05$), with *TC/R* scores increasing with the number of reported abuse incidents. Interestingly, there was a significant positive correlation between *AgPast* scores in the DSA group and the therapist's ratings of intensity of violence or sadism associated with the sexual abuse: (r = .51; p < .05).

The general findings of this study suggest that the *TC/R* and *AgPast* possess good construct validity in discriminating frequency and sadistic/violent qualities of sexual abuse. A classification rule of *T/CR* $\geq$ .25 produced a reasonable sensitivity and specificity, with 77% correct classification of clients with definite sexual abuse and 30% false positive classification. A rule of *TC/R* $\geq$ .30 yielded even better specificity but worse sensitivity, with 45% true positives and 16% false positives. When using the lower cutoff (*T/CR* $\geq$ .25) score to examine the false positive cases, the authors found that 31% of the clients erroneously classified as sexually abused had experienced some other form of trauma (e.g., physical abuse, medical trauma, natural disaster). The authors recognize that the small sample size in this study means that these figures should be cross-validated with larger samples before they can be applied to clinical and forensic work.

This study by Kamphuis et al. (2000) investigated the construct validity of an index of CS variables and one non-CS variable related to sexual abuse that has been reported and verified. The authors explicitly warn that their results should not be interpreted as detection of past sexual abuse when it has not been reported. As with many tests, unfounded claims using Rorschach results may exist that violate the rules of interpretation, but this is not considered a commonplace occurrence with the CS (Plake & Impara, 2001). Nevertheless, regarding detection of past sexual abuse, opponents allege that the Rorschach is "frequently used for this purpose" (Lohr et al., 2002, p. 5). In reply, proponents have stated that they are unaware of any CS studies or practitioners claiming the ability to detect past sexual abuse (Weiner et al., 2002). Anecdotally, leading CS researchers claim that they do not know of a single instance in which a Rorschach teacher or scholar has recommended using the CS to learn whether a child has been sexually abused (Weiner et al.).

Baity and Hilsenroth (1999) investigated the construct validity of six Rorschach variables of aggression commonly associated with psychopathology found in forensic populations. They hypothesized that five Rorschach aggression variables (*A1*, *A2*, *MOR,* *AgC*, and *AgPast*) would be significantly related to one another, to the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed. [DSM-IV]; American Psychiatric Association, 1994) Cluster B personality disorder criteria, and to self-report measures of anger, aggression, and antisocial behavior. This hypothesis was based on the rationale articulated by Gacono and Meloy (1994): "…when aggressive impulses produce intrapsychic tension . . . they are more likely to be articulated on the Rorschach" (p. 270).

The Rorschach literature reveals a lack of agreement on the application of standard measures or variables for assessing aggression, resulting in the development of several non-CS aggression related variables. Although construct validity of Exner's aggression variables (*AG* and *MOR*) is the focus of this literature review, construct validity of non-CS aggression variables are also relevant in this case insofar as they improve the soundness of the CS variables.

The first set of variables (*A1* and *A2*) was first introduced in 1977 prior to the CS. Primary process aggression (*A1*) is defined as intense, overwhelming, murderous, or palpably sadomasochistic aggression. Secondary process aggression (*A2*) revolves around hostility or aggression that is more socially tolerated and usually non-lethal. The second set of aggression variables was Exner's (1993) aggressive movement (*AG*) and morbid content (*MOR*), which are the most widely used variables in Rorschach aggression research. *AG* is defined as any movement response in which the action is clearly aggressive. *MOR* is defined as the identification of an object as dead, destroyed, ruined, spoiled, damaged, injured, or broken. It also includes the attribution of dysphoria to an object. Gacono and Meloy (1994), who have extensively researched forensic utility of the Rorschach, developed the last set of variables used in this study. Aggressive content (*AgC*) is defined as any content popularly perceived as predatory, dangerous, malevolent, injurious, or harmful. Aggressive past (*AgPast*) is defined as any response in which an aggressive act has occurred or the object has been the target of aggression.

All the participants used in this study were taken from an archival search of files at a university-based outpatient clinic, which was accomplished by an exhaustive search of about 800 cases seen at the clinic over a 7-year period. Data collection occurred in three phases. In the first phase, 217 out of 800 patients were initially identified as having a personality disorder as diagnosed by a clinical team. The second phase included rating the 217 cases for the presence or absence of a *DSM-IV* personality disorder diagnosis by four advanced level doctoral students in an American Psychological Association (APA) approved clinical program. These ratings were based on a review of patient records that included evaluation reports, session notes from the assessment, session notes from the first 12 weeks of therapy, and a 3-month treatment review. Rorschach data was not available to the raters during these reviews.

In the third stage, the authors were concerned with establishing good interrater agreement. Thirty-one cases were randomly selected from the pool of 217 cases. A kappa coefficient of .90 was obtained for the presence or absence of a *DSM-IV* personality disorder diagnosis. Results revealed 91 cases that met this diagnostic criterion, of which 65 met the criteria for Cluster B personality disorders according to the following breakdown: Antisocial (ANPD) = 20, Borderline (BPD) = 25, Histrionic (HPD) = 5, and Narcissistic (NPD) = 15. These 91 cases were rated again for the presence or absence of *DSM-IV* criteria for a Cluster B personality disorder using the same methods in the second phase. Interrater agreement was calculated again using a randomly selected sample of 25 patient records. Kappa coefficients were calculated individually for each of the four Cluster B personality disorders with the following results: ANPD = .86, BPD = .80, HPD = .90, NPD = .90. These values were considered good to excellent. Results for Exner's CS variables revealed the following: *MOR* fell in the excellent range (> .74; *MOR* = .79), and *AG* fell in the average to good range ($\geq$ .60 - .74; *AG* = .64). Individuals diagnosed with BPD had the highest mean response for the *MOR* variable (2.70). The *MOR* variable is related to BPD, using stepwise regression, by significantly predicting the total number of criteria for BPD. It is the only aggression variable in this study to demonstrate this ability.

The authors found that all aggression variables in this study were significantly related to one another, including the CS variables, *AG* and *MOR*. Of particular interest, *AG* was highly related to *A2* and *AgC*[8], and *MOR* was significantly related to *AgPast* and

[8] *r* = .38 and .47, respectively.

A1.[9] To investigate the relation of the six aggression variables with criteria for a *DSM-IV* personality disorder, stepwise regression analyses were performed. The Rorschach aggression variables served as independent predictors, while the total number of criteria for each Cluster B personality disorder was used as a criterion variable. These results indicate that the aggression variables in this study are not interchangeable, and that there is an apparent uniqueness to each of the individual scores.

The results of this study revealed adequate construct validity for the Exner variables (*MOR* and *AG*). The authors demonstrated that all six Rorschach aggression variables could be scored reliably, and that these variables are related to one another in significant ways. Additionally, some of these aggression variables are related to the total number of *DSM-IV* criteria for two Cluster B personality disorders (ANPD and BPD). Factor analysis of the six variables formed two distinct factors, and these factors accounted for 77% of the total variance. Factor I appears to represent aggression at objects that might indicate a more primitive level of organization associated with more intense aggression. These Factor I responses (*MOR, AgPast,* and *A1*) may help indicate how the individual experiences and interacts with the outside world. Factor II (*AG, AgC,* and *A2*) appears to represent a higher-level, less primitive aggression that is more socially tolerable. Baity and Hilsenroth (1999) have built upon past empirical studies exploring the construct validity of different Rorschach aggression variables (Exner, 1993; Fowler et al., 1995; Gacono & Meloy, 1994). Their research appears to be an important step in understanding aggression as it relates to character traits common in forensic populations.

[9] r = .79 and .63, respectively.

This article is one of the few that comprehensively examines all aggression variables in one study.

