General Article

THE COMPREHENSIVE SYSTEM FOR THE RORSCHACH: A Critical Examination

By James M. Wood,[1] M. Teresa Nezworski,[2] and William J. Stejskal[3]

[1]*Department of Psychology, University of Texas at El Paso;* [2]*School of Human Development, University of Texas at Dallas; and* [3]*Woodbridge Psychological Associates, Woodbridge, Virginia*

The Comprehensive System (Exner, 1993) is widely accepted as a reliable and valid approach to Rorschach interpretation. However, the present article calls attention to significant problems with the system. First, contrary to common opinion, the interrater reliability of most scores in the system has never been demonstrated adequately. Second, important scores and indices in the system are of questionable validity. Third, the research base of the system consists mainly of unpublished studies that are often unavailable for examination. Recommendations are made regarding research and clinical use of the Comprehensive System.

By the end of the 1960s, heated controversy had developed among psychologists regarding the Rorschach technique. On the one hand, eminent critics (e.g., Eysenck, 1959; Jensen, 1965; Zubin, Eron, & Schumer, 1965) had identified numerous problems with the Rorschach, including (a) lack of standardized rules for administration and scoring, (b) poor interrater reliability, (c) lack of adequate norms, (d) undemonstrated or weak validity, and (e) susceptibility to situational influences.

On the other hand, defenders of the Rorschach questioned the methodology and clinical relevance of existing research and cited the consensus of clinicians regarding the test's value. There was a feeling that many criticisms of the test were "naive and unjust, often fomented from bias, ignorance, or simply a misunderstanding of the method and the principles that led to its exploration by Rorschach" (Exner, 1993, p. 3).

The controversy remained unresolved until the publication of *The Rorschach: A Comprehensive System* (*TRACS;* Exner, 1974). That volume and its subsequent extensions and revisions (Exner, 1978, 1986, 1991, 1993; Exner & Weiner, 1982) appeared to accomplish the remarkable feat of satisfying both sides in the Rorschach controversy. Satisfying the technique's defenders, *TRACS* presented an approach that preserved and strengthened the long clinical tradition regarding the test. The Comprehensive System for administration, scoring, and interpretation borrowed features from the various Rorschach systems already in use (Exner, 1969, 1993). Furthermore, the Comprehensive System incorporated many interpretive terms congenial to psychodynamic conceptualizations (e.g., dependency, narcissism, flight into fantasy).

At the same time, *TRACS* provided enough standardization and empirical data to satisfy the most exacting of scientific critics. Over a period of 20 years, the various editions of *TRACS* (a) established detailed, objective rules for administration, scoring, and interpretation of the Rorschach; (b) catalogued extensive data regarding the interrater reliability of the scales; (c) provided norms and reference data for numerous psychiatric and nonpsychiatric groups, including children; and (d) cited numerous empirical studies to support the validity of Comprehensive System scores.

Thanks to these accomplishments, both sides in the Rorschach controversy appeared to have "won." With the Comprehensive System, the clinicians had their test and the empiricists their data. As Anastasi (1988) commented, "The availability of this system, together with the research completed thus far, has injected new life into the Rorschach as a potential psychometric instrument" (p. 599).

Perhaps because the Comprehensive System represents a "middle ground" (Groth-Marnat, 1990, p. 281), however, it has been scrutinized less carefully than might have been expected for so popular an assessment procedure. The present article focuses attention on three specific issues regarding the Comprehensive System: interrater reliability, validity, and nature of the research base.

Address correspondence to James M. Wood, Department of Psychology, University of Texas at El Paso, El Paso, TX 79968.

INTERRATER RELIABILITY

Interrater Reliability in *TRACS*

Standard textbooks on clinical assessment have commented favorably on the interrater reliability of the Comprehensive System (Erdberg, 1985; Groth-Marnat, 1990; Kaplan & Saccuzzo, 1993). For example, Groth-Marnat (1990) reported,

> During the development of Exner's Comprehensive System, Exner gave particular attention to reliability in developing his different scoring categories. No category was included unless it achieved a minimum .85 level between different scorers. . . . (p. 279)

By incorporating detailed, explicit scoring rules, the Comprehensive System appears to have overcome the problems of rater disagreement that plagued earlier applications of the Rorschach.

The positive opinions of commentators are supported by numerous tables and discussions in *TRACS* (Exner, 1993) that address interrater reliability in the form of "percentage of agreement." For example, percentages of agreement for 32 scores in two different studies are provided in a table titled "Percentage of Coder Agreement for Two Reliability Studies" (Exner, 1993, p. 138). The percentages appear quite high, ranging from 88% to 99%.

