The Rorschach: Facts, Fictions, and Future

Donald J. Viglione
Alliant International University

Mark J. Hilsenroth
Adelphi University

A large body of empirical evidence supports the reliability, validity, and utility of the Rorschach. This same evidence reveals that the recent criticisms of the Rorschach are largely without merit. This article systematically addresses several significant Rorschach components: interrater and temporal consistency reliability, normative data and diversity, methodological issues, specific applications in the evaluation of thought disorder and suicide, meta-analyses, incremental validity, clinician judgment, patterns of use, and clinical utility. Strengths and weaknesses of the test are addressed, and research recommendations are made. This information should give the reader both an appreciation for the substantial, but often overlooked, research basis for the Rorschach and an appreciation of the challenges that lie ahead.

Reliability

Interrater Reliability

Individual research studies published over the past 20 years in peer-reviewed journals (Acklin, McDowell, & Verschell, 2000; McDowell & Acklin, 1996; Meyer, 1997a, 1997b; Viglione, 1999), in Exner’s volumes (Exner, 1993), and in the publication process (Meyer et al., in press) provide conclusive empirical evidence of strong interrater reliability for the great majority (95%) of Comprehensive System (CS) variables. An extensive amount of data reveals that well-trained raters can score both high or low base-rate CS variables with good (K and intraclass correlation [ICC] >.60) to excellent (.75-.80) reliability (Garb, 1998; Shrout & Fleiss, 1979).1

Conclusion: Empirical Data Support Rorschach Interrater Reliability

The great majority (95%) of the individual variables are coded with good or excellent interrater reliability. Interrater reliability for form quality, individual cognitive special scores, form versus color dominance, shading subtypes, DOV and DOV+, and some content categories could be enhanced. Only a few very low base-rate variables (e.g., MQ none) occasionally produce poor reliability coefficients, but they have not yet been analyzed with sufficiently large samples given their miniscule base rates. As in the past (Exner, 1974, 1995; Viglione, 1997), Rorschach researchers have recognized the limitations of the test. Improvement in training for the coding of these variables is the focus of a new reference book currently being prepared for publication (Viglione, 2001).

Empirical Facts Regarding Temporal Consistency Reliability

There is general agreement (Meyer, 1997a; Viglione, 1999) with Wood and Lilienfeld’s (1999) statement that “Exner and his colleagues have reported test–retest reliability coefficients and that these coefficients are generally quite good” (p. 344). The great majority of the CS variables with such strong temporal consistency reliability are central interpretive variables (Exner, 1991, 1993). In

1 Our review of interrater reliability only addresses CS reliability. Many other studies have demonstrated good to excellent reliability data for non-CS variables such as the Mutuality of Autonomy (e.g., Fowler, Hilsenroth, & Handler, 1996), Aggressive Content (e.g., Gacono & Meloy, 1994), and the Rorschach Oral Dependency scales (e.g., Fowler et al., 1996).
other words, these variables are listed among the 68 nonredundant variables on the lower portion, “below the line” in the CS structural summary blank. In turn, the great majority of variables without test–retest coefficients (e.g., those identified by Garb, Wood, Nezworski, Grove, & Stejskal, 2001) are not included among these central, interpretive ratios, percentages, and derivations. Thus, the variables supposedly without reliability coefficients are not as central to interpretation. Of the 51 variables identified by Garb et al. (2001) as having test–retest coefficients, 46 (90%) are central, interpretive ratios, percentages, and derivations. Consequently, the great majority of the most important variables have adequate data for temporal consistency reliability evaluation.

Garb et al. (2001) presented two tables addressing the issue of CS test–retest reliability. Their Table 1 presents the central interpretive variables that have test–retest reliability coefficients. Among the 30 variables that Garb et al. identified (in their Table 2) as not having test–retest coefficients, there are 3 variables (FC:CF + C, a:p, EB) that actually do have adequate data. FC:CF + C, a:p, EB are each ratios of two other variables, and all six of the component scores are listed by Garb et al. in their Table 1 as variables for which test–retest coefficients have been reported. The text in a study by Exner, Armbruster, and Viglione (1978) clearly reveals that evidence of good reliability for EB, EA:ep (now called EA:es), and FC:CF + C cannot be attributed to chance. With respect to other variables that Garb et al. said lacked retest data, one, EBPer, is actually just a cutoff score of EB, two of the listed scores do not exist [F = % and H + (Hd):A + (Ad)], and one variable is listed (blends), even though Exner (1999) has reported on its retest stability. In addition, Garb et al. listed the An + Xy score twice, listed the W variable twice, first as W and then as W:M, and listed the variable M (in the W:M comparison), even though it is also listed in their Table 1 as one of the scores that has test–retest data. The variables CONTAM, PSV, and CP are listed, but they have such incredibly low base rates that correlational, parametric statistics are unstable without huge samples. Among the 30 variables that Garb et al. (2001) alleged to lack test–retest correlation coefficients, there are only 12 variables that actually do not have test–retest data.

Table 1 is a more complete listing of the available test–retest coefficients. These test–retest coefficients compare favorably to the Wechsler scales, second version of the Minnesota Multiphasic Personality Inventory (MMPI-2) and other tests. Table 2 presents data culled from many sources, comparing CS test–retest correlations with those from the other scales. Table 2 presents the recent meta-analyses that address this issue, an exhaustive selection of recent MMPI and MMPI-2 data, and data on two other representative tests, one a performance test and the second a commonly used research instrument.

Memory effects cannot account for strong temporal consistency reliability findings. With respect to retest intervals of 6 months or more, psychometric tests (e.g., Anastasi & Urbina, 1996) minimize the possibility that memory effects act as confounds. It is certainly not an issue in 3-year retest intervals (Exner et al., 1978). In addition, Haller and Exner (1985) addressed the memory issue empirically. They instructed experimental respondents to give different answers on the second testing, but they gave no such instructions to the control groups. The experimental group produced adequate temporal consistency reliability coefficients (mean correlation = .66; Mdn = .71), which were almost identical to those of the control group (mean correlation = .72; Mdn = .74). Nonetheless, a review of the responses revealed that experimental participants gave many more new responses in the second testing than did control participants. Thus, even though the content of the responses changed, the formal scores for the central interpretive variables largely remained the same.

The six actuarial indexes in the CS (i.e., Schizophrenia [SCZI], depression [DEPI], suicide constellation [S-CON], coping deficit [CDI], hypervigilance [HVI], and obsessive style [OBS]) lack definitive temporal consistency data, but available findings support their stability. The actuarial indexes are derived from 58 variables, 47 of which possess or adequate test–retest data in the research reports. Six of the 11 scores without test–retest data (FQo, FQ-, FQu, Level 2 special scores, and FABCOM2) are subcomponents of other variables (X = %, X+% Ego Impairment Index [EII], WSUM6, SUM6, INC + FAB) with good to excellent test–retest data.

Garb et al. (2001) cited several studies that seemed to demonstrate less acceptable temporal consistency reliability for the CS, but they did not report their serious methodological flaws. Thus, these flawed studies include research with a nonstandard form of administration (Schwartz, Mebane, & Malony, 1990); a dissertation involving only 17 older adults with an unspecified test–retest interval (Erstad, 1996); and a sample of schizophrenic patients with various illness courses, medications, hospitalizations, psychotherapies, and test–retest intervals ranging from 1 to 18 years (averaging 6.4 years; Adair & Wagner, 1992).

Conclusion and Future Research

The empirical data support the conclusion that the great majority of CS central interpretive ratios, percentages, and derivations possess impressive temporal consistency. As demonstrated in Table 2, the test–retest coefficients for CS Rorschach variables compare quite favorably with those for other tests. For the limited number of variables without published retest coefficients, most are of secondary interpretive significance. Typically, they enter into calculations to produce the central interpretive variables, and the latter have demonstrated adequate or better temporal consistency. The CS has more than adequate temporal consistency reliability in all respects, especially in the context of comparing Rorschach data to other established methods of assessment.

Nonetheless, more test–retest reliability research should be undertaken. The Rorschach and the CS, like many other psychological assessment measures (e.g., MMPI/MMPI-2 and Wechsler Adult Intelligence Scale—Revised [WAIS-R/Wechsler–III]), continue to evolve; therefore, temporal consistency information should be updated and include variables for which no data are available. There were only 22 ratios, percentages, and derivations in the CS structural summary and approximately a dozen other variables central to interpretation (Exner, Weiner, & Schuyler, 1976) at the time that the original temporal consistency data were processed (Exner, Armbruster, & Viglione, 1978). Future research could utilize short test–retest intervals and ensure minimal change in psychiatric, environmental, and target constructs. The sample should include both nonpatients and a variety of patients to avoid restrictions of range on actuarial indices and pathology indicators.

To analyze data with moderately low base rates, a large sample
Table 1
Test–Retest Correlation Coefficients for Comprehensive System Structural Summary Variables Among Adults