A critique of this study reveals that the authors demonstrated adequate diligence in utilizing examiners that were well trained in the CS. Adequate scoring accuracy and interrater reliability contributed significantly to a baseline agreement level greater than a kappa coefficient of .80, which is a very respectable level. Although review of this article has focused primarily on the construct validity of the two CS variables of aggression (*MOR* and *AG*), the overall utility of all the aggression variables indicates even better utility in the sense that none of them are significantly related to each other and are not interchangeable. Assuming that the Rorschach meets legal admissibility standards, it appears that this study contributes uniquely to the assessment of aggression for forensic patients in a manner that compliments the MMPI-2. Further research on aggression should continue to integrate the CS and non-CS aggression variables to improve the construct validity of all six variables individually and in combination with each other.

Reliability

Statistical reliability is the consistency of a set of measurements that are used to describe a test. Reliability is the quality of measurement in terms of the consistency or repeatability of a test. It has to do with precision, whereas validity is a measure of accuracy. The less measurement error, the better the reliability. There are various types of reliability, and each type estimates the reliability of a test in different ways as demonstrated in the articles reviewed below.

Interrater reliability. According to Acklin et al. (2000), interrater reliability refers to the degree measurement error is absent from the data. Lower measurement error is associated with higher data consistency. When considering observers as a source of error, data consistency is assessed by measuring interrater reliability. Interrater (also referred to as intercoder or interobserver) reliability is a main area of controversy, as reliability of diagnostic tests is prominent in psychometric theory (Acklin et al.).

Researchers investigating interrater reliability must choose the type of statistic to use and the level at which the analysis should focus. Potential statistics used in past psychological research include: (a) percentage of exact agreement; (b) percentage of exact agreement determined only when at least one rater assigns a score, not counting agreement on the absence of a score; (c) measures of association, such as the Pearson correlation; (d) measures of "chance-corrected" agreement, such as Cohen's kappa (κ), and (e) the Intraclass Correlation Coefficient (ICC; Meyer, 1999a).

Cohen's κ is considered the method of choice in current CS research on interrater reliability at the response level (Acklin et al., 2000). It defines chance agreement by the relative frequency, or base rate, with which each rater assigns each score option. Cohen's κ is determined by multiplying each scorer's base rate for a score option and summing the product across all options in the score category.[10] Cohen's κ is often used with the ICC to take into account behavior prevalence and chance agreement and give credit for similar behaviors that are not in strict agreement (Meyer, 1999a).

[10] The formula is κ = *(observed agreement – chance agreement)/(1 – chance agreement)*.

Rorschach critics have alleged that the original CS interrater reliability data is insufficient because it is based on percentage agreement versus the more sophisticated Cohen's κ or ICC (Wood et al., 1996). This appears to be a legitimate criticism, and some Rorschach researchers have recognized problems with inconsistent research methodology. Acklin et al. (2000) attempted to address this weakness by investigating methods of CS reliability research using the latest principles of observational methodology. They hypothesized that CS codes, coding decisions, and summary scores would yield at least acceptable levels of reliability using Cohen's κ and the ICC. As of the date of this study, a standard approach for assessing the reliability of the Rorschach CS has yet to be established.

Acklin et al. (2000) examined a previous non-patient sample of 20 protocols (*n* = 412 responses) randomly selected from a larger sample of protocols provided by students at a Midwestern university from 1987 to 1989. A new clinical sample of 20 protocols (*n* = 374 responses) diagnosed with certain research diagnostic criteria (RDC) were randomly selected from a larger sample of protocols obtained from psychiatric inpatients at a 56-bed general hospital over a 4-year period. RDC were based solely on chart review independent of Rorschach data consisting of: schizophrenia, unspecified functional psychosis, bipolar mania, intermittent depressive disorder, major depression, minor depression, labile personality, drug use disorder, antisocial personality, and other psychiatric disorders.

Participants were assessed by graduate clinical psychology students trained in the CS by a leading Rorschach researcher, Marvin Acklin, PhD (Acklin et al., 2000), who also provided supervision. Claude J. McDowell II, PhD and a graduate student with advanced training in the CS independently re-scored all 40 protocols. Both possessed at least three years of experience in using the CS.

Acklin et al. (2000) admitted that organization of CS data for reliability analysis was highly complex, and they did not have the advantage of any previous well-designed approaches. Thus, the authors developed a new approach for evaluating CS reliability in order to improve dialogue among Rorschach researchers. They conducted a highly sophisticated analysis of multiple levels of CS data, including summary-level and response-level codes and coding decisions, ratios, percentages, and derivations from the CS Structural Summary.

The authors used a criterion-referenced measurement method, which tends to be more stringent than a norm-referenced measurement method. The criterion-referenced method tends to yield lower interrater reliability data due to inclusion of examiners as a source of systematic observation error. Thus, the authors expected lower than normal interrater reliability on the response level. Acklin et al. (2000) adopted Meyer's (1997) rationale that reliability of data representing summary scores is at least as important as data representing response-level scores; therefore, both scores are calculated to determine overall interrater reliability. Cohen's κ was used for response-level scores, and ICC was the method of choice for assessing summary scores.

For the non-patient sample, Cohen's κ was interpreted as the proportion of possible agreement that is achieved by raters beyond chance agreement. Values ranged from -1.0 to 1.0. Scores greater than zero indicated that raters agree more often than predicted by chance, a value of zero indicated that interrater agreement was no better than chance, and a negative value indicated that raters agree less often than predicted by chance. For the clinical sample, the ICC was interpreted as the proportion of total variance in observers' ratings that is attributable to true variation among target verbal behaviors. Values ranged from 0 to 1.0, with 1.0 indicating perfect observer agreement.

Results for the non-patient sample indicated that well-trained and experienced raters could apply the majority of CS codes consistently. Of the 88 individual codes and coding decisions that met base-rate inclusion criteria, κ values revealed the following: 41% demonstrated excellent reliability ($\kappa \geq .81$), 36% demonstrated substantial reliability ($.61 < \kappa < .81$), and 25% demonstrated unacceptable reliability ($\kappa < .61$). Mean and median κ values in this sample were in the upper range of substantial reliability across all codes combined, across determinant codes exclusively, and across content codes exclusively. Special Scores and other low frequency scores were expectedly more problematic.

Results for the clinical sample also indicated that well-trained and experienced raters could apply the majority of CS codes consistently. The results reveal that of the 89 individual codes and coding decisions that met base-rate inclusion criteria: 47% demonstrated excellent reliability, 44% demonstrated substantial reliability, and 9% demonstrated unacceptable reliability.[11] Mean and median κ values were in the upper range of substantial reliability across all codes combined, as well as across determinant

[11] Individual codes and coding decisions in the unacceptable range included *ALOG*, Deviant Verbalization, *DR2, DV1, FQu, Hh, FY,* and *Y*.

codes exclusively. Mean and median κ values were in the middle range of substantial reliability across Special Scores exclusively, and they demonstrated excellent reliability across content codes exclusively.

Regarding protocol-level reliability for the non-patient sample, results also indicated that well-trained and experienced raters produced consistent CS protocols across the majority of aggregate codes, percentages, ratios, and derivations from the Structural Summary. Of the 82 variables that met base-rate inclusion criteria in this sample, ICC values revealed the following: 55% demonstrated excellent reliability (ICC ≥ .81), 29% demonstrated substantial reliability (.61 ≤ ICC < .81), and 16% demonstrated unacceptable reliability (ICC < .61).[12] Mean ICC values in the non-patient sample fell within the upper range of substantial reliability (.780), and the median ICC values fell within the range of excellent reliability (.825). In the clinical sample, 85 variables met the base-rate inclusion criteria, with ICC values in the following ranges: 62% demonstrated excellent reliability (ICC ≥ .81), 28% demonstrated substantial reliability (.61 < ICC < .81), and 9% demonstrated unacceptable reliability (ICC < .61). Mean and median ICC values in the clinical sample fell within the range of excellent reliability.