However, percentage of agreement has long been recognized as a potentially inadequate and misleading measure of reliability (Cohen, 1960, 1968; Fleiss, 1981; Light, 1971; see also Jensen, 1965). The problem is that percentage of agreement makes no adjustment for agreement by chance and can therefore yield inflated estimates of true consistency among raters. In some circumstances, raters can achieve a very high percentage of agreement even if they score completely at random.

As an example, consider *m* (i.e., the perceived movement of an inanimate object, such as "a swaying tree"), which occurs in about 5% of Rorschach responses (Exner, 1993, p. 260). Imagine that two raters independently rate a large number of Rorschach protocols and randomly assign a score of *m* to 5% of responses. Even though the two raters score at random, they will agree that *m* is present in about .0025 (.05 × .05) of responses and absent in about .9025 (.95 × .95). By chance alone, therefore, the total percentage of agreement between the two raters will be .9050 (.0025 + .9025). This 90% agreement signifies random rather than good interrater reliability.

It is instructive to reexamine the *TRACS* reliability table that was just cited (Exner, 1993, p. 138). The table shows that 20 coders achieved 93% agreement, and 15 raters achieved 95% agreement, when scoring *m*. These percentages are only somewhat higher than the 90% that two coders would be expected to achieve by chance.

This example illustrates why percentage of agreement is an unacceptable measure of interrater reliability. Yet it is the only measure provided in *TRACS* (Exner, 1993) for most scores. Despite the considerable data regarding interrater reliability in *TRACS*, therefore, the statistical approach is inadequate and potentially misleading. If the reliability of Comprehensive System scores is to be evaluated properly, appropriate statistics must be provided, such as kappa, phi, Spearman's rho, or Pearson's *r* (Fleiss, 1981; Nunnally & Bernstein, 1994).

Three additional difficulties may be noted regarding the treatment of interrater reliability in *TRACS*. First, it is unclear how percentage of agreement was calculated for many Comprehensive System scores. Although two computational methods are described (Exner, 1991, pp. 459–460), neither is appropriate for calculating percentage of agreement for individual scores such as *m*.

Second, the percentages of agreement reported in *TRACS* are primarily for individual responses, not total scores. For example, *TRACS* reports that raters achieved 93% to 95% agreement when scoring individual responses for *m*. However, the total number of *m* responses in a protocol appears to be the most clinically relevant figure, insofar as it serves as the basis for Comprehensive System interpretations (Exner, 1991, p. 169). No reliability information regarding total *m* is provided in *TRACS*. The same is true for many other Comprehensive System scores.

Third, for a number of important scores, *TRACS* (Exner, 1991, 1993) provides no measure of interrater reliability, not even percentage of agreement. For example, no reliability information is provided for the Suicide Constellation, Schizophrenia Index, Depression Index, or most individual content categories. The problems noted here are particularly important because the *Standards for Educational and Psychological Testing* of the American Psychological Association (1985, p. 20) state that the reliability of test scores should be fully reported.

Field Interrater Reliability

A distinction may be made between a test's ideal interrater reliability and its field interrater reliability. The ideal reliability is demonstrated by highly trained experts who are performing at their best under optimal conditions. By contrast, field reliability is demonstrated by practitioners who are performing under the time constraints and conditions typical of their work.

The ideal and field reliabilities of a test can be substantially different, as illustrated by recent controversies

in forensic medicine regarding the DNA test for tissue samples. Although reliable when carried out by scientists with appropriate training and equipment, the DNA test can be unreliable when performed under the conditions of some commercial laboratories (Annas, 1992).

The field reliability of the Comprehensive System has received little attention from researchers. However, one study (Exner, 1988) seems relevant. Quizzes were administered to more than 300 alumni of the Rorschach Workshops to evaluate scoring accuracy. The results were "disconcerting" (Exner, 1988, p. 5). Error rates ranged from around 15% for Determinants to about 27% for Special Scores. Exner commented:

> The Rorschach is a good test from which to derive information about personality organization and functioning and it is reasonably easy to interpret, but if the bulk of the interpretation is generated from a Structural Summary that has average error rates similar to those in Table 1, the results will be misleading, and even totally wrong in some cases. (p. 5)

Surprisingly, the results of this study have not been discussed in subsequent editions of *TRACS* (Exner, 1991, 1993).

Reliability of Administration and Recording

Although the present discussion has focused on scoring, the reliability of test administration and recording also merits comment. Before the advent of the Comprehensive System, researchers found that Rorschach scores could be inadvertently contaminated by situational factors (Masling, 1960). For example, the interpersonal style of the test administrator could influence the subject's responses to the cards (Lord, 1950).

To guard against such extraneous influences, the Comprehensive System (Exner, 1993) requires that the test be administered according to narrowly defined procedures and that the administrator write down the subject's responses verbatim. However, the degree to which examiners using the Comprehensive System in clinical settings actually adhere to the standardized administration procedures, or are capable of recording subjects' responses verbatim, has not been studied empirically.