<table>
<thead>
<tr>
<th>Variablesa</th>
<th>Groupb</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>N</td>
<td>50</td>
</tr>
<tr>
<td>Retest period</td>
<td>≥ 1 year</td>
</tr>
<tr>
<td>Participants</td>
<td>Adult nonpatients</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Complexity, style, controls</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>.86</td>
<td>.79</td>
<td>.84</td>
<td>.77</td>
<td>.80</td>
</tr>
<tr>
<td>Lambda</td>
<td>.78</td>
<td>.82</td>
<td>.76</td>
<td>.82</td>
<td>.78</td>
</tr>
<tr>
<td>EB/4 (M: WSumC)</td>
<td>.84</td>
<td>.87</td>
<td>.83</td>
<td>.75</td>
<td>.77</td>
</tr>
<tr>
<td>M</td>
<td>.82</td>
<td>.86</td>
<td>.83</td>
<td>.68</td>
<td>.69</td>
</tr>
<tr>
<td>WSumC</td>
<td>.83</td>
<td>.85</td>
<td>.84</td>
<td>.71</td>
<td>.71</td>
</tr>
<tr>
<td>EM</td>
<td>.77</td>
<td>.72</td>
<td>.72</td>
<td>.28</td>
<td>.81</td>
</tr>
<tr>
<td>M</td>
<td>.26</td>
<td>.39</td>
<td>.34</td>
<td>.84</td>
<td>.06</td>
</tr>
<tr>
<td>C</td>
<td>.73</td>
<td>.67</td>
<td>.67</td>
<td>.77</td>
<td>.69</td>
</tr>
<tr>
<td>V</td>
<td>.87</td>
<td>.81</td>
<td>.89</td>
<td>.82</td>
<td>.92</td>
</tr>
<tr>
<td>T</td>
<td>.91</td>
<td>.87</td>
<td>.95</td>
<td>.86</td>
<td>.91</td>
</tr>
<tr>
<td>Y</td>
<td>.31</td>
<td>.23</td>
<td>.41</td>
<td>.69</td>
<td>.13</td>
</tr>
<tr>
<td>eb (FM + M: Sum Sh)</td>
<td>.69</td>
<td>.66</td>
<td>.71</td>
<td>.74</td>
<td>.71</td>
</tr>
<tr>
<td>Sum Sh</td>
<td>.64</td>
<td>.72</td>
<td>.59</td>
<td>.69</td>
<td>.73</td>
</tr>
<tr>
<td>D-Score</td>
<td>.91</td>
<td>.83</td>
<td>.88</td>
<td>.82</td>
<td>.85</td>
</tr>
<tr>
<td>Adj es</td>
<td>.82</td>
<td>.85</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Affect</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>FC</td>
<td>.86</td>
<td>.86</td>
<td>.92</td>
<td>.71</td>
<td>.72</td>
</tr>
<tr>
<td>CF + C</td>
<td>.81</td>
<td>.79</td>
<td>.83</td>
<td>.74</td>
<td>.72</td>
</tr>
<tr>
<td>C + Cn</td>
<td>.56</td>
<td>.51</td>
<td>.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Afr</td>
<td>.82</td>
<td>.90</td>
<td>.85</td>
<td>.57</td>
<td>.76</td>
</tr>
<tr>
<td>S</td>
<td>.74</td>
<td>.71</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ideation</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>.83</td>
<td>.86</td>
<td>.87</td>
<td>.71</td>
<td>.83</td>
</tr>
<tr>
<td>p</td>
<td>.72</td>
<td>.75</td>
<td>.85</td>
<td>.51</td>
<td>.78</td>
</tr>
<tr>
<td>2AB + Art + Ay (Intell.)</td>
<td>.84</td>
<td>.88</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sum6 special scores</td>
<td>.81</td>
<td>.88</td>
<td>.81</td>
<td>.84f</td>
<td>.68f</td>
</tr>
<tr>
<td>WSum6</td>
<td>.86</td>
<td>.86</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mediation</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>.83</td>
<td>.73</td>
<td>.81</td>
<td>.76</td>
<td>.82</td>
</tr>
<tr>
<td>X+%</td>
<td>.86</td>
<td>.80</td>
<td>.87</td>
<td>.82</td>
<td>.85</td>
</tr>
<tr>
<td>X-%</td>
<td>.92</td>
<td>.80</td>
<td>.88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xu%</td>
<td>.85</td>
<td>.89</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Information processing</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Zf</td>
<td>.85</td>
<td>.83</td>
<td>.89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zd</td>
<td>.85</td>
<td>.83</td>
<td>.89</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Interpersonal</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>COP</td>
<td>.81</td>
<td>.88</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AG</td>
<td>.82</td>
<td>.86</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IsolateR</td>
<td>.84</td>
<td>.83</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Self-perception</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3r + (2 VR)</td>
<td>.89</td>
<td>.87</td>
<td>.90</td>
<td>.74</td>
<td>.73</td>
</tr>
<tr>
<td>Fr + rF</td>
<td>.82</td>
<td>.89</td>
<td>.89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FD</td>
<td>.88</td>
<td>.90</td>
<td>.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOR</td>
<td>.71</td>
<td>.83</td>
<td>.87</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---


b For those variables listed as missing by Garb et al. (2001), and for which we could find no published retest coefficients, we asked Exner to provide the coefficients. Garb et al. (2001) did not list M—, DQ+, DQv, and Level 2 cognitive special scores as missing. Accordingly, we did not process coefficients for these variables and do not present them in the table. CP and Mnone are listed as missing by Garb et al. (2001). They have extremely low base rates. Only 5 of the approximately 13,392 responses in the nonpatient reference group of 600 adults (Exner, 2001) were assigned a CP and only 4 were assigned a Mnone. Accordingly, a huge sample is needed to stabilize the retest coefficients for these two variables and values were not requested for them. Test–retest coefficients are still needed for Ma, Mp, Level 2 cognitive special scores, W, D, Dd, DQ+, DQv, Fd, Pure H, (H) + Hd + (Hd), An + X>, and the new comprehensive system variable (Exner, 2001).

c EBPer is merely a cutoff score of EB. It is listed as missing by Garb et al. (2001).

d Adj D is merely a transformation of EA—Adj es, both of which have coefficients.


f At the time of these publications (Haller & Exner, 1985; Exner, Thomas, & Mason, 1985), there were only five cognitive special scores.
A Comparison of Test–Reetest Correlations Among Frequently Used and Highly Regarded Tests

<table>
<thead>
<tr>
<th>Tests variables</th>
<th>Correlation</th>
<th>Retest interval</th>
<th>M</th>
<th>Mdn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple studies</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Meta-analysis—Self-report, observer, and projectivea</td>
<td></td>
<td>1 year</td>
<td>.55</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 years</td>
<td>.52</td>
<td>—</td>
</tr>
<tr>
<td>Meta-analysis—8 self-report testsb</td>
<td></td>
<td>1 year</td>
<td>.78</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 years</td>
<td>.58</td>
<td>—</td>
</tr>
<tr>
<td>Meta-analysis—MMPIb</td>
<td></td>
<td>5 years</td>
<td>.57</td>
<td>—</td>
</tr>
<tr>
<td>MMPI—7 studiesc</td>
<td></td>
<td>1 year</td>
<td>.57</td>
<td>.58</td>
</tr>
<tr>
<td>Single studies</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MMPI⁰</td>
<td></td>
<td>1 day to 2 years</td>
<td>.68</td>
<td>.68</td>
</tr>
<tr>
<td>MMPI-2</td>
<td></td>
<td>1 week</td>
<td>.80</td>
<td>.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 months</td>
<td>.73</td>
<td>.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 years</td>
<td>.72</td>
<td>.73</td>
</tr>
<tr>
<td>California Verbal Learning Test</td>
<td></td>
<td>1.3 years</td>
<td>.47</td>
<td>.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 year</td>
<td>.53</td>
<td>.54</td>
</tr>
<tr>
<td>Attributional Style Questionnairec</td>
<td></td>
<td>5 weeks</td>
<td>.63</td>
<td>.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 years</td>
<td>.79</td>
<td>.80</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 weeks</td>
<td>.80</td>
<td>.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 year</td>
<td>.76</td>
<td>.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 weeks</td>
<td>.82</td>
<td>.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 year</td>
<td>.82</td>
<td>.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 year⁶</td>
<td>.75</td>
<td>.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 to 4 days</td>
<td>.71</td>
<td>.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 to 4 days</td>
<td>.66</td>
<td>.71</td>
</tr>
</tbody>
</table>

Note. Dashes indicate measure was not applicable. MMPI = Minnesota Multiphasic Personality Inventory.


(N = 80 or more) should be collected. Interpretive temporal consistency based on established interpretive cutoffs should be included. Systematically assigning multiple raters to recode records would allow one to discriminate the effects of temporal inconsistency from interrater disagreement in accounting for changes over time.

Normative Data and Diversity Issues

The detailed presentation of large reference samples from nonpatients, various patient groups, and children from ages 5 to 16 has contributed to a resurgence of the Rorschach over the past 20 years (American Psychological Association, [APA] 1998; Exner, 1974, 1993; Exner et al., 2001; Wood & Lilienfeld, 1999). To limit our discussion of normative data, we focus here on the adult, nonpatient CS reference sample, consisting of 600 adults, which is a relatively large normative sample (Exner et al., 2001). It is stratified for geographic distribution and partially stratified for socioeconomic level. The first of these nonpatient participants was collected before 1975, followed by various additions and revisions through 1991 (Exner, 1993). All respondents were volunteers, with the majority coming from places of employment. The gender distribution is balanced, but the sample is rather young; 73% of the participants were under 36 years of age. Nonetheless, there are some differences between this adult group and the 15- and 16-year-old sample (Exner et al., 2001). The adult group is relatively well-educated; the median grade level is over 13 years, and only 5% had less than 12 years of schooling. Eighteen percent of the respondents were non-Caucasian, which is consistent with other normative samples, although this percentage could be increased. For example, the minority percentage for the WAIS-III is 21% (The Psychological Corporation, 1997) and for the MMPI-2 it is 19% (Hathaway et al., 1989).

Rorschach researchers (Brinderson, 1995; Kates, 1994; Shaffer, Erdberg, & Haroian, 1999; Viglione, Gaudiana, & Gowri, 1997; Weiner, 2001b) have questioned the continued suitability of these CS norms. For reasons similar to the emergence of concerns about the aging of MMPI norms (e.g., Colligan, Osborne, & Offord, 1984), one might contend that social and cultural shifts, cohort differences, and the gradual evolution of the test justify collection of a new sample.

Research Data as Approximations of Nonpatient Sampling

To approximate a contemporary nonpatient reference sample, we reviewed the literature for available adult control and nonpatient groups. Shaffer et al. (1999) published the most comprehensive and notable effort to produce normative approximation data. To construct Table 3, we selected central interpretive variables that appeared to deviate from CS normative expectations in the Shaffer et al. study. We then supplemented the Shaffer et al. findings with data from five other published samples known to us and one previously presented sample (Viglione et al., 1997). These data allowed us to address some issues regarding deviation from CS reference sample expectations.

R. The data in Table 3 suggest that R is quite variable. Whether or not such variability presents a problem for the Rorschach has been debated for many years. The distributional parameters (mean is greater than median, a large standard deviation, and a minimum value of 14 with no predetermined maximum value) suggest that a great deal of variability is due to a small proportion of records with excessively high R. From a clinical utility perspective, long records consume administration and coding time and increase cost without necessarily increasing interpretive yield.

Lambda. The data for Lambda are incomplete, but suggest that median values are close to CS reference values. The Shaffer et al. (1999) study provides the only sample that seems to deviate from the expected distribution. Nonetheless, on the basis of standard deviations, within each sample there are records that are much more form dominant than would be expected. The occasional high Lambda record does not pose much of a problem for routine interpretation. On the other hand, low R–high Lambda records may lack negative predictive power and may reduce interpretive yields and cost-benefit ratios.
so that this call 20% of Rorschach responses among nonpatients distorted. We construct of perceptual distortion. One could justify 5% or 10% of then we have a problem with the operational definition for the cutoff for the "distortion" interpretation of X-% may be closer X-%, as a measure of perceptual distortion, need to be modified. (1999).

of the responses in nonpatient records are distorted (i.e., FQ-), percentages to address these issues. However, if it is true that 20% of possible false-positive findings. Solely by form quality should be closely examined to rule out

Interpretive relationships. Viglione, Gaudiana, & Gowri (1997), nonpatient volunteers, without history of psychiatric treatment, deny substance abuse. b Because of significant problems with skew for this ratio as the denominator, Pure F, approaches zero, means and standard deviations may be misleading. The large standard deviations with means greater than medians suggest that skew is present and that the means are artificially elevated by outliers or skewed high values. 