Acklin et al. (2000) recognize this study as a first step towards standardizing reliability research, and that much larger replication studies are encouraged for the purpose of "…enhancing the Rorschach's validity as a clinical and research

[12] Variables that fell within the unacceptable range of reliability included *AdjD, DQv/+, F+%, DV1, FAB2, FC:CF+C,* Level 2, *S-%, SCON, SCZI,* s*T*, Sum *V*, and *Xu%.*

measure" (p. 43). Although this study reveals some legitimate concern over the reliability of some variables, the authors do not interpret this as proof of unacceptable interrater reliability of the CS as a whole. They certainly do not consider these weaknesses as grounds for a moratorium on the Rorschach that opponents have called for. The authors state:

> We believe that this study provides strong evidence for the reliability of the Rorschach Inkblot Test across multiple levels of Comprehensive System data. Furthermore, these results are consistent with conclusions drawn from the majority of previously reported reliability studies for the Rorschach Comprehensive System. (p. 43)

The results of this study appear to be consistent with meta-analytic reviews and studies with patient and non-patient samples that have yielded Kappa values ranging from .79 to .88 across various CS coding categories, which is considered excellent (Acklin et al., 2000; Plake & Impara, 2001). Similar results have been found for the ICC. Meyer et al. (2002) found mean and median interrater coefficients of .92 and .90, respectively, for 164 structural summary variables in two independent ratings of 219 protocols containing 4,7611 responses.

A critique of the study reveals that there are a number of CS variables that do not meet acceptable levels of reliability. A possible factor that may have lowered reliability was a low prevalence of some variables in the majority of CS data. For example, the average base rates of response-level codes and coding decisions were 10.4% for the non-patient sample and 10.8% for the clinical sample. Only 13% and 12% of non-patient and clinical codes and coding decisions, respectively, occurred at a base rate of 20% or greater. Relatively small sample size may have also contributed to error for both non-patient and clinical samples.

Despite these shortcomings, there are also several significant outcomes of this study that appear to strengthen CS interrater reliability. Under the stringent methods of this study, a significant majority of CS variables demonstrated substantial to excellent reliability. Compared to the non-patient sample, the number of variables in the clinical sample that fell in the unacceptable range was relatively low. This suggests that interrater reliability is better for clinical norms, and future research should focus on the potential problems with non-patient populations. Regarding statistical procedures, the authors appear to make a strong case for replicating their methodology for assessing CS reliability in future studies. The variables that demonstrated unacceptable reliability should be the focus of future research to either improve the reliability of these variables or exclude them from the CS.

Shaffer, Erdberg, and Meyer (2007) investigated the ability of the CS to demonstrate minimally acceptable interrater reliability in a non-patient population. This study was in response to criticism of CS adult non-patient norms in the U.S. This study did not specifically address any known CS cross-cultural sensitivity issues; nonetheless, they expected to explore any significant culture-specific results revealed by this study. The data is the result of their 3-year normative study that attempted to replicate U.S. demographics as of 1996: Caucasians comprised 72.1 % of the population; African-Americans 12.6%; Hispanic-Americans 10.7%; and Asian-Americans 3.7%. The data was subjected to an extensive coding review process according to the most recent version of the CS.

The authors presented this recent study as an update on their previously reported normative data (Shaffer, Erdberg, & Haroian, 1999) on adult non-patients living in central California gathered from 1994 to 1995. Data from the past study revealed the performance of 123 non-patients on the MMPI-2 and the Rorschach. In the newer study, Shaffer et al. (2007) used identical procedures on an additional 160 non-patients who were administered the Rorschach for a substantial total sample size of 283 protocols. This sample was mostly representative of individuals functioning at a high level of mental health.

In an effort to ensure good interrater reliability for the newer study, the authors chose examiners who had successfully completed an advanced Rorschach assessment course and a one-year assessment practicum. This included instruction in the administration, coding, and interpretation of the Rorschach using the CS. Additionally, many examiners completed unspecified advanced training in the Rorschach, and all examiners had at least weekly access to supervisory assistance from senior students and the authors.

The sample from the previous study consisted of 123 non-patients (Shaffer et al., 1999), and 52 records were randomly selected from this sample to determine interrater reliability. The sample from the newer study consisted of 160 protocols, and interrater reliability findings for this sample were obtained from 40 randomly selected protocols. The combined total of randomly selected protocols contained a total of 1,874 responses.

Percent agreement and kappa were calculated for all CS variables at the response level. Additionally, iota was calculated for all CS variables at the protocol level. (Iota is a chance corrected reliability coefficient that is equivalent to Cohen's κ, but can be computed for multivariate data.)

Results (percent agreement and iota) were presented for the basic scoring segments of each Rorschach response. Percentage of agreement, with a range of .74 to .98 for a number of responses, was in the moderate to excellent range. The iota values ranged from .57 (within the moderate range of agreement) to .89 (within the excellent range) for the same response categories. In summary, the percentage agreement for major response level categories was impressive. Iota for the same category was expectedly less, with one particularly lower value of iota for special scores (.57).

A critique of this study reveals several concerns regarding whether it is possible to generalize these results with confidence. First, all participants were from a single geographic area of the United States. Second, African-Americans were underrepresented in relation to the U.S. census data (3.5% versus 12.2%). The authors attempted to address these potential limitations by anchoring the findings to normative data from two other widely used assessment instruments, the WAIS-R and the MMPI-2. The similarity of the data on these two instruments, when compared to the data from this study, suggests that the results of this study are typical. However, some caution should be used in generalizing these results to the U.S. population.

Test-retest reliability. Test-retest reliability, also referred to as stability in this review, is foundational in psychological assessment procedures. It is the ability of a test to measure a trait that is supposed to be stable over time (Plake & Impara, 2001). Rorschach opponents have stated that the stability of the CS is not established based on the fact that published retest correlations are available for only a portion of CS variables (Garb, Wood, Nezworski, Grove, & Stejskal, 2001). Proponents have responded by pointing out that the alleged lack of data involves either composite variables for which data are available for their component parts, or individual variables that occur too infrequently to allow for meaningful test-retest research (Viglione & Hilsenroth, 2001).

According to Sultan, Andronikof, Réveilléré and Lemmel (2006), large short-term reliability correlations are necessary to determine acceptable levels of stability. Long-term stability is expected to be related to stable personality characteristics. Based on previous CS research on stability, the authors state that reasonable expectations for stability should be in the .70 to .80 range for most interpretively significant variables. When the CS is compared to other existing personality tests, the expected range of stability correlations for intermediate retest intervals should be in the .60 to .70 range. This is based on a previous meta-analysis of 23 studies over 13 years (Sultan et al.).

Traditionally, the Rorschach literature has interpreted low stability levels as reflecting state characteristics, whereas high stability levels have been assumed to reflect trait-like features. Sultan et al. (2006) investigated stability based on the following assumptions gained from previous studies on stability. First, almost all the CS variables that supposedly relate to trait characteristics have exhibited substantial stability in adults both in the short and long term. Second, lower stability is associated primarily with the inanimate movement variable (*m*) and diffuse shading variable (*Y*), which are considered to be state related. The authors explored the possibility that external self-report measures of state or trait characteristics are capable of moderating Rorschach stability levels. This was the first study of its kind to use this strategy.

Sultan et al. (2006) presented the following hypotheses for their study: (a) moderate to high levels of stability were expected for most Rorschach variables - a high level of stability over intermediate levels was defined as exceeding .70, and a moderate level as exceeding .50; (b) stability levels were expected to be higher for personality, cognitive or self/relational construct-related variables[13] than for emotional, coping or state-related variables;[14] (c) discrepancies between baseline test (T1) and retest (T2) were expected to be related to instability. They hypothesized that changes in distress could account for "error variance" in state-related emotional variables. These changes in distress would be measured by an external criterion, the General Health Questionnaire (GHQ-12); (d) task engagement (TE), which measures the number and complexity of Rorschach responses, was expected to moderate stability such that controlling for TE would increase stability in CS variables known to be related to this factor; (e) large variations in TE were expected to be related to lower stability levels for state-related variables and negative emotion markers; and (f) it was assumed that interaction and relational dynamics with the examiner would be related to TE, and that some of the instability could also be due to effects related to the examiner's administration of the test.

[13] *M*, *a*, *EA*, *EGO*.

[14] *m*, *Y*, *D*, shadings.

The authors chose an intermediate time interval (3 months) for retest, with the empirically based assumption that this time interval permits changes in some of the CS constructs measured. Participants consisted of 75 non-patient individuals recruited from an ongoing French-language normative project. They were tested twice between November 2001 and March 2002. As is normal in CS research, all the protocols used in the study contained at least 14 responses. Twelve examiners participated in this study. No examiner tested the same person twice, which is a procedure consistent with previous research. All the examiners were clinical psychologists (which is the equivalent to a Master of Arts degree in France) and had previously been trained in the CS with the training equivalent of Rorschach Workshop Level I and II.

Variables used in this study were divided into two categories. The first set of variables was related to personality, cognitive, and self or relational variables.[15] The second set of variables was related to emotional, coping, or *m* and *Y* influenced variables.[16] Additional variables included were the total number of positive *DEPI* and *S-CON* scores, which were also used as markers of negative emotions. Finally, GHQ-12 scores were used as an external criterion for measuring distress. The GHQ-12 is a 12-item self-report instrument that measures aspects of psychological distress and social dysfunction. It is intended to be a screening measure for mental disorders in the community and in non-psychiatric clinical settings.

[15] *R*, *Zf*, *F*, *M*, *a*, *p*, *W*Sum*C*, *L*, *EA*, *EGO*, *W*Sum*6*.

[16] *P*, *FM*, *m*, *FC*, Sum *T*, Sum *Y*, Sum *V*, *Adj es*, *D*, *Afr*.

The first priority in this study was to establish interrater reliability as a necessary condition for obtaining accurate stability. To ensure the best interrater reliability, the area coordinator scored all protocols, which were then re-scored blind by an independent rater. These scores were compared, and each area coordinator then decided on the adoption of the final scores. Of the 150 protocols for T1 and T2, 25% (40 protocols = 20 tests and 20 retests) were randomly selected and re-scored independently by one of three other psychologists who were blind to the initial consensus scoring. The total number of responses in the 40 protocols was 1,027.

Interrater agreement was calculated at the protocol level of summary scores using the exact agreement for a single-rater ICC according to a one-way random effects model. Results indicated an ICC mean and median of .86 and .89 respectively. The authors concluded that the ICC was reasonable for *C* (ICC = .40 to .59) and good for *CF*, *MOR*, and *X-%* (ICC = .60 to .74). For all other variables, it was excellent (ICC = .75 to 1.00). Regarding TE, results generally indicated that respondents were less puzzled by the task at T2 and became involved in the task more easily (e.g., better form quality, lower special scores).

Results revealed that among the 47 variables studied, 9 had stability coefficients above .70 (high stability), and 21 had coefficients above .50 (moderate stability). However, the overall level of stability across all variables was somewhat lower than expected (< .69). The correlations for most variables were approximately .10 to .15 lower than those reported by Exner (2003). Results also revealed that the majority of the categories defined by cutoff scores were unstable. Additionally, several index scores

(*DEPI* and *CDI*) did not produce evidence of stability when they were initially positive. This reveals that the Rorschach stability was lower than expected in the first hypothesis. The authors were careful to mention that, due to the wide range of stability levels for variables within the CS, variables should be evaluated individually rather than judging the Rorschach as a whole.[17]

The authors offer several cautions in interpreting the data. Although interrater reliability was generally excellent, relatively low *C* and *CF* values (ICC = .45 and .65 respectively) may be responsible for some of the error variance between test and retest. Infrequent codes may also compromise stability coefficients. For example, in Exner's reference sample, 74% of respondents had at least one texture response, whereas the proportion for texture in this study was only 55%. The authors also explain that the external state measure (GHQ-12) was designed as a screening test, and was not sensitive or specific enough to detect important aspects of state variance. Finally, this study focused on emotional states only, and not on other states and processes measured by the Rorschach. The authors highly recommend that larger sample sizes be used in future studies.

Sultan et al. (2006) presented their stability data with the assistance of expert CS researchers Gregory Meyer and Donald Viglione. Compared to Exner's 2003 published norms, results of this study do not appear to significantly strengthen the CS as a whole. However, even the weakest variables (*C* and *CF*) are not considered unacceptable by the

[17] For example, *R*, *Zf*, *F*, *M*, *S*, *Pair*, *Lambda*, *EA*, and *EGO* exhibited acceptable stability compared to other instruments.

authors when compared to meta-analyses and other empirical studies. The results of this study are generally consistent with other stability research studies revealing retest correlations of at least .75, with 19 core variables that have 1-year and 3-year retest correlations of .85 or higher (Viglione & Hilsenroth, 2001). These statistics have been evaluated in the most recent edition of the *Fourteenth Mental Measurements Yearbook* (Plake & Impara, 2001), which states, "the test-retest reliability of the Comprehensive System is impressive" (p. 1034).

Standardized Procedures

The CS provides detailed and specific standardized administration procedures for the Rorschach. All psychometric instruments must clearly articulate standardized procedures to ensure uniformity of administration in research and clinical settings. This is especially true for the Rorschach due to the highly interactive nature of the testing task compared to self-report measures.

A common criticism of the Rorschach is often focused on the poor validity obtained when respondents offer too few responses according to CS guidelines. In the most recent edition of the *Fourteenth Mental Measurements Yearbook*, Plake and Impara (2001) point out that a person providing 14 responses will naturally tend to have more whole responses than a person who hypothetically gives 33 responses, potentially confounding results across subjects. Brief records, which are more common for inpatient and forensic populations, lack sensitivity and negative predictive power with reduced interpretive yield. If brief records cannot provide incremental validity over other assessment methods, then it lacks clinical utility (Dean et al., 2007). This is a well-known problem for the Rorschach, and it calls into question the adequacy of current CS standardized procedures used to regulate response range.

Standardized procedures are potentially problematic for the Rorschach due to the interactional task of examiner and examinee. Rorschach researchers and clinicians have long expressed concerns about the validity of brief and lengthy records. However, no recent study has addressed the specific relationship between protocol length and the validity of the Rorschach in assessing pathology. Dean et al. (2007) observed that brief records are typically less informative and reliable than longer records due to low sensitivity and susceptibility to high false negative rates. On the other hand, excessively long protocols are typically vulnerable to skewed data, scoring inaccuracy, and misinterpretation.

Dean et al. (2007) propose that there is an optimal response range that is likely to be associated with increased response complexity, leading to an interpretive yield that is less likely to be either impoverished or excessively cumbersome.[18] This study was designed to examine the effects of protocol length and complexity on Rorschach validity. The authors predicted that restricting the response range by decreasing the number of excessively short and long protocols would improve the validity of the Rorschach. They introduced a nonintrusive method for constraining responses by prompting for an extra response when only one is offered per card. The new method also includes removing the

[18] Rorschach response complexity is described in terms of cognitive flexibility, sophistication, and problems solving skills. Viglione (1999) defines complexity as the amount of productivity, precision, differentiation, and integration involved in the aggregate of all Rorschach responses.

card after four responses are given. The authors considered complexity to be associated with an optimal number of responses that fall in the optimal range of 18 to 28 responses. This alternative administration method was intended to increase *R* in records that were likely to be too short, and decrease *R* in records that were likely to be excessively long.

Dean et al. (2007) hypothesized that providing patients with extra response prompts would increase protocol length and produce a more complex interpretive yield. They predicted that this could be accomplished while preserving the validity of the CS. The authors specifically investigated the relationship between their new administration method and the prediction of thought disorder criteria. They hypothesized that validity for predicting thought disorder criteria would be maximized with protocols containing 18 to 28 responses. Finally, they predicted that when response complexity (e.g., multiple determinants versus only one) was considered as a moderator variable in exploratory analyses, the association between Rorschach variables and thought disorder scores would improve.