Form quality. X+% and X-% results suggest that FQu and FQ- are more numerous than anticipated, but they are consistent with some previous observations (Viglione, 1989). Interpretive ranges for X+%, as a measure of overall conventionality, and X-%, as a measure of perceptual distortion, need to be modified. The data in Table 3 may underestimate X+% and overestimate X-% because one cannot be sure that pathological respondents were excluded from the samples. Nevertheless, the results do suggest that a more appropriate cutoff for the "unconventional" interpretation of X+% may be approximately X+% < .50 and that the cutoff for the "distortion" interpretation of X-% may be closer to X-% > .25. The data suggest that the cutoffs for children should also be more liberal (Hamel, Shaffer, & Erdberg, 2000; Kates, 1994). With improvements in coding FQ, as suggested in the interrater reliability section in this article, these interpretive bands might tighten. It is also important to recognize that FQ has some of the best validity data of any coding CS variable (Dawes, 1999; Viglione, 1999). Cases with elevations in the SCZI produced solely by form quality should be closely examined to rule out possible false-positive findings.2

Exner (2000) has made some changes in the form quality percentages to address these issues. However, if it is true that 20% of the responses in nonpatient records are distorted (i.e., FQ-), then we have a problem with the operational definition for the construct of perceptual distortion. One could justify 5% or 10% of the responses being distorted, but it does not make sense for us to call 20% of Rorschach responses among nonpatients distorted. We may need to narrow the operational definition of FQ-, so that this code truly represents distorted perception as outlined by Dawes (1999).

WSUM6. For cognitive special scores in the form of WSUM6, the large standard deviations suggest that a small proportion of individuals provide many of the elevated values. Given that these respondents are not patients (though it is possible that some had been prior to testing), one cannot be sure whether these results are associated with poor sampling, idiosyncratic responses, idiosyncratic coding, or overly long, elaborated, and complex records. On the other hand, central tendencies are consistent with the CS data. The standard deviations for WSUM6, as well as R and Lambda, suggest that these samples contain more exceptional or extreme cases than does the CS reference group.

**Cultural and International Implications**

These normative approximation data also have implications for cultural and international issues. Research data on the Rorschach and diversity issues support the use of this measure with diverse populations (Viglione, 1999). Large-scale cultural diversity and international projects have been under way for some time, and expertise among the project managers around the world is improving (Erdberg & Schaffer, 1999). These data have the potential of teaching us a great deal about cultural and geographic issues. In terms of thinking about cultural differences and expectations from

2 The SCZI is being replaced by a dimensional index of thought disorder, the Perceptual Thought Index (Exner, 2001). Such an index more accurately reflects the available data (Hilsenroth, Fowler, & Padawer, 1998; Kleger, 1999) and a modern understanding of thought disorder as existing on a continuum from optimal to most disordered within and across diagnoses (Viglione, 1999).
a normative perspective, the data emerging from these studies are more consistent with the normative approximation data given in Table 3 than they are with the CS reference sample.

Moreover, emerging childhood reference group data from United States samples (Hamel et al., 2000; Kates, 1994) and international samples (Erdberg & Schaffer, 1999) are consistent with a normative approximation sample of African American children that was collected about 20 years ago (Krall et al., 1983). Garb et al. (2001) cited the Krall et al. study in their challenge of the diversity applications of the Rorschach. However, this is quite puzzling because a close examination of the data indicates that the only problematic difference between Krall et al.'s findings and the norms available at that time (Exner & Weiner, 1982) is for a form quality score (i.e., a lower $F+$%), which is no longer being used in the CS (Exner, 2000). Krall et al. reported a range for $F+$% of .48 to .69, which is consistent with the normative approximation results presented in Table 3 and the emerging international data. Also, differences in popular responses, for example, are as expected given cultural variations (Mattlar, Carlsson, & Forsander, 1993). Thus, supposed problematic cultural differences evaporate when one considers the emerging international data. This conclusion is supported by the very limited differences between nonpatient African American and Caucasian American adults matched on important demographic variables (Presley, Smith, Hilsenroth, & Exner, in press). Finally, Meyer (in press) extended these findings by conducting a series of analyses to explore potential ethnic bias in Rorschach CS variables with a patient sample. After matching on several salient demographic variables, ethnicity revealed no significant findings, and principal-components analyses revealed no evidence of ethnic bias in the Rorschach's internal structure.

A New Normative Sample?

The Shaffer et al. (1999) and other normative approximation data are relatively consistent with CS reference group data (Exner et al., 1995). Adjusting one's expectations for form quality, as outlined above, should eliminate the possibility of overpathologizing adults.

An alternative point of view about these reference data with nonpatients is also emerging (Meyer, 2001; Weiner, 2001b). This point of view suggests that there should be a careful consideration of the guidelines used for collecting this Rorschach reference data. Figural in this understanding of differences is the recognition that Exner's (1993) reference sample consists of socially-vocationally functioning nonpatients. An alternative point of view about these reference data with nonpatients is also emerging (Meyer, 2001; Weiner, 2001b). This point of view suggests that there should be a careful consideration of the guidelines used for collecting this Rorschach reference data. Figural in this understanding of differences is the recognition that Exner's (1993) reference sample consists of socially-vocationally functioning nonpatients. In addition, it is equally important to note that CS data collected from both United States and worldwide samples are somewhat more healthy than Exner's reference samples of people who are starting outpatient psychotherapy or different patient groups. The interested reader is directed to Weiner (2001b), who proposed guidelines for research when obtaining reference samples from patients and nonpatients. Meyer (2001) also provided an extensive and in depth review of possible factors affecting normative shifts in Rorschach data. Both of these authors presented rational and empirical data to support their position that CS norms do not overpathologize potential examinees.

Nonetheless, collecting new reference samples, with increased attention to cultural and other diversity issues, should improve reliability and validity, and enhance utility. Like the MMPI, the CS and Rorschach have evolved since they were introduced (Exner, 1974). The findings with $R$, and the possibility of related variability among the other indices, suggest that constraining $R$ through statistical comparisons or procedurally through minor administrative modifications deserves some consideration.

Meta-Analyses

Do meta-analytic results support the validity of the Rorschach in general terms? There are at least six recent and original meta-analyses that address Rorschach validity (Atkinson, Quarrington, Alp, & Cyr, 1986; Bornstein, 1996, 1999; Hiller, Rosenenthal, Bornstein, & Berry, 1999; Meyer & Handler, 1997; Parker, Hanson, & Hunsley, 1988). All concluded with empirically based statements, supporting the validity of the Rorschach. Weiner (2001a) summarized these empirical data and offered them as convincing support of the overall validity of the Rorschach. Opposing these five meta-analyses is one re-analysis of the Parker et al. data (1988) by Garb, Florio, and Grove (1998). The research in this meta-analyses was published from 1954 through 1982 and contains only 15 Rorschach studies (compared with 75 utilizing the MMPI), only 5 of which utilized the CS. As suggested in the original authors' (Parker, Hunsley, & Hanson, 1999) reply to the re-analysis of their data, the relevance of the Garb et al. (1998) re-analysis is highly questionable and adds little to our contemporary understanding of these two clinical instruments. The Rorschach research and techniques have advanced quite a bit over the last 20 years. Likewise, the MMPI-2 was not available to be evaluated in either of these two meta-analytic studies.

Analysis of moderator variables in the Hiller et al. (1999) meta-analysis in this Special Series revealed that the MMPI had larger validity coefficients than the Rorschach for self-report criteria and psychiatric diagnosis "whereas the Rorschach had higher validity coefficients in studies using objective outcomes as criterion variables" (Hiller et al., 1999, p. 291). This is consistent with Viglione's (1999) finding that the Rorschach variables were valid and useful when one emphasizes ecologically valid behavioral criteria, rather than self-report criteria. Thus, as the criterion improves, so do the Rorschach validity coefficients. In contrast, the MMPI does best with criteria that share self-report method variance. Hiller et al. speculated that impression management plays less of a part in the Rorschach than in self-report testing. "In these instances, it may be that the Rorschach is more objective than so-called objective instruments" (p. 291). Although Hiller, Rosenenthal, and their colleagues' methods, data, and conclusions have been critiqued, the support for the Rorschach from their work, as well as from all the other original meta-analyses, are quite clear and accurate. As a whole, these meta-analyses reveal that the Rorschach produces valid data to measure many objective, behavioral criteria.

Incremental Validity

Research data from a variety of sources reveals considerable emerging empirical support for the incremental validity of the Rorschach. This research is not exhaustive, but it is clear that Rorschach variables do possess incremental validity over other well-known tests and collateral information. As such, the evidence indicates that the Rorschach adds meaningful data to the assess-
ment process. Accordingly, negative characterizations, such as "The frequent claim that the Rorschach adds meaningful assessment information to other data has not been supported in any published study" (Hunsley & Bailey, 1999, p. 271), are false because they ignore empirical findings.

**Empirical Data That Address Incremental Validity**

Research presented in Weiner’s (2001a) and Viglione’s (1999) reviews of the literature contain empirical data consistent with the conclusion that Rorschach variables possess incremental validity over other tests, demographic data, and other information (see Table 4). These data were not sufficiently recognized by those who adopted a critical view of the Rorschach in this Special Series. On the other hand, consistent with Weiner, Garb (1999) did confirm incremental validity support to the EII, the Rorschach Oral Depression scale, and the Rorschach Prognostic Rating scale (PRS).