Three Rorschach variables were used to measure thought disorder: the Schizophrenia Index (*SCZI*), the Perceptual Thinking Index (*PTI*), and the Ego Impairment Index-2 (*EII-2*). Although these three variables contain some of the same CS variables, the authors believed that each index might be unique enough to reveal different facets of thought disorder. Participants in this study included 61 adults ages 18 to 66, with a mean age of 37 and an average of 11 years of education. All patients were in long-term residential treatment at either a state psychiatric facility or a state prison, and all carried an Axis I diagnosis according to the *DSM-IV* (APA, 1994). The participants were randomly assigned to one of two groups. There were no differences between the two groups in mean age[19] or education.[20]

The standard administration group (*n* = 31) received the traditional CS administration (Exner, 1993). If necessary, participants were given a single prompt for more responses on Card I only. The experimental group (*n* = 31) received the alternative administration, which involved prompting participants for another response whenever only one response was provided to a card. Examiners gave this prompt up to three times per card, if necessary, except to cards V and IX. If participants provided only one response to a card after the three prompts were offered, no additional prompts were given for that card. If the participant provided only one response to any card except cards V and IX, the examiner said, “Take your time and look some more. I’m sure you’ll find something else too” (Exner, 1995, p. 6). The alternative administration also attempted to reduce lengthy protocols by allowing only four responses per card for all cards. After the respondent provided the fourth response to any card, the examiner removed the card.

The alternative administration method prevented the need to re-administer the test due to fewer than 14 responses, which effectively eliminated the chance for human error introduced by re-administration. This is in contrast to the commonly held concern that changing administration procedures would compromise test validity. In the combined sample (standard administration group and alternative administration group), the combination of Rorschach variables was significantly associated with thought disorder

[19] Standard group *M* = 36.58, *SD* = 9.79; alternative group *M* = 37.17, *SD* = 9.99.

[20] Standard group *M* = 11.58, *SD* = 2.32; alternative group *M* = 10.70, *SD* = 2.14.

scores (Adjusted R = .43; p < .01). However, when analyzing the standard group, the three Rorschach variables did not reach significance in predicting thought disorder (Adjusted R = .27; p < .05). In the alternative administration group, the same three variables revealed significant improvement in ability to predict thought disorder (Adjusted R = .52; p < .01).

Dean et al. (2007) expected that the alternative administration would significantly increase the number of protocols that fall within the optimal 18 to 28 response range. The results were not statistically significant;[21] however, an examination of frequencies indicated a trend toward significance: F(29,30) = 1.32, p = -.230. The alternative group had 20 protocols of 18 or more responses, versus 11 with fewer than 18. The standard group had 13 protocols with 18 or more responses, and 17 protocols with fewer than 18. The authors concluded that the effect of the alternative administration on protocol length might reach significance with a larger sample.

The authors also expected the alternative group to yield significantly fewer brief protocols (R < 14) than the standard group. Of the 31 protocols in the experimental group, 21 were given more than one prompt. On average, these respondents produced 3.24 more responses during the initial response phase, with a mode of 17 responses, compared with a mode of 12 in the standard group. These results were statistically significant. The alternative administration procedures also significantly reduced

[21] First hypothesis results according to a nonparametric 2 X 2 Pearson chi-square analysis with a one-tailed Fischer's Exact Test: χ^2 (1, n = 61) = 2.755, p = .080; r = .18.

administration time by eliminating the need for protocol re-administration due to an initial $R < 14$ ($r = .33$).

Finally, the authors predicted that when complexity was considered as a moderator variable in exploratory analyses, the association between Rorschach variables and thought disorder scores would improve. An analysis of the alternative administration group revealed significantly higher validity than the standard administration group for the *PTI*. When the *PTI* was the sole predictor,[22] it contributed significantly to the prediction of thought disorder by increasing the Adjusted *R* from .381 to .457. These figures support the hypothesis that complexity is a moderating variable in Rorschach validity.

Critique of this study reveals a significant potential improvement in Rorschach standardized procedure. Specifically, examiners can use the alternative administration procedure when giving the Rorschach to an individual at risk for producing a brief protocol. Prompting for extra responses when only one is given can improve test validity and significantly reduce the possibility of re-administration due to fewer than 14 responses. In this study, the alternative method eliminated the need for protocol re-administration (effect size $r = .33$). In contrast, 23% of the respondents in the standard group produced fewer than 14 responses, requiring re-administration according to CS guidelines. The issue of excessively long protocols was not as thoroughly addressed in this study. (For the new administration group, removing cards after four responses only occurred on four protocols.) It is a relevant issue, as the literature reveals that

[22] The complexity beta weight was .253 ($p = .034$).

excessively long protocols are prone to lack of specificity, poor positive predictive power, scoring error, and consumption of time to administer and score.

The clinical utility of these results are significant. Rorschach clinicians often encounter problems with re-administration in terms of time requirements, frustration of both examiner and examinee, and the questionable validity of results when the response phase is repeated and the respondent does not offer the original responses again. This automatically increases the possibility of interpretive error, which is a common criticism of the Rorschach. This study convincingly demonstrated that additional data obtained by eliciting extra responses significantly enhanced the validity of the Rorschach. This appears to challenge the common belief that prompting for additional responses reveals information that is less meaningful or representative of the respondent.

The authors conclude that the new administration procedure did not reinforce types of answers, but rather encouraged valid responses when the respondent failed to extrapolate from the initial prompt. Dean et al. (2007) recommend that future research of the alternative method should focus on improving the validity and clinical utility of the Rorschach while reducing cost and time investment with respondents who typically produce brief and lengthy protocols. It appears that expansion of this study and implementation of this new administration procedure could significantly improve the Rorschach's overall psychometric soundness.

Normative Data

Non-patient norms remain a paramount concern in validation research when attempting to compare base rates of pathology in various treatment groups with the general population. Rorschach opponents have alleged that the currently available normative data for the CS are outdated and likely to overpathologize by indicating the presence of pathology in supposedly non-patient individuals (Wood et al., 2001). Proponents have responded by stating that the research findings cited by opponents were methodologically compromised by small samples, inadequate systematic research procedures, and scoring inaccuracy introduced by inexperienced examiners (Weiner, 2001a). A review of literature reveals that there is indeed little if any substantial empirical studies that expose a serious and consistent problem with overpathologizing in normative populations.

The more poignant debate in the current controversy is centered on criticism of Exner's CS norms for lacking cross-cultural sensitivity. Some of this criticism comes from proponents (Acklin et al., 2000), who concede that there is a lack of established norms for minorities. Specifically, there is concern on both sides of the debate about the possibility of African-Americans, Hispanics, Native-Americans, and non-Americans scoring differently on core variables (Wood & Lilienfeld, 1999). However, Acklin et al. (2000) do not agree with the opponents who assert that lack of cross-cultural sensitivity is evidence of poor psychometric soundness of the CS as a whole. They object to the assertion that lack of appropriate cross-cultural norms is tantamount to lack of clinical utility with American minorities, especially when relatively few studies have been conducted in this area. Proponents argue that the discovery of differences in cross-cultural normative data, should they exist, may reflect actual cultural differences that are being accurately measured by the Rorschach (Weiner et al., 2002).

Shaffer et al. (2007) recently introduced a special supplement in the *Journal of Personality Assessment* on international normative data research. This cooperative effort represents a dramatic expansion of international efforts to strengthen CS normative data in the United States and abroad. As part of this effort, empirical studies were conducted starting in 2002 in the following countries: Argentina, Australia, France, Greece, Israel, Italy, Japan, Netherlands, Portugal, Romania, Spain, and the United States. It is important to note that Shaffer et al. did not intend to perform empirical cross-cultural comparisons between Exner's normative data and international reference samples. Many of these studies were exploratory; nonetheless, criticism of the CS for shortcomings in cultural sensitivity appears to be reasonable. However, contrary to the opinion of Wood and Lilienfeld (1999), it also appears reasonable for proponents to conduct further research in this area as opposed to discontinuing its use in the assessment of minorities altogether.

Presley, Smith, Hilsenroth, and Exner (2001) investigated the issue of cross-cultural sensitivity with African-Americans based on the fact that the establishment of African-American norms has historically received minimal attention in CS research. They recognized that possible significant differences in patterns of responses of a normative group of African-Americans would be valuable for clinicians who may be unaware of the role of race in assessment. The rationale of this study is based on the assumption that African-Americans must adapt to and cope with a White-American worldview. This includes negotiating daily struggles with their environment due to race. The authors hypothesize that these race-related struggles would translate into some unique patterns of

Rorschach responses that are significantly different from the current CS normative data, which is based primarily on White-American demographics.

A review of the literature revealed only seven studies published on this topic in over 60 years. Most of these studies were hindered by restricted sampling and an absence of consistent Rorschach interpretive methodology (i.e., not all the studies used the CS). However, a consistent pattern of responding has emerged from these previous studies, revealing that African-Americans consistently provide fewer responses (*R*) than their White-American counterparts regardless of which interpretive system was used. Based on this observation, the authors proposed that White-Americans and African-Americans would display significant differences in their perceptions of the world, resulting in differences in Form Quality and Popular responses. The authors also proposed that semantic language differences might have the potential of reducing the reliability of the deviant verbalization (*DV*) Special Score. Clinically, new normative data would allow an examiner's awareness of African-American idioms to reduce misinterpreting some unique African-American responses as a form of cognitive slippage.

In this study, Presley et al. (2001) explored structural data from a normative sample of African-Americans included in the broader normative sample of the CS (Exner, 1993). African-American normative data were separated from the broader data to establish a separate set of norms. This is the first study to establish an African-American normative sample in the form of descriptive statistics consistent with Exner's norms. The authors predicted significant differences in response productivity (*R*) and perception (*F*+%, *P*, *X*+%, *X*-%). In addition, potential differences regarding social or cultural aggression, four primary Rorschach indexes, and other selected interpersonal and self-perception variables were explored.

The Rorschach responses of 88 participants were drawn from the normative data sample of Exner's (1993) CS. This sample consisted of 700 protocols for non-patient adults, which were broken down by gender (350 women and 350 men with a mean age of 32.36 years). Of these patients, 332 were married, 192 were single, and the remaining patients were separated, divorced, widowed, or living with a significant other. The average years of completed education were 13.25. Socioeconomic status (SES) was categorized according to upper, middle, and lower class categories. The majority of participants came from a middle class SES level. The 88 participants used for this study consisted of 44 adult non-patient African-Americans and 44 non-patient White-Americans. An analysis of variance (ANOVA) revealed no significant difference in age, SES, and education between the two samples.

The results were drawn from 23 CS variables that were chosen a priori to examine statistical differences between the African-American and White-American groups by means of univariate ANOVA. A comparison of means, standard deviations, *F* tests, effect sizes, and *p* values for the two groups were analyzed. Comparisons that reached both a statistical level of significance ($p < .05$) and represent a medium effect size ($\geq .5$) were reviewed for clinical significance. There were three statistically significant differences ($p < .05$) between groups involving white space (*S*) responses, *SCZI*, and *COP*. Specifically, the African-American group revealed significantly higher frequencies of *S* responses,[23] and higher *SCZI* scores[24] than White-Americans. However, African-Americans had significantly less *COP* than White-Americans. Effect sizes for these three variables were .47, .52, and .64 respectively.

Interpretation of these results indicates a close similarity between the African-American and White-American samples on most of the Rorschach variables. Only 3 (*S*, *SCZI*, and *COP*) of the 23 variables reached statistical significance ($p < .05$). Although a higher frequency of *S* was found for African-Americans, the obtained effect size of .47 was below the suggested level of ≥ .5. Therefore, this did not amount to a clinically meaningful difference.

Although no significant differences were found between the African-Americans and White-Americans on the remaining 21 Rorschach variables, trends toward significance were apparent for the *P*, Sum *C'*, *DEPI*, and *CDI* variables. (See Table 2.)

Table 2

Mean Scores of P, Sum C', DEPI, and CDI

	P[25]	Sum *C'*[26]	*DEPI*	*CDI*
African-Americans	6.20	1.67 (*SD*=1.21)	2.75 (*SD*=1.18)	1.16 (*SD*=1.13)
White-Americans	6.82	1.16 (*SD*=1.10)	2.25 (*SD*=1.38)	1.20 (*SD*=1.05)

[23] $F(1,87) = 4.88, p = .03$

[24] $F(1,87) = 5.83, p = .02$

[25] $F(1,87) = 3.40, p = .07$

[26] $F(1,87) = 3.40, p = .07$ for Sum *C*, *DEPI, and CDI*

A significant difference was also obtained on the *SCZI* with an effect size of .52, initially suggesting that this finding may be clinically significant. However, the difference on the *SCZI* scale was < 1, and the mean for each group was < 1 (African-American = .48, White-American = .16). This was well below the established clinical cutoff score of ≥ 4 necessary for the *SCZI* to be interpretively relevant. This result strongly suggests that this difference has no clinical utility. Similarly, there was no clinical relevance regarding the differences for the *P*, Sum *C'*, *DEPI*, and *CDI* variables.

The results concerning the frequency of *COP* responses were the most meaningful. The obtained effect size of .67 was well above the cutoff point of ≥ .5, indicating clinically significant difference. The White-American sample revealed a mean of almost two *COP* responses compared with one for the African-American sample. Compared to the African-American frequencies for *COP*, only 23% of White-Americans had a score of 0, and 45% had a score of > 2. In contrast, 43% of African-Americans had a score of 0, and only 16% had a score of > 2.

The clinical utility of these findings is consistent with Exner's (1993) study that reviewed follow up data for 100 first admission psychiatric inpatients. Exner (1993) found that 81% of the patients with 2 or more *COP* responses at discharge reported favorable progress. However, only 66% of patients who provided one *COP* response reported favorable progress. Based on the assumption that African-Americans produce significantly fewer *COP* responses than White-Americans, this data suggests that African-Americans may be less likely to perceive or anticipate positive interactions among others as commonplace events. Other than *COP* responses, African-Americans were very similar to White-Americans in terms of gender, age, education, SES, and marital status variables. The authors note that all the examiners in this study were White. The authors suggest that the low *COP* scores may be related to the anticipation of racial discrimination when faced with a White-American examiner. This may also be related to lower response frequency for African-Americans. These assumptions are not based on empirical data, but would be useful for future research.

Critique of this study reveals that relatively small sample size limited the degree to which one can draw inferences about culture differences in the general population. In addition, the authors point out that African-Americans are not a homogeneous group, and within-race differences may be as significant as between-race differences in understanding the impact of culture on mental health. Nevertheless, this is one of the few studies investigating racial differences between African-American and White-American differences on Rorschach variables. The results of this study are supported by Meyer (2002), who found no association between ethnicity and any of 188 Rorschach scores among demographically matched European-American, African-American, Hispanic-American, Asian-American, and Native-American respondents.

Chapter Three: Summary and Conclusions

This literature review has attempted to examine the current controversy over the Rorschach by evaluating a broad sample of recent empirical research from 1998 to 2008. A general summary of the articles examined is provided below in response to the question posed in the introduction: Does the Rorschach demonstrate acceptable psychometric properties according to the same scientific standards applied to other instruments? Regarding this question, the Rorschach empirical research critiqued in this review demonstrated acceptable (a) validity (incremental, convergent, and construct), (b) reliability (interrater and test-retest) (c) standardized procedures, and (d) normative data. These studies represent meaningful contributions to the extant CS literature and offer a legitimate basis for further research on the psychometric properties of the Rorschach.

The samples of CS variables investigated in this review generally demonstrate adequate to excellent psychometric soundness using current and well-designed methodology. This review does not attempt to generalize these findings to all Rorschach variables and indexes. Nevertheless, the findings appear to contradict claims that the Rorschach as a whole is based on inadequate methodology and insufficient data. Literature reviews, theoretical position papers, and meta-analytic studies involving the CS offer varying interpretations of empirical data; however, specific critique of non-empirical literature is beyond the scope of this review.