Meyer (2000) has addressed incremental validity issues in two meta-analyses. Within six studies, using both the Rorschach PRS and the MMPI Ego Strength scale (Es), the Rorschach PRS demonstrated considerable incremental validity for predicting treatment outcome. The uncorrected and corrected effect sizes were $r = .32$ and $r = .48 (N = 229)$, respectively. The uncorrected and corrected effect sizes, after excluding the outlier results from one

<table>
<thead>
<tr>
<th>Author and year</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archer &amp; Gordon (1988)</td>
<td>Optimal overall correct classification (OCC hit rate) of individual inpatients with schizophrenia for Rorschach Schizophrenia (SCZI = 5) = .80. Optimal OCC for Minnesota Multiphasic Personality Inventory (MMPI) Sc scale (Sc = 75) = .76. Utilizing traditional cutoff scores, OCC rates were as follows: ScZI ≥ 4, OCC = .69; Sc ≥ 65, OCC = .48; and Sc ≥ 70, OCC = .60.</td>
</tr>
<tr>
<td>Archer &amp; Krishnamurthy (1997)</td>
<td>In classification of depression in adolescents with a stepwise discriminant function analysis, Rorschach variables VISTA and Afr added $R^2 = .05$ beyond combined $R^2 = .14$ for MMPI-A scales AD-P and Ma. These four variables had highest positive predictive power over any single variable or combination of variables.</td>
</tr>
<tr>
<td>Bornstein et al. (1997)</td>
<td>Rorschach Dependency scores significantly correlated with both number of significant interpersonal events ($r = .84$) and impact ratings of these events ($r = -.38, p &lt; .01$). Self-report measure of dependency was not significantly related to either ($r = -.11$ and $r = .16$, respectively).</td>
</tr>
<tr>
<td>Cooper et al. (1991)</td>
<td>The Rorschach Defense scales provided unique prediction of outcome GAF (Global Assessment of Functioning, Health–Sickness) ratings in regression equations beyond initial GAF and borderline personality self-report scale.</td>
</tr>
<tr>
<td>Hilsenroth et al. (1995)</td>
<td>Rorschach variables were able to significantly differentiate ($p = .008$) those patients prematurely terminating from psychotherapy versus those continuing in treatment, whereas the MMPI-2 was unable to do so ($p = .56$). Specifically Rorschach scores from the interpersonal–relational cluster had a mean effect size (ES) of .57, while the MMPI-2 content scale Negative Treatment Indicators had an ES of −.14.</td>
</tr>
<tr>
<td>Holzman et al. (1974)</td>
<td>The classification of a recent schizophrenia diagnosis (hospitalized less than 6 months) and deviant eye tracking was greater (65% accuracy) utilizing Rorschach data alone than a clinical team diagnosis (58%).</td>
</tr>
<tr>
<td>O’Connell et al. (1989)</td>
<td>Rorschach data (Thought Disorder Index (TDI)) predicted the development of psychotic and psychotic-like symptoms over a 2–3 year period in a sample of Axis II and affective disorder patients over and above information from clinical interview on lifetime prevalence of psychotic and psychotic-like symptoms (combined $R^2 = .21$, TDI-beta = .32, $p &lt; .03$), schizotypal symptoms (combined $R^2 = .42$, TDI-beta = .32, $p &lt; .009$), or schizotypal and borderline symptoms (combined $R^2 = .51$, TDI-beta = .31, $p &lt; .006$). Initial TDI scores also demonstrated clinical utility in prediction of psychotic and psychotic-like symptoms at follow-up.</td>
</tr>
<tr>
<td>Perry &amp; Braff (1994)</td>
<td>Human Experience variable component of the Ego Impairment Index (EII) significantly related to neuropsychological markers of schizophrenia ($r = -.42, p &lt; .01$; $r = -.31, p &lt; .025$; $r = -.35, p &lt; .025$), whereas thought disorder scales based on clinical interview (Schedule for Positive Symptoms and Schedule for Negative Symptoms) were not ($p &gt; .05$).</td>
</tr>
<tr>
<td>Perry &amp; Viglione (1991)</td>
<td>Rorschach EII predicted outcome Beck Depression Inventory (BDI; $p &lt; .0002$) and Carroll Rating Scale for Depression ($p &lt; .01$) scores in depressed patients treated with tricyclic antidepressants beyond variance accounted for by gender and baseline scores on BDI and EII. Other demographic variables were also considered but did not affect outcome.</td>
</tr>
<tr>
<td>Russ (1980)</td>
<td>Rorschach measures of adaptive regression (AR) and defensive effectiveness (DE) were significantly related to academic achievement, independent of IQ (AR: $r = .45, p &lt; .01$; DE: $r = .40, p &lt; .01$, respectively).</td>
</tr>
<tr>
<td>Russ (1981)</td>
<td>Rorschach measure of AR was significantly related to reading and overall academic achievement independent of IQ ($r = .51, p &lt; .001$), and $r = .41, p &lt; .001$). Index AR scores were significantly predictive of reading achievement 1 year later ($r = .29, p &lt; .05$).</td>
</tr>
<tr>
<td>Shapiro et al. (1990)</td>
<td>Rorschach Depression Index significantly differentiated sexually abused African American girls from controls ($p &lt; .005$). Children’s Depression Inventory scores for sample were not significantly different from controls ($p &gt; .05$), consistent with incremental validity of the Rorschach relative to self-reported depression. The groups did differ on the Internalization scale of the Child Behavior Checklist ($p &lt; .0001$).</td>
</tr>
<tr>
<td>Skelton et al. (1995)</td>
<td>The dependent variable was a ratio of Rorschach TDI over a TDI derived from the Wechsler. This ratio was 2.46 times higher in a group of 25 identity-disordered adolescents than it was among 35 conduct-disordered and oppositional defiant adolescents ($p &lt; .01$).</td>
</tr>
</tbody>
</table>
study (Fiske, Cartwright, & Kirtner, 1964), were $r = .40$ and $r = .59$ ($N = 187$), respectively. Effect sizes for the ESs were essentially nil (range, –.03 to 03). Clearly, the Rorschach PRS demonstrates incremental validity over the MMPI ES scale. Nine studies ($N = 358$) included both the Rorschach PRS and an intelligence test. Across these studies, the Rorschach PRS demonstrated considerable incremental validity over and above intelligence in the prediction of future outcome. After correcting for methodological artifacts, the unique contribution for the PRS was $r = .48$; and without such corrections, $r = .36$. Blatt’s (Blatt et al., 1994) findings also demonstrate incremental validity of Rorschach scales over IQ in the prediction of treatment outcome, as do a number of studies summarized in Viglione’s (1999) synthesis of the literature (see Table 4).

Blais, Hilsenroth, Castlebury, Fowler, and Baity (2001) explored the incremental validity of the Rorschach in discriminating Diagnostic and Statistical Manual of Mental Disorders (4th ed.; DSM-IV; American Psychiatric Association, 1994) personality disorders (PD). In a re-analysis of existing data, select Rorschach variables and the MMPI-2, PD scales were used to predict chart review ratings of DSM-IV antisocial, borderline, histrionic, and narcissistic personality disorders in a sample of treatment-seeking outpatients. Hierarchical regression analyses showed that both MMPI-2 and Rorschach data added incrementally to the prediction of DSM-IV borderline and narcissistic PD criteria. The findings were less clear for the incremental value of Rorschach and MMPI-2 data in predicting the total number of DSM-IV histrionic PD and antisocial PD criteria. Rorschach data were the best predictors of DSM-IV histrionic criteria whereas MMPI-2 data were best at predicting the DSM-IV antisocial criteria.

Dawes’s (1999) Results Actually Support the Criterion and Incremental Validity of the EII and Rorschach

Earlier in this Rorschach Special Series, Dawes (1999) “present[ed] two methods for assessing the incremental validity of a particular Rorschach variable” (p. 297), namely the EII, with the expressed purpose of demonstrating how incremental validity of the Rorschach might be investigated. He investigated two data sets from: (a) Meyer and Resnick (1996), which includes the Rorschach, MMPI, and impairment as approximated by severity defined by diagnosis (DxMax); and (b) Perry, Moore, and Braff (1995), which contains the Rorschach, measures of social competence and thought disorder, and another psychiatric rating scale. His conclusions are quite critical of the Rorschach but do not fit the data he presents. The analytic strategies he used reduced the probability of finding support for the Rorschach and EII and biased the analyses against the Rorschach. Therefore, we take issue with his methods and conclusions in this article.

Reducing the power of the analyses. To re-evaluate these data sets, Dawes (a) entered the number of Rorschach responses (R) and his new form quality measure, XQUAL, before the EII in the hierarchical regression; and (b) used a unit weighting scheme in addition to multiple regression. The argument for using the number of responses and his new measure, XQUAL, is that these Rorschach variables are easier to score and involve less projection by the clinician. There is no established body of research that supports the number of responses to predict impairment; therefore, there is no a priori reason to justify it as a predictor of impairment in a regression analysis. Also, data from the study do not support including R as a predictor: The number of responses is not significantly correlated with the criteria (Meyer’s DxMax, $r = .13$, ns and in the wrong direction; and Perry’s Social Competence Index, $r = .16$, ns).

Similarly, Dawes’s decision to create a new variable, XQUAL and to require the EII to predict beyond XQUAL in a regression equation cannot be justified. The EII contains form quality (FQ—) and R among its subcomponents. XQUAL and R are, thus, structurally redundant with EII and produce multicollinearity problems within the regression, reducing the power of the analyses. Among the arguments for the unit weighting system is that a clinician’s intuition does not reliably “capture optimal weighting systems” (p. 299). Its effect in the analysis is to reduce power once again.

Garb et al. (2001) argued that Dawes’s (1999) and Meyer and Resnick’s (1996) use of the severity of clinical diagnosis as a criterion is only a gross approximation of impairment or psychological disturbance. This is true, as it is for the Social Competence Index criterion in the Perry, Moore, and Braff (1995) data set. These measures do not share a great deal of variance with impairment, the intended target construct of the EII. This limited overlap between measure and target construct stacks the cards against the Rorschach by limiting maximal effect sizes and the power of the analyses. Small effect sizes are to be expected, even in the best circumstances, with such a global criterion. Under such circumstances, the maximum multiple Rs of .49 and .35, reported by Dawes, are quite impressive.

Dawes also introduced DxSum, the sum of the impairment scores for up to three diagnoses, as a criterion measure. It was used without precedent, supporting data, or a straightforward rationale. The DxSum makes the sum of mild diagnoses much more impaired than the most virulent but pure schizophrenia or bipolar disorder with psychotic features and adds systematic error to DxMax. We obtained the Meyer and Resnick (1996) data set ($N = 187$) and found that 36 individuals in the data set had a diagnosis with moderate severity, namely a DxMax score of 3 or less, but a DxSum of 6 or more. For these 36 patients, even though they had been assigned several diagnoses of mild to moderate severity, their DxSum score made them appear more impaired than individuals with a single diagnosis of schizophrenia ($N = 16$). In addition, the data do not support Dawes’s argument: All correlations in Dawes’s Table 1 between meaningful predictors (the MMPI mean elevation, the MMPI Goldberg index, XQUAL and EII) are higher for DxMax than for Dawes’s DxSum. Clearly, the DxSum variable adds unacceptable error to the criterion and only functions to bias the results against the Rorschach.

One may presume that some researchers and clinicians might disagree with the statement that form quality distortions (FQ—) “involve neither empirical research to establish their validity nor training in the Rorschach to score” (Dawes, 1999, p. 297). However, it appears that Dawes accepts that certain FQ— responses can be seen as clear, in vivo problematic behaviors, consistent with understanding the Rorschach as behavioral sample of cognitive problem-solving behaviors (Viglione, 1999).

Dawes (1999, p. 298) stated that “the authors considered the average rating for up to three diagnoses and the sum of the ratings for up to three diagnoses.” Meyer (personal communication, March 6, 2000) calculated the average and sum of diagnoses, but he did not recall considering them to be valid criterion variables and did not include them in his analyses.
With respect to the Perry et al. data set, the only relevant statistic is the univariate correlation relating the EII to the Social Competence Index \((r = .31, p < .05)\). Also, on the basis of the additional measures included in the Perry data set, the EII was correlated with the Schedule for Positive Symptoms global score \((r = .367, p = .007)\) and with a thought disorder subscale of the Brief Psychiatric Rating scale \((r = .334, p = .014)\). However, as expected, it was not correlated with the Schedule for Negative Symptoms global score \((r = -.092, ns)\).

The Meyer data set allows us to address incremental validity. For the reasons given above, the only defensible hierarchical analytic strategy for this purpose is to add the EII and the other Rorschach variables after the MMPI variables. These analyses are presented in Table 5. The correlation between EII and DxMax is .35. Given the limitations of DxMax as a measure of impairment, the incremental \(R_s\) for the EII and the Rorschach are quite impressive and may approach the maximum validity of DxMax as a measure of impairment.

**Conclusion regarding Dawes (1999).** Thus, Dawes's (1999) analyses support the incremental and criterion validity of the Rorschach. The EII provided significant unique predictive power over MMPI variables for psychological impairment or disturbance as grossly estimated by the maximum severity of diagnosis. Analytic decisions by Dawes reduce the sensitivity of the analyses and, as a result, minimize the potential to support the Rorschach. Re-analyses of these data sets provide clear empirical support for criterion validity and incremental validity of the EII and the Rorschach.

**Challenges to Incremental Validity of the Rorschach**

Recent claims (Garb, 1999; Garb et al., 2001; Hunsley & Bailey, 1999) that the Rorschach has poor incremental validity are based largely on old studies that confound clinical judgment with incremental validity (Garb, 1994; see Garb, 1998, for a reproduction of these studies), a point we return to later. These studies also do not have the benefit of the CS scoring system and established interpretive routines, in addition to lessons learned from clinical judgment studies (Exner, 1993; Garb, 1994, 1998; Meyer et al., 1998). Accordingly, these old experiments now lack ecological validity.

**Table 5**

**Incremental Validity of Ego Impairment Index (EII) and Rorschach From Dawes (1999) With Meyer and Resnick's (1996) Data Set Predicting Ratings of Severity of Impairment of Dx (DxMax)**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>(R)</th>
<th>Change in (R^2)</th>
<th>Incremental (r) over MMPI only</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMPI</td>
<td>.366</td>
<td>.134</td>
<td></td>
</tr>
<tr>
<td>MMPI + EII</td>
<td>.424</td>
<td>.058</td>
<td>.241</td>
</tr>
<tr>
<td>MMPI + Rorschach</td>
<td>.492</td>
<td>.108</td>
<td>.329</td>
</tr>
</tbody>
</table>

*Note.* The Minnesota Multiphasic Personality Inventory (MMPI) and Rorschach variables for analysis were those selected by Dawes: For the MMPI, mean elevation clinical scales and Goldberg Index; and for the Rorschach, EII, number of responses, and XQUAL.

**Incremental Validity and Clinical Utility**

The clinical utility of a test emerges within individual cases that pose idiosyncratic questions and require idiographic decisions (Kleinmuntz, 1990; Meehl, 1957; Meyer et al., 1998). Another complicating factor is that the "yield" from other sources of data and tests varies from case to case. For example, a Rorschach may be too sparse to adequately address interpersonal perception or thought disorder. In one case, the MMPI may be too exaggerated to address depression. For these and other reasons, the usefulness of a particular test varies greatly from case to case and question to question. "Field" incremental validity, therefore, only manifests in actual clinical practice in the form of clinical usefulness and the impact of different tests in decision making within individual cases. Weiner (1999) and Stricker and Gold (1999) provided compelling rationales and case examples that clearly demonstrate the clinical "field" incremental validity of the Rorschach in individual cases.

**Future Research**

As with all personality measures, more research that demonstrates incremental validity is needed. More sophisticated research studies might address complexes of Rorschach variables with real life behavioral and cognitive criteria. It would also be helpful to look at response styles, response sets, and contextual moderator variables that might influence incremental validity. In other words, Rorschach variables may generalize best to contexts that resemble the Rorschach problem, such as complex situations in which there is little external guidance so that the individual has to make his or her own way. Also, the adequacy of criterion variables and statistical power present technical challenges to such research.

**Two Applications of Special Interest to Practitioners**

**Thought Disorder and Implications for the Course of Schizophrenic Spectrum Disorder**

A variety of Rorschach summary scores have traditionally been used to assess and conduct research on thought disorder. Early research in this area used the Delta Index of Thought Disorder from which the Thought Disorder Index (Johnston & Holzman, 1979) was developed. Both indexes share many features with CS cognitive special scores and their weighted sum, the WSUM6. There are abundant data that support the irrefutable conclusion that these scales, as well as similar summary scores from the CS and other systems, are related to thinking disturbance, psychoses, and schizophrenic spectrum disorders (e.g., Acklin, 1999; Cadenhead, Perry, & Braff, 1996; Exner, 1991, 1993; Hilsenroth, Fowler, & Padawer, 1998; Kleiger, 1999; Perry & Braff, 1994; Perry, Geyer, & Braff, 1999; Perry, Moore, & Braff, 1995; Viglione, 1999; Weiner, 1966: Perry, Viglione, & Braff, 1992). In addition, research summarized below also associates Rorschach measures with schizophrenic spectrum disorders, genetic and environmental risk factors, psychophysiological processes in schizophrenia, and subsequent outcome.

**Empirical findings.** The Thought Disorder Index is associated with multiple criteria related to biological aspects of schizophrenia and psychosis proneness. Two studies associated the Thought Disorder Index with the Chapman Psychosis Proneness Scales.
(Coleman, Levy, Lenzenweger, & Holzman, 1996; Edell & Chapman, 1979), a measure of magical thinking, perceptual aberration, and social anhedonia (Chapman, Chapman, & Kwapił, 1995; Chapman, Chapman, & Raulin, 1976, 1978; Eckblad, Chapman, & Chapman, 1982). In another study, similar Thought Disorder Index elevations were produced by unaffected relatives of schizophrenics (Shenton, Solovay, et al., 1989). The Thought Disorder Index also predicted the development of psychotic and psychotic-like symptoms over a 2 to 3 year period in a sample of Axis II and affective disorder adult patients over and above information from clinical interviews (O’Connell, Cooper, Perry, & Hoke 1989). Initial Thought Disorder Index scores also demonstrated clinical utility to predict psychotic and psychotic-like symptoms at follow-up. The Thought Disorder Index was about three times greater among psychotic children and high-risk children compared with normal children (Arboleda & Holzman, 1985). In this study, Thought Disorder Index scores in nonpsychotic, hospitalized children were not different from those of normal children. In a family adoptive study, Wahlberg et al. (1997) found that genetic and environmental factors interacted to produce elevated Thought Disorder Index scores.

Empirical data also associate Rorschach measures of thought disorder with more direct biological measures of brain functioning. For example, Daniels, Shenton, et al. (1988) found that patients with right hemisphere damage displayed fragmented thinking, manic patients displayed playfull thinking (as measured by subcomponents of the Thought Disorder Index), and schizophrenic patients used idiosyncratic thinking. A series of studies (Perry & Braff, in press; Perry & Braff, 1994; Perry et al., 1999) demonstrates relationships between (a) the EII and its subcomponents and (b) neuropsychological criterion measures of “gating.” (Gating is the ability to filter out irrelevant information in the environment, and it is considered to be a deficit in the neural circuitry of schizophrenia.) In one study, these information-processing criterion variables and Rorschach responses were measured in a near-simultaneous computer paradigm (Perry et al., 1999). Information-processing deficits were not correlated with symptom rating scales but were correlated with form quality minus (X—%) and cognitive special scores (WSUM66). X—% independently accounted for 57% (p < .001) of the variance in the neuropsychological measure and contributed 35% (p < .01) of the unique variance beyond the clinician-rated Schedule for Positive Symptoms Delusion subscale score. The Schedule for Positive Symptoms Delusion scale was selected among five symptom scales, because it had the greatest correlation with the neuropsychological measure. Thus, X—% demonstrated incremental validity over a variety of symptom scales. In this study the EII was not calculated because of the restriction of the number of responses, due to the demands of the computer-generated paradigm. Nevertheless, these results support the hypothesis that the EII and its subcomponents tap into the cognitive dysfunction that is observed among schizophrenia patients.

Challenges, conclusions, and future research. Garb et al. (2001) took exception to Viglione’s (1999) claim that Rorschach variables can be helpful in determining biological vulnerability for schizophrenia and that these variables have been linked to biological processes associated with schizophrenia. To do so, they isolated two individual studies from the mass of findings in the literature. The research with various Rorschach measures of thought disorder, with and outside of the CS, clearly reveals that the Rorschach can produce data related to the biological aspects of thought disorder and psychosis proneness. Although the Thought Disorder Index administration typically differs from the CS administration and conclusive data need to be collected in regard to the impact of these administrative differences, similar findings across the many Rorschach summary scores of thought disorder suggest that these positive findings should generalize to the EII and other CS-based variables measuring disordered thinking. Of course, research is needed to confirm this logical expectation.