The general findings of this review are consistent with meta-analyses conducted by Meyer, Mihura, and Smith (2005), who found that the CS has good to excellent psychometric properties overall as measured by general validity, reliability, and normative data. The *Fourteenth Mental Measurements Yearbook* (Hess, Zachar, & Kramer, 2001) is a widely recognized publication that presents objective reviews of all currently available assessment instruments. The editors' conclusion also appears to be consistent with the findings of this review:

> The Rorschach, employed with the Comprehensive System, is a better personality test than its opponents are willing to acknowledge . . . Many of the recent criticisms of the Rorschach should help CS proponents improve research and interpretive fidelity regarding the Rorschach. But opponents confuse failures with flaws. (p. 1037)

In a relatively recent attempt to provide a consensus on the Rorschach, the Society of Personality Assessment (SPA) commissioned an independent blue-ribbon panel to obtain an impartial meta-analytic summary of the Rorschach validity evidence (Smith et al., 2005). The Board of Trustees of the SPA endorsed former Harvard professor Robert Rosenthal, PhD, to lead the panel. Rosenthal, a highly respected statistician, meta-analytic researcher, and methodologist, compared the validity of the Rorschach and the MMPI-2. The results, reported initially by Hiller, Rosenthal, Bornstein, Berry, and Brunell-Neuleib (1999) and confirmed by a follow up study by Rosenthal et al. (2001), revealed that the Rorschach and the MMPI-2 have validity effect sizes of "substantial magnitude": unweighted mean *r* of .29 and .30, respectively. It is noteworthy that Rosenthal had not previously conducted research on either instrument and had no personal or professional investment in the outcome. Based on the data from these studies, the SPA Board of Trustees issued a statement supporting the credibility of the Rorschach as an empirically sound instrument comparable to the MMPI-2.

However, as recently as 2006, opponents attempted to discredit the SPA's official statement by stating that Rosenthal's work was untrustworthy because it represents the work of only a few Rorschach "friends" (Weiner et al., 2002). This opinion, which appears to imply that the work of Rorschach proponents is likewise untrustworthy, was not based on scientific examination of Rosenthal's methodology or results. Rorschach proponents defended the legitimacy of the panel's results, pointing out that the SPA statement was drafted by a 14-member Board of Trustees whose stated intent was to exclude as many sources of bias as possible. Smith et al. (2005) contend that qualifications of board members should be evaluated based on substantial publications on a wide range of topics that extend well beyond Rorschach research.

Suggestions for Further Research

This literature review reveals that Rorschach scoring accuracy using the CS was cited as a major area of concern in a majority of articles reviewed. This warrants future research targeted at establishing a standard for examiner training in CS research (Acklin et al., 2000). Prior to the current controversy, Exner (1988) expressed concerned about the paucity of research on scoring accuracy. Poor scoring accuracy among inadequately trained and inexperienced examiners may jeopardize the clinical utility of the Rorschach regardless of other sound psychometric properties. He recognized that high error rates could easily lead to faulty or completely incorrect interpretive conclusions. Acklin et al. (2000) also expressed concern about scoring accuracy. The authors concluded in their study that their attempt to establish interrater reliability was not as robust as it could have been due to possible scoring errors of omission and commission.

Some studies reviewed in this paper used examiners who arguably did not have the same level of expertise as professional Rorschach clinicians. For example, Fowler et al. (2001) used advanced psychology graduate students to score protocols in control groups without cross validation of scorer accuracy. Future research should stringently control for expertise of CS examiners, who should ideally be graduates with a substantial level of postgraduate experience with the CS. This recommendation is consistent with a study on scoring accuracy (Guarnaccia, Dill, Sabatino, & Southwick, 2001), which revealed less than acceptable accuracy in their exploratory research even with good interrater reliability. Guarnaccia et al. concluded that lack of examiner training is most likely related to scoring accuracy problems due to the labor-intensive nature of the Rorschach.

Concluding Remarks

Rorschach research continues to yield promising and increasingly sophisticated data as evidenced by the sample of recent research reviewed in this paper. In addition to documenting progress in the development of CS variables, these studies also clearly identify problematic variables that demonstrate relative statistical weakness. Rather than presume that these weaknesses equate to failure of the entire instrument, it appears far more reasonable to assume that some negative evidence, especially in exploratory studies, is actually a prerequisite for the ongoing process of scientific improvement.

Positive evidence for the CS significantly outweighs the negative in this review; therefore, the claim by leading opponents that the Rorschach is “devoid” of scientific merit (Lohr et al., 2002, p. 8) does not appear to be justified. Positive or negative bias toward any instrument should not supersede scientific judgment. In the midst of the current controversy, weaknesses and shortcomings of the CS should not be used to justify an immediate moratorium without accounting for the data in support of the Rorschach generated by a considerable and established body of empirical publications. Weiner et al. (2002) suggest that it is highly unlikely that so many Rorschach assessors have been using the instrument for so long, in so many contexts, on the basis of illusory correlation. Nevertheless, a small cadre of opponents still appears determined to rid professional psychology of the Rorschach (Weiner et al., 2002). This discourages collaborative scientific dialogue and creates an oppositional relationship with many in the psychological assessment community that appears to extend beyond the normal parlance of scholarly debate.

The potential for improvement of psychometric properties of the Rorschach appears promising, but the process will continue to be laborious and time consuming given the extraordinary scrutiny of the CS in recent years. Opponents who intend to prove the Rorschach is illegitimate are under the onus of producing a significant volume of negative data to support their hypothesis. This can only be accomplished through the process of generating and replicating well-designed studies on par with current CS research before presenting credible opinions to consumers of Rorschach data. Meanwhile, educators, clinicians, and researchers should exercise due diligence in evaluating all sources of primary empirical data to minimize the effects of negative as well as positive research bias.

It is important to consider that this literature review reveals no controversy over the legitimacy of the Rorschach among a large majority of qualified and credentialed Rorschach practitioners and researchers. However, even the perception of controversy can undermine public trust in evaluators who incorporate the Rorschach in their practices, erode education and training in the CS, and diminish research interest in this unique and widely used instrument. Claims that the Rorschach is “well-nigh worthless” (Wood et al., 2003, p. 1) may continue to create negative perceptions of the test and cast doubt on its utility, but the supporting evidence for this claim appears scant. Nevertheless, consumers of Rorschach data should not discount the value of intense scrutiny that can potentially sharpen the Rorschach into a more valid and reliable instrument by exposing its flaws and strengthening its merits. Clinicians, researchers, and educators must invariably take extra care to avoid opinions that are not based on solid scientific evidence concerning the psychometric properties of the CS.

REFERENCES

Acklin, M. W., McDowell II, C. J., Verschell, M. S., & Chan, D. (2000). Interobserver agreement, intraobserver reliability, and the Rorschach Comprehensive System. *Journal of Personality Assessment*, *74*(1), 15-74.

American Psychiatric Association (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., rev.). Washington, DC: Author.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Archer, R. P. (Ed.) (1999). Perspectives on the Rorschach [Special Section]. *Psychological Assessment*, *6*(4), 307-351.

Baity, M. R., & Hilsenroth, M. J. (1999). Rorschach aggression variables: A study of reliability and validity. *Journal of Personality Assessment*, *72*(1), 93-110.

Blais, M. A., Hilsenroth, M. J., Castlebury, F., Fowler, J. C., & Baity, M. R. (2001). Predicting DSM-IV cluster B personality disorder criteria from MMPI-2 and Rorschach data: A test of incremental validity. *Journal of Personality Assessment*, *76*(1), 150-168.

Bornstein, R. F. (2001). Clinical utility of the Rorschach Inkblot Method: Reframing the debate. *Journal of Personality Assessment*, *77*(1), 39-47.

Dean, K. L., Viglione, D. J., Perry, W., & Meyer, G. J. (2007). A method to optimize the response range while maintaining Rorschach Comprehensive System validity. *Journal of Personality Assessment*, *89*(2), 149-161.