In addition, further research should be conducted to specify the association between the Thought Disorder Index and CS summary scores of thought disturbance, using protocols collected according to CS administrative procedures. Other research goals might be to conduct prospective, longitudinal studies to confirm that these Rorschach variables can predict (a) the emergence or return of psychoses, (b) negative outcomes in the course of schizophrenic spectrum disorders, and (c) the development of these disorders among high-risk children. It would be useful to know whether these associations are moderated by Rorschach variables and collateral non-Rorschach information. For example, simple or preservative records may be associated with false-negative Rorschach thought disorder findings, given that the perseveration evident in Rorschach records is poorly captured by existing scores. Improvements in the measurement of thought disorder within the CS are possible (e.g., Kleiger, 1999; Perry, Potterat, Auslander, Kaplan, & Jeste, 1996), including the possibility of incorporating aspects of the Thought Disorder Index, adding the EII to the CS, developing more meaningful measures of perseveration, and refining procedures for coding ALOG and DR so that they more selectively identify problem-solving failures. This research should also address the sensitivity of Rorschach variables to problem-solving limitations in relatively well-functioning individuals.

Suicide

The CS S-CON is an actuarial variable that yields a single score to assess suicide risk from a composite of 12 variables and ratios. On the basis of all the available data and the cost-benefit of false-positive versus false-negative errors, continued use of the S-CON is in order (Viglione, 1999). It should not be used to rule out suicide risk, but it can be used to increase awareness about self-destructive behavior and suicide.

The Exner and Wylie (1977) study still qualifies as one of the more methodologically rigorous studies of suicide prediction. In the original study of 59 completed adult suicides, the authors found a total S-CON score of 8 positive indices correctly identified 75% of the patients who subsequently died as a result of their suicide attempt. Fowler, Pier, Hilsenroth, Holdwick, and Padower (2001) examined the predictive relationship between the CS S-CON and lethality of suicide attempts within 60 days following administration of the Rorschach. This study uses an ecologically valid and compelling behavioral criterion. Patient records were reliably rated as nonsuicidal (n = 37), parasuicidal (n = 37), or near lethal (n = 30), based on the presence and lethality of self-destructive acts. The S-CON differentiated patients who would later engage in near-lethal suicide attempts from those who would engage in parasuicidal behavior and from those who did not engage in self-destructive behaviors. An S-CON score of 7 or more success-
fully predicted which patients would engage in near lethal suicidal activity relative to parasuicidal patients (overall correct classification [OCC] rate .79), nonsuicidal inpatients (OCC .79), and college students (OCC .89). Logistic regression analysis revealed that an S-CON score of 7 or more was the sole predictor ($p < .01$) of emergency medical transfers due to suicide attempts from among nine demographic, IQ, psychiatric, and Rorschach predictors. Some of these predictors included clinician-rated global psychiatric severity, GAF scores, and the presence of DSM-IV Axis I and II disorders. Thus, this study demonstrates the incremental validity of this Rorschach scale beyond clinician interview and clinician ratings. The total S-CON score also predicted which patients would have a subsequent drug overdose ($r = .32, p = .001$) within 60 days of the Rorschach administration.

To challenge the S-CON, Garb et al. (2001) focused on two studies that used postattempt suicide criteria (Meyer, 1993; Silberg & Armstrong, 1992) and one presentation that has been described only in reviews by Eyman and Eyman (1987, 1991, 1992). However, direct examination of these reviews reveals empirical support for the efficacy for two single sign suicide predictor variables—transparency responses and color-shading blends. These scores are incorporated into shading, form dimension, vista, and color-shading subcomponents of the S-CON. The Rorschachs in the Eyman and Eyman study were administered in the Rapaport-Schafer system (Rapaport, Gill, & Schafer, 1946; Schafer, 1954) but recoded in the CS. The Rapaport-Schafer system does not contain all the same variables as the CS so that inquiry questions do not elicit all relevant information. Translation of these protocols contain all the same variables as the CS so that inquiry questions do not elicit all relevant information. Translation of these protocols to CS variables on an actuarial basis with specific cutoffs may not be justified without further research.

As a whole, the empirical data support the continued use of the S-CON. The decision-making utility of the S-CON may not have been exhaustively evaluated, but based on substantial positive results one cannot ignore the S-CON data if the false-negative error means losing a life. The most enlightened clinical interpretation of an elevated S-CON is to assert that the Rorschach indicates additional suicide risk that might not be recognized through other means. Practice ethics and an S-CON of 7, 8, or more should lead clinicians to consider suicidal risk and to collect additional information about self-destructive risk. Of course, the S-CON may produce false positives in some contexts because of the low base rates of actual suicide (Fimm & Kamphuis, 1995). In real-life clinical decision making, which takes into consideration a wide range of contextual and collateral information, the cost of this false-positive error is minimal compared with the cost of false-negative error (i.e., missing an imminent suicide).

**Clinician Judgment**

**Ecologically Valid Clinical–Actuarial Judgment**

The goal of a large portion of the decisions in formulating the administration procedures of the CS was to minimize examiner and contextual effects identified in earlier research (Exner, 1974). Interrater reliability and temporal consistency data summarized earlier in this article, as well as other research data (e.g., Brinderson, 1995; Cheyette, 1992; Perry, Sprock, Schaible, McDougall, et al., 1995; Viglione & Exner, 1983), suggest that these problems have been minimized. Similarly, a large portion of the interpretive tactics and routines, particularly the interpretive search routines, were developed to minimize problems identified in clinical judgment studies.

The current CS combination of actuarial and clinical methods (Exner, 1991, 1993; Stricker & Gold, 1999; Weiner, 1999) is consistent with thinking in the field of judgment and decision making (e.g., Garb, 1989). This clinical–actuarial method uses procedures whose goal is to minimize cognitive judgment errors and distortions. These procedures are probably ideal in addressing unique questions in clinical contexts where no two cases are alike.

In contrast to the dated clinician judgment analogue research, in which the participant clinician interprets a limited, predefined, and idiosyncratic set of data, the clinician in the field uses data to develop, to reformulate, and to refine referral questions. The clinician also uses expertise to gather or to select accurate data relevant to those questions. In doing so, the examiner constantly interacts with emerging data sets, to change existing questions, probabilities, and base rates (Meyer et al., 1998). Along the way, the examiner brings extreme base rates nearer to levels in which testing data would be more efficient in decision making. In that sense, the single event, low base analogy found in the literature, for example, “Is this client thought disordered?” as addressed by Hunsley and Bailey (1999, p. 272), is an oversimplified analogue to clinical decision making in the field.

In actual clinical practice, decisions with low base rate phenomena rest on the consideration and synthesis of many small data points. An example of such a situation may be the evaluation of a schizophrenic-like thought disorder in outpatient or forensic contexts. Synthesizing such data as history, symptoms, observations of thought disorder, and related characteristics can eventually lead to optimal base rates closer to 50% (Hunsley & Bailey, 1999). In other words, in a patient without the history and observations that might lead to a sufficiently high probability of diagnosing a schizophrenic-like thought disorder, the clinician with positive evidence of confused thinking on the Rorschach might search for historical or behavioral information consistent with (a) excessively elaborated, idiosyncratic, and overly abstract thinking (Franklin & Cornell, 1997; Gallucci, 1989) or (b) thinking problems associated with intrusions of traumatic imagery (Holaday & Whitenberg, 1994; Sloan, Arsenault, Hilsenroth, Harvill, & Handler, 1995). These are alternative conditions that may lead to indices of confused thinking on the Rorschach that have been supported by research.

Such a clinical–actuarial model is consistent with the threshold model described by Rogers (1997), which mitigates errors in evaluating malingering. It also explains why configurational, synthetic approaches, as described by Stricker and Gold (1999) and as demonstrated by Weiner (1999), are fundamental to the clinical–actuarial method. In this form, this method minimizes the possibility of a false-positive actuarial interpretation. This clinical—
actuarial method with the Rorschach uses an ecologically valid and informed understanding of interactive probabilities to increase accuracy of in vivo decision making in applied psychology.

Meehl and Rosen (1955) addressed this issue in the form of the “successive-hurdles” approach. Their example considered separate tests for a brain tumor decision, and they described an increase in hit rates. Meehl and Rosen reported small relative increases in hit rates over base rates and large relative drops in false-negative errors with this procedure. For example, successive tests with (a) 90% true positive and 15% false positive for Test 1, followed by (b) Test 2 with 80% true positive and 10% false positive would increase hit rates from 90% to 96%. In this example, overall errors are cut by more than half, and false-negative errors are reduced by more than three times, from 10% to 3%. The improvement in clinical utility is a function of the cost of these false-positive and false-negative errors. In psychological evaluation, this process contains many more interactive steps. Thus, illness course, symptom report, Rorschach scores for thought disorder, observations of thought-disordered behavior in interview, and other Rorschach variables and observations come together. On the basis of each of these “tests,” the examiner chooses to observe more data to continue to test the hypotheses. Both a yes–no decision and an idiographic, qualitative description of the problem emerge in the typical clinical or forensic Rorschach application.

The clinical–actuarial judgment method is further complicated when one adds the strengths, weaknesses, and predictions of the individual examiner (Stricker & Gold, 1999). Along these lines, and in response to criticisms by Wood, Lilienfeld, Garb, and Nezworski (2000), Garfield (2000) emphasized the skill of the clinician as an important determinant when using the Rorschach in applied practice. To study these issues, more creative and ecologically valid ways of studying the clinician and improving in vivo applied judgment need to be developed both within the quantitative tradition, but also in terms of qualitative research (e.g., Schaich, 1999). This approach is consistent with the APA Psychological Assessment Work Group report (Meyer et al., 1998).

Clinical Judgment: Future Research

Further research on clinical judgment is needed not only for the Rorschach but also for other types of decision making. It might be illuminating to conduct new studies with the CS and non-CS variables, using clinicians who are aware of errors and pitfalls in clinical judgment. As Garb (1998) stated, “most of the judgment studies on the use of projective tests are over 20 years old” (p. 232), with virtually no studies with the modern CS and other Rorschach variables. Such analogue clinical judgment research would have to meet ecological validity challenges, namely, how to reproduce the clinical–actuarial method in the lab, how to motivate clinicians and enforce the real life cost-benefit ratios for accurate versus inaccurate judgments, and how to make the questions and decisions sufficiently idiographic, meaningful, and malleable. Designs must also incorporate better criterion measures. Reproducing the cogwheeling interactive approach (Meyer et al., 1998), vicarious functioning of the respondent and interpreter in facing unbounded problems (Hammond, 1955; Levy, 1963) in an ecologically valid way is critically important.

Lack of ecological validity, poor criteria, and the failure to motivate clinicians all reduce potential effect sizes. If incremental validity is to be addressed quantitatively, an additional challenge is power. Sample sizes must be very large or multiple clinical decisions must be evaluated to be able to rule out false-negative results.

Utility, Cost-Benefit, and the Rorschach

The Rorschach allows us to study samples of behavior collected under similar conditions in different native languages around the world (Erdey & Schaffer, 1999). There is an efficiency to sampling behaviors with a single technique to develop one’s understanding and appreciation of developmental changes across the age span and across all types of disorders and problems.

To make cost-benefit comparisons ecologically valid, one would have to envision a full range of costs and benefits equivalent to the Rorschach. The alternative to the Rorschach would have to address a wide range of clinical, personality, forensic, developmental, cognitive, and neuropsychological constructs and applications across the life span and around the world. The Rorschach allows one to develop a common, experientially based database for the problem-solving practices and predictions of clients from age 5 through old age.7 One could identify many other single purpose scales within a specific content–construct area that might compete well with the Rorschach, but to produce the same benefits, one would have to master and monitor developments in perhaps 50 alternative instruments. Another cost is the expense and hassle of purchasing kits, test blanks, and computer programs and paying per use fees. Also, the Rorschach provides an efficient way to collect a behavior sample outside of interview behavior. Psychol-

---

7 In addressing clinical utility, Garb et al. (2001) presented a factually incorrect and misleading picture of Viglione’s (1999) review of the literature on clinician judgment: “Half of the studies cited by Viglione were published in the 1960s and 1970s, with the remainder of the studies published in the 1980s” (p. 437). Viglione cited five research reports, which met a priori inclusion and exclusion criteria, providing evidence that clinicians can make valid and useful judgments from Rorschach data (Biletz, Jones, & Whitaker, 1982; Cerney, 1984; Dana & Back, 1983; Dudek & Marchand, 1983; Holzman et al., 1974). Viglione cited three additional works, not as research reports, but merely as descriptions for scales used in the five cited studies: (a) Urist (1977) for the Mutuality of Autonomy Scale, (b) Johnston and Holzman (1979) for the Thought Disorder Index, and (c) Masling, Rabid, and Blondheim (1967) for the Rorschach Oral Dependency Scale. Garb et al. (2001) argued that Viglione (1999) consistently ignored negative findings and then listed seven studies supposedly ignored by Viglione (Albert, Fox, & Kahn, 1980; Bilet et al., 1982; Cochrane, 1972; Gadol, 1969; Golden, 1964; Perez, 1976; Turner, 1966). Only the first two of the studies were among the 445 studies selected for review by the editor of this Special Series. These studies addressed the incremental validity of judgments and the relationship between the validity of judgments and confidence or training, rather than simply the issue of judgment. Garb et al. (2001) also asserted that “in fact, positive results have never been obtained for the Rorschach in studies on clinical judgment and...” (italics added) incremental validity, confidence and (italics added) validity, and training and (italics added) validity” (p. 437). To clarify this statement, no single Rorschach study simultaneously offers (a) a demonstration of valid clinician judgment with incremental validity techniques, (b) a demonstration of an association between judges’ avowed confidence and accuracy of judgments, and (c) a demonstration of an association between judges’ training and accuracy of judgment.

---

8 The Rorschach has applications below age 5 for testing developmental issues (Leitchman, 1996).
ogists’ time is valuable, but so is the client’s time; therefore, it is another cost to be entered into cost-benefit analysis. Mastering one test, the Rorschach, is efficient because it eliminates the need for many content specific tests. As a result, we can pick and choose more judiciously and master select instruments within an assessment battery, rather than misusing a large number of tests for all the potential purposes that the Rorschach addresses.

Use of the Rorschach: Past, Present, and Future

Training, Research, and Clinical Use

Repeated surveys of psychological test use over the past 40 years have shown a substantial, consistent, and sustained use of the Rorschach in academic training, research, and clinical settings. Surveys consistently indicate that over 80% of graduate programs teach the Rorschach and that students regard this training as important in developing other clinical skills and in understanding their patients better, and as useful in their practicum-internship training (e.g., since 1995, Camara, Nathan, & Puente, 2000; Hilsenroth & Handler, 1995; Watkins, Campbell, Nieberding, & Hallmark, 1995). Although some surveys report that the Rorschach takes more of the clinician’s time than many other tests, they also report that it is one of the most frequently used (e.g., Camara et al., 2000). Also, 90% of clinical practitioners working in the field expressed a belief that clinical students should be competent in Rorschach assessment (Watkins et al., 1995). This attitude was shared by internship directors (Durand, Blanchard, & Mindell, 1988). In a recent but relatively small survey of 84 (19%) of the 445 APA-approved internship programs in 1997 (Piotrowski & Belter, 1999), internship directors reported that the MMPI-2/MMPI-A, WAIS-R/WAIS-III, and Rorschach were the top three assessment instruments used in internship programs and were considered essential for practicing psychologists. Twenty-seven training directors reported decreased emphasis on projective assessment, whereas 12 reported an increased emphasis. Also, hours of seminar time required by internship programs were virtually identical for projective and self-report assessment measures. Finally, the Rorschach was used more than all five of the behavioral assessment instruments listed by Piotrowski and Belter (1999). A recent survey of practitioners’ test use with adolescents suggests that testing has declined because of managed care constraints, but that the Rorschach has retained its status as the second most used instrument (Archer & Newsom, 2000).

The two most comprehensive recent surveys of predoctoral internships, one including 329 (Stedman, Hatch, & Schoenfeld, 2000) and another including 324 (Clemence & Handler, 2001) Association of Psychology Postdoctoral and Internship Centers sites (most of which were programs accredited by the APA), revealed that internship training directors greatly value the Rorschach as well as integrated test batteries. Again, training directors reported a desire for incoming interns to have had courses or a good working knowledge of the Rorschach. Academic departments that are making decisions about the place of the Rorschach in their curriculum should closely attend to these data.

The Rorschach is the second most frequently researched personality assessment instrument. This consistently high rate of empirical investigation led the authors of a recent review to state the following: “Whether viewed from the perspective of research attention or practical usage, the Rorschach inkblot technique continues to be among the most popular personality assessment methods and predictions about the technique’s demise would appear both unwarranted and unrealistic” (Butcher & Rouse, 1996, p. 91).

Forensic Use

Weiner, Exner, and Schiara (1996) reviewed responses from 93 practitioners who testified in over 4,000 criminal cases, over 3,000 custody cases, and 858 personal injury cases, for a total of almost 8,000 cases. In only 6 of these cases (0.08%) did these respondents report that the integrity of the test was challenged and in only 1 case (0.01%) was it ruled inadmissible. Although Weiner’s sampling technique may have limitations, these data provide a “good empirical basis for concluding that, contrary to whatever opinion may be voiced, the Rorschach is welcome in the courtroom” (pp. 423–424).

To address the question of the legal weight, or the effect the Rorschach has on court judgments, Meloy, Hansen, and Weiner (1997) reviewed court of appeals citations from 1945 to 1995. In 90% of the 194 cases, “the admissibility and weight of Rorschach data were not questioned by either the appellant or the respondent and were important enough to be mentioned and discussed in the legal ruling by the court of appeal” (p. 60). Typically when questions arose, the practitioner’s interpretation or findings derived from the test, rather than the test itself, were challenged. In only two cases (0.8%) was the test criticized as a psychological measure. These data demonstrate that the Rorschach “has authority, or weight, in higher courts of appeal in the United States” (p. 61). In a systematic, comprehensive legal analysis of the admissibility of the Rorschach, McCann (1998) agreed with the courts. He concluded that a focus on the structural summary from the CS “meets professional and legal standards for admissibility of psychometric evidence and expert testimony” (p. 140).

Challenges

In an earlier article in this Special Series, Dawes (1999) referred to the “deficiency” (p. 301) of the Rorschach in relation to the criterion of reasonable certainty used in forensic applications. His argument, concerning the forensic use of the Rorschach, appears to be based on some elusive juxtaposition of incremental validity statistical analysis with the forensic notion that expert testimony should be “incremental” in the forensic sense of aiding the trier of facts. Contrary to his argument and his own data, the evidence from the courts suggests that more direct self-report assessment methods may not be as valuable in forensic adversarial evaluation contexts; therefore, the incremental value of the Rorschach in terms of assisting the trier of facts increases. In fact, Dawes’s own findings indicated that whatever deficiency the Rorschach possessed, the MMPI also possessed because the Rorschach improved on predictions made by the MMPI.

Given these data, the recent task force report by Division 12 (Clinical) of the APA (1999), concerning the development of a

---

9 Citations before 1995 that support the assertions in this section are excluded for the sake of brevity.
model assessment curriculum for graduate training in assessment, is confusing and misguided. In this report, Rorschach training was ranked 95th in importance out of the total 105 topics. The available reliability and validity data, coupled with other contemporary survey material reported earlier, suggest that there is vast discrepancy between the Division 12 Task Force future model curriculum for assessment and actual graduate training needs. Should these changes in graduate assessment training occur, students and future clients will have to pay for the misconceptions of graduate faculty.

Also Hunsley and Bailey (1999) cited a dated survey (Wade & Baker, 1977) of practicing clinicians to claim that practitioners disregard scientific evidence. Actually, there is empirical evidence to show quite the contrary. Beutler, Williams, Wakefield, and Entwistle (1995) recently found that clinicians in private practice value research and put forth effort to understand scientific findings. There is no empirical evidence that indicates that the manner in which psychologists typically use the Rorschach is different from the standardized methods in which they were trained, and no evidence that psychologists who are using the Rorschach are unaware of its research findings.

### Conclusion

The data on the use and perceived value of the Rorschach clearly demonstrate that psychologists find the Rorschach to be of value in routine clinical practice. Some psychologists ask why so many practitioners persist in their use of the Rorschach despite current time pressures and limited fees by third party payers. This becomes even more perplexing when self-report questionnaires and structured interviews may require less of the clinician's time. Considering the persistent use of the Rorschach in the context of the available reliability and validity information leads to some obvious conclusions: One must speculate that clinicians and those in charge of clinical training find it helpful because of its reliability and validity relevant to challenges in applied settings. In other words, clinicians continue to use the test despite its time requirements because of its clinical utility.

A substantial body of research demonstrates that individuals exhibit a defensive bias and manipulate their responses when asked to self-report aspects of their personality or psychopathology. For example, individual differences in self-serving biases have been demonstrated in relation to attributions, self-descriptions, inferences, personality traits, and avoidance of negative affect. Rorschach research supports the view that the Rorschach is most useful in (a) contexts in which the respondent may be unwilling or unable to self-report problems and (b) those in which clinicians are trying to predict real-life behavior (Hiller et al., 1999; Viglione, 1999). In addition to routine clinical contexts, the data suggest that the Rorschach may be especially useful under conditions that might induce response manipulation in self-report. These might include employment, forensic, and custody evaluations, and other settings that involve adversarial examiner-respondent relationship components (Rogers, 1997).