Exner, J. E. (1988). Problems with brief Rorschach protocols. *Journal of Personality Assessment*, *52*, 640-647.

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

Exner, J. E. (1995). *A Rorschach workbook for the Comprehensive System* (4th ed.). Asheville, NC: Rorschach Workshops.

Exner, J. E. (2003). *The Rorschach: A Comprehensive System: Vol. 1. Basic foundations* (4th ed.). New York: Wiley.

Fowler, J. C., Piers, C., Hilsenroth, M. J., Holdwick, D. J., & Padawer, J. R. (2001). The Rorschach suicide constellation: Assessing various degrees of lethality. *Journal of Personality Assessment*, *76*(2), 333-351.

Fowler, C., Hilsenroth, M. J., & Handler, L. (1995). Early memories: An exploration of theoretically derived queries and their clinical utility. *Bulletin of the Menninger Clinic*, *59*, 78-98.

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

Ganellen, R. J. (1996). Comparing diagnostic efficiency of the MMPI, MCMI-II, and Rorschach: A review. *Journal of Personality Assessment*, *67*(2), 219-243.

Garb, H. N., Wood, J. M., Nezworski, M. T., Grove, W. M., & Stejskal, W. J. (2001). Toward a resolution of the Rorschach controversy. *Psychological Assessment*, *13*(4), 433-448.

Guarnaccia, V., Dill, C. A., Sabatino, S., & Southwick, S. (2001). Scoring accuracy using the Comprehensive System for the Rorschach. *Journal of Personality Assessment*, *77*(3), 464-474.

Hess, A. K., Zachar, P., & Kramer, J. (2001). Rorschach. In B. S. Plake & J. S. Impara (Eds.), *Fourteenth mental measurements yearbook* (pp. 1033-1038). Lincoln: University of Nebraska Press.

Hiller, J. B., Rosenthal, R., Bornstein, R. F., Berry D. T. R., & Brunell-Neuleib, S. (1999). A comparative meta-analysis of Rorschach and MMPI-2 validity. *Psychological Assessment*, *11*, 278-296.

Hilsenroth, M. J., & Stricker, G. (2004). A consideration of challenges to psychological assessment instruments used in forensic settings: Rorschach as exemplar. *Journal of Personality Assessment*, *83*(2), 141-152.

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

Kamphuis, J. H., Kugeares, S. L., & Finn, S. E. (2000). Rorschach correlates of sexual abuse: Trauma content and aggression indexes. *Journal of Personality Assessment*, *75*(2), 212-224.

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.

Lohr, J. M., Fowler, K. A., & Lilienfeld, S. O. (2002). The dissemination and promotion of pseudoscience in clinical psychology: The challenge to legitimate clinical science. *The Clinical Psychologist*, *55*, 4-10.

Meloy, J. R., Acklin, M. W., Gacono, C. B., Murray, J. F., & Peterson, C. A. (1997). *Contemporary Rorschach interpretation*. Mahwah, NJ: Erlbaum.

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

Meyer, G. J. (1999a). Simple procedures to estimate chance agreement and kappa for the interrater reliability of response segments using the Rorschach Comprehensive System. *Journal of Personality Assessment*, *72*(2), 230-255.

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

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

Meyer, G. J. (2001). Introduction to the final special section in the special series on the utility of the Rorschach for clinical assessment. *Psychological Assessment*, *13*(4), 419-421.

Meyer, G. J. (2002). Exploring possible ethnic differences and bias in the Rorschach Comprehensive System. *Journal of Personality Assessment*, *78*(1), 104-129.

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

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

Meyer, G. J., 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., 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.

Plake, B. S., & Impara, J. C., (2001). *The fourteenth mental measurements yearbook*. Lincoln: The University of Nebraska Press.

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

Rorschach, H. (1921). *Psychodiagnostik*. Bern, Switzerland: Bircher (Transl. Hans Huber Verlag, 1942).

Rosenthal, R., Hiller, J. B., Bornstein, R. F., Berry, D. T. R., & Brunell-Neuleib, S. (2001). Meta-analytic methods, the Rorschach, and the MMPI. *Psychological Assessment*, *13*(4), 449-451.

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.

Shaffer, T. W., Erdberg, P., & Meyer, G. J. (2007). Introduction to the JPA special supplement on international reference samples for the Rorschach Comprehensive System. *Journal of Personality Assessment*, *89*(1), S2-S6.

Smith, B. L., Boss, A. L., Brabender, V., Evans, B. F., Handler, L., Mihura, J. L., & Nichols, D. (2005). The status of the Rorschach in clinical and forensic practice: An official statement by the board of trustees of the society of personality assessment. *Journal of Personality Assessment*, *85*(2), 219-237.

Sultan, S., Andronikof, A., Réveillère, C., & Lemmel, G. (2006). A Rorschach stability study in a nonpatient adult sample. *Journal of Personality Assessment*, *87*(3), 330-348.

Smith, K., Conroy, R. W., & Ehler, B. D. (1991). Lethality of suicide attempt rating scale. In B. Bongar's (Ed.), *The suicidal patient: Clinical and legal standards of care* (pp. 257-284). Washington, DC: American Psychological Association.

Viglione, D. J. (1999). A review of recent research addressing the utility of the Rorschach. *Psychological Assessment*, *11*, 241-265.

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

Weiner, I. B. (2001a). Advancing the science of psychological assessment: The Rorschach Inkblot Method as exemplar. *Psychological Assessment*, *13*(4), 423-432.

Weiner, I. B. (2001b). Considerations in collecting Rorschach reference data. *Journal of Personality Assessment*, *77*(1), 122-127.

Weiner, I. B., Exner, Jr., J. E., & Sciara, A. (1996). Is the Rorschach welcome in the courtroom? *Journal of Personality Assessment*, *67*(2), 422-424.

Weiner, I. B., Speilberger, C. D., & Abeles, N. (2002). Scientific psychology and the Rorschach Inkblot Method. *The Clinical Psychologist*, *55*(4), 7-12.

Wood, J. M., & Lilienfeld, S. O. (1999). The Rorschach inkblot test: A case of overstatement? *Psychological Assessment*, *6*(4), 341-349.

Wood, J. M., Lilienfeld, S. O., Nezworski, M. T., & Garb, H. N. (2001). Coming to grips with negative evidence for the comprehensive system for the Rorschach: A comment on Gacono, Loving, & Bodholdt; Ganellen; and Bornstein. *Journal of Personality Assessment*, *77*(1), 48-70.

Wood, J. M., Nezworski, M. T., & Stejskal, W. J. (1996). The comprehensive system for the Rorschach: A critical examination. *Psychological Science*, *7*(1), 3-9.

Wood, J. M., Nezworski, M. T., Lilienfeld, S. O., & Garb, H. N. (2003). *What's wrong with the Rorschach? Science confronts the controversial inkblot test*. San Francisco: Wiley.

Woody, R. H. (2002). Clinical psychology in the courtroom. *The Clinical Psychologist*, *55*, 13-18.

VITA

NAME:	Kevin Neil Park			
EDUCATION:	Rosemead School of Psychology Clinical Psychology	PsyD		(Cand.)
	Rosemead School of Psychology Clinical Psychology	MA		2001
	Dallas Theological Seminary Counseling	MA		1998
	Texas A&M University Biology	BA		1992
INTERNSHIP:	Naval Medical Center San Diego San Diego, CA	2003	-	2004
PRACTICA:	L.A. County - USC Medical Center Neuropsychological Assessment Clerkship	2001	-	2002
	Biola Counseling Center Adult Outpatient Practicum	2000	-	2001
	California High School Child/Adolescent Practicum	2000	-	2000
EMPLOYMENT:	Veterans Administration Readjustment Therapist	2009	-	present
	The Curtice Center (Private Practice) Psychotherapist	2007	-	present
	United States Navy Clinical Psychologist	2003	-	2007
	Biola Counseling Center Staff Therapist	2001	-	2003