### Methodological Challenges to Validity of the Rorschach

#### Challenges

Rorschach researchers (Exner, 1995; Viglione, 1997) have offered recommendations to improve Rorschach research and, in turn, improve the validity and utility of the test. Others have also repeatedly offered methodological recommendations and standards for Rorschach research, and they have used these standards to reject findings from individual studies supportive of the Rorschach (Garb, 1999; Garb et al., 2001; Hunsley & Bailey, 1999; Wood et al., 2000; Wood, Nezworski, & Stejskal, 1996a, 1996b). These same standards are not fully applied when evaluating studies cited as yielding negative evidence for the Rorschach. Also, impartiality would require applying the same standards to all tests, rather than reserving them specifically for the Rorschach (Garfield, 2000). In the first set of publications for this Special Series in Psychological Assessment, Viglione (1999) met a number of the challenges posed by Rorschach critics. Nevertheless, the positive research findings summarized by Viglione (1999) were met with the same repeated methodological criticisms (Garb et al., 2001). Similarly, other research reports (e.g., Meyer, 1997a; Hilsenroth, Fowler, Padawer, & Handler, 1997) have not been recognized as positive responses to criticisms. In the section below, we examine Garb, Wood, and their colleagues' applications of these criticisms about methodology in Rorschach research.

#### Within and Across Diagnoses

Viglione's (1999) synthesis of the empirical research supports the conclusion that "the Rorschach is useful in examining many clinically relevant criteria and behavior within and across diagnoses" (p. 260). However, in reaction to Stricker and Gold's (1999) summary of the literature, endorsing the usefulness of the Rorschach in evaluating depressed, schizophrenic, or acutely suicidal patients, Garb et al. (2001) asserted that "empirical research suggests that the Rorschach is not useful for all these tasks" (p. 433) and cited only Wood et al. (1996a, 1996b) to bolster this statement. This claim overlooks a great deal of evidence in support of the usefulness of the Rorschach in these disorders and associated phenomena. There are numerous examples of research that support the validity of the Rorschach applied to various problems within and across diagnoses (e.g., see Viglione, 1999).

#### Empirical Data in Peer-Reviewed Journals, Accessible to All

Some critics have maintained that too much of the empirical support for the Rorschach contained within studies in Exner's volumes is not published in peer-reviewed journals and is therefore not available for review. In response to this criticism, Viglione (1999) synthesized research findings from peer-reviewed journals only. He reported considerable empirical support for the test. Viglione focused on the Methods and Results sections, rather than on the theorizing of the authors. Thus, a careful review of the empirical data published in peer-reviewed journals supports the utility of the Rorschach.

#### Selection of Studies

Rorschach critics (e.g., Garb et al., 2001; Wood et al., 2000; Wood, Nezworski, Stejskal, Garven, & West, 1999) have claimed that Rorschach researchers selectively present findings. Meyer (1999) requested that the contributors of this Special Series address 445 articles in five targeted journals published between 1977
and 1997. In response, Viglione (1999) applied objective, meta-analytic type selection criteria to select 128 research reports. In contrast, Hunsley and Bailey (1999) confined their review to 28 studies, whereas Garb et al. (2001) reviewed only 44 of these same 445 articles. In addition, Viglione's selection criteria resulted in the exclusion of supportive Rorschach evidence found not only in Exner's texts, but also in review articles, meta-analyses, and books (e.g., Blatt et al., 1994; Exner, 1993; Gacono & Meloy, 1994; Kleiger, 1999; Mastling, 1986; Weiner, 1966). Most of the negative studies selected by Garb, which Viglione was said to have ignored, were not among those identified by the series editor. On the other hand, Garb et al. uncovered virtually no research supportive of the Rorschach, even though they were provided with the same 445 target articles as the other series contributors. Viglione included negative findings when they occurred.10

Experiment-Wise Alpha Levels

Garb et al. (2001) maintained that the Rorschach literature fails to control experiment-wise alpha levels. For example, Garb et al. (2001) viewed Lerner and St. Peter's (1984a, 1984b) positive findings about borderline personality structure and Rorschach human representations as a flawed study because of the failure to control for false-positive findings. Garb et al. (2001) did not report that 41 of 64 analyses were significant at $p < .05$, and 33 of 64 were significant at $p < .01$.11 Mathematically, these positive results cannot be explained by chance. All other analyses were legitimate, post hoc group contrasts for the 41 significant comparisons.

Criterion Contamination and Diagnosis

Critics have faulted Rorschach researchers for using clinician diagnosis as a criterion (Garb et al., 2001; Wood et al., 2000). For example, Garb (1998) questioned the use of clinicians as criterion judges: "Historically, the clinicians making the indicator ratings have been called criterion judges and their ratings criterion ratings, but these terms are inaccurate because the ratings are fallible" (p. 14). Nonetheless, these same critics have emphasized studies by Archer and his colleagues (Archer & Gordon, 1988; Archer & Krishnamurthy, 1997) to argue against the incremental validity of the Rorschach. The Archer and Krishnamurthy (1997) study used the clinical diagnosis from just one clinician as the sole basis for the diagnosis. Moreover, Archer and Gordon (1988) noted that "psychological test results were frequently available and potentially involved in the diagnostic decisions" (p. 279) so that criterion contamination and bias are possible. Both studies lack reliability statistics for their diagnostic criteria. These studies do not meet the methodological standards imposed by the critics on studies with evidence supportive of the Rorschach. Yet, these studies are cited as negative evidence for the validity of the Rorschach. This is clearly evidence of a double standard used by the critics of Rorschach research.

Implications and Conclusions

We have addressed the Rorschach's strengths and weaknesses and offered research recommendations in 10 focal areas relevant to the clinical utility of the test. In addressing these issues, it is our intention to give readers an informed appreciation for the substantial but often overlooked research basis for the reliability, validity, and utility of the Rorschach. From the broadest point of view, we have demonstrated that examining the empirical findings can only lead to the conclusion that the Rorschach is a valuable test (Viglione, 1999; Weiner, 1996). Interrater reliability by well-trained coders both in the field and in the lab are very strong. Temporal consistency data are good, as are the currently available normative data, though both could be improved. Criterion validity for the test is very strong, and incremental validity and meta-analytic data are present with more data emerging. The current sweeping criticisms of the Rorschach are largely without merit, uninformed, and biased against the test. Thus, the Rorschach, by virtue of the accumulated research, is now "better prepared to stand the tests of reliability and validation" (Exner, 1974, p. 16).

Clinicians persist in using the test because it continues to provide comprehensive and valuable information about their patients. Thus, these clinicians are living proof of the application of these research findings. In turn, the research findings mandate that the test continue to be used. We have broadened the scope of the cost-benefit issues and utility analysis in the application of the test to be more consistent with clinical practice. In doing so, we have identified other obvious benefits to the Rorschach. These include the use of a single test for many applications and the opportunity to collect quantifiable data while simultaneously collecting a behavioral sample in a context that maximizes individual differences in problem solving. The Rorschach serves those who are interested in having a valuable tool to help understand an individual comprehensively and to synthesize information across many domains (Stricker & Gold, 1999). Undoubtedly, the behavior sampling opportunity that occurs during administration promotes these goals and provides another foundation to generalize outside of the consulting room.

Acclaim for the Rorschach has elicited criticism (Wood & Lilienfeld, 1999). Recent critical commentaries of the Rorschach have identified some areas (e.g., normative data and methodological problems in some studies) in which the Rorschach can be improved. Like other tests, the Rorschach has its weaknesses. However, these recent criticisms are undermined by methodological double standards, confirmatory bias, incomplete coverage of the literature, failure to integrate positive contributions over the last 5 years that have specifically addressed earlier criticisms from these authors, and a lack of any original data to support their positions. For example, calls for a moratorium on the use of the Rorschach (Garb, 1999; Garb et al., 2001) clearly contradict empirical findings. Some clinical utility criteria (Hunsley & Bailey, 1999) set such a high and unrealistic standard that implementing them would result in abandoning all testing, interviews, and observation for assessment. Useful criticism and reasonable clinical validity criteria challenge future research rather than mandate our

11 Viglione (1999) is criticized for not including Lerner and St. Peter's (1984a) first article. To be factual, the earlier article was published in Psychiatry, a journal excluded from the series editor's targeted list. It was never considered for inclusion, according to the objective inclusion criteria adopted in this review.
forsaking psychological and neuropsychological assessment altogether. However, an important contribution from critical reviews has been to improve Rorschach research (Acklin, 1999), something that is happening much more quickly than was the case after Rorschach researchers called for similar improvements a few years ago (Exner, 1995; Viglione, 1997).

The informed use of the Rorschach will lead to a more sensitive and helpful understanding of our clients as well as more effective interventions. Continuing research and development of the test as well as the provision for sound training and continuing education in the Rorschach method, as with any assessment measure, are also essential. Nevertheless, the extant empirical research findings necessitate that any scholar must accept the fact that the Rorschach test method, its administration procedures, and its objective rules for quantifying behavior produce valid data and observations.

References


Parker, K. C., Hanson, R. K., & Hunsley, J. (1988). MMPI, Rorschach, and
Perry, W., & Braff, D. L. (in press). The simultaneous measurement of sensorimotor gating and thought disturbance measured in close temporal proximity in schizophrenic patients. *Archives of General Psychiatry*.
Viglione, D. J., & Exner, J. E. (1983). The effects of state anxiety and


Received July 27, 2000

Revision received February 26, 2001

Accepted July 3, 2001

---

Low Publication Prices for APA Members and Affiliates

**Keeping you up-to-date.** All APA Fellows, Members, Associates, and Student Affiliates receive—as part of their annual dues—subscriptions to the *American Psychologist and APA Monitor.* High School Teacher and International Affiliates receive subscriptions to the *APA Monitor,* and they may subscribe to the *American Psychologist* at a significantly reduced rate. In addition, all Members and Student Affiliates are eligible for savings of up to 60% (plus a journal credit) on all other APA journals, as well as significant discounts on subscriptions from cooperating societies and publishers (e.g., the American Association for Counseling and Development, Academic Press, and Human Sciences Press).

**Essential resources.** APA members and affiliates receive special rates for purchases of APA books, including the *Publication Manual of the American Psychological Association,* and on dozens of new topical books each year.

**Other benefits of membership.** Membership in APA also provides eligibility for competitive insurance plans, continuing education programs, reduced APA convention fees, and specialty divisions.

**More information.** Write to American Psychological Association, Membership Services, 750 First Street, NE, Washington, DC 20002-4242.