VALIDITY

Strictly speaking, it is imprecise to ask if the Comprehensive System for the Rorschach is valid. The system yields 54 percentages and ratios in a Structural Summary, in addition to numerous other scores, and the validity of each must be established separately. No single article can examine the validity of all scores in the system. Therefore, the present discussion focuses on a subset of scores that, according to *TRACS*, are related to clinically important phenomena, such as psychological symptoms or disorders, level of functioning, or level of stress. Because such scores can influence decision making in important contexts (e.g., clinical and forensic settings), their validity is particularly important.

Exner (1991) has warned that "the Rorschach interpreter usually should not anticipate the discovery of direct diagnostic evidence in the data of the test" (p. 129). However, the Comprehensive System includes several scores (e.g., Egocentricity Index, Adjusted D, Depression Index, Suicide Constellation) that bear directly on clinical decision making. Empirical evidence regarding the validity of these scores is often scant or negative.

The Egocentricity Index and Reflections

In the Comprehensive System, a Rorschach response is scored as a *pair* if it refers to two of the same object (two bears, two lobsters) and as a *reflection* if it refers to a mirror image or reflection ("trees reflected in a lake"). Scores on the Egocentricity Index (EGOI) are derived by multiplying the number of reflection responses by 3, adding the number of Pair responses, and dividing by the total number of Rorschach responses (Exner, 1974, 1993).

According to *TRACS* (Exner, 1974, 1993), the EGOI and reflection responses are indicators of self-esteem, self-focus, and narcissism. In a review of empirical studies, however, we (Nezworski & Wood, 1995) concluded that the EGOI and reflection responses are probably unrelated to self-focus or self-esteem, and that their relationship to narcissistic personality disorder has not been established. The most recent edition of *TRACS* (Exner, 1993) fails to cite numerous negative research findings regarding the EGOI.

D and Adjusted D

Scores for D and Adjusted D are derived by an algorithm from the number of Rorschach responses involving movement, color, achromatic color, texture, shading, and *vista*. According to *TRACS* (Exner, 1993), D and Adjusted D measure the presence of situational stressors and the individual's ability to cope with them. However, in a recent review, Kleiger (1992, p. 293; but see Exner, 1992b) noted "two broad problem areas" with research regarding D and Adjusted D. First, about half of the empirical studies on major structural concepts in the Comprehensive System are unpublished reports and have not appeared in refereed journals. Second, the findings of the published studies appear "equivocal."

The research reviewed by Kleiger (1992, pp. 293–294) included treatment outcome studies, laboratory studies,

and normative data on children (Exner, 1974, 1978, 1986; Exner & Bryant, 1975, 1976; Weiner & Exner, 1991; Wiener-Levy & Exner, 1981). Kleiger noted that some data (Exner, 1974) seemed to be "incomplete" or "described in a confusing manner," making it "difficult for the reader to evaluate adequately the inferences drawn from the research" (p. 293). In addition, Kleiger found that the results of a published study (Wiener-Levy & Exner, 1981) had been interpreted as support for the validity of D and Adjusted D, but in fact contradicted earlier findings (Exner & Bryant, 1975, 1976).

The Depression Index

Like other indices included in the Comprehensive System, the Depression Index (DEPI) is derived by combining scores on several Rorschach variables according to an algorithm. The DEPI has existed in two versions. The first (Exner, 1986) was found to miss a large proportion of depressed patients (Exner, 1991, pp. 22–26; Viglione, Brager, & Haller, 1988). The DEPI has therefore been revised, and the second version appears in the most recent edition of *TRACS*, Volume 2 (Exner, 1991).

According to data from normative and reference samples (Exner, 1993, pp. 260–264, 309–311), the new DEPI is both sensitive and specific. About 75% of inpatient depressives, but only 3% of nonpatient adults, have scores of 5 or higher on the DEPI. By contrast, other researchers have generally failed to find a relationship of the new DEPI to diagnoses or self-report measures of depression among child, adolescent, or adult patients (Ball, Archer, Gordon, & French, 1991; Meyer, 1993). Additionally, in a dissertation study of 109 adult inpatients, Sells (1990/1991) found that scores on neither the original nor the new DEPI were significantly correlated either with diagnoses of depression, assigned according to the third revised edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-III-R; American Psychiatric Association, 1987), or with scores on Scale 2 (Depression) of the Minnesota Multiphasic Personality Inventory (MMPI).

The question arises why the new DEPI performed well with the depressed reference sample in *TRACS* but poorly in replications. Three possibilities may be identified. First, it is not clear that the depressed patients described in *TRACS* were diagnosed accurately. *TRACS* (Exner, 1991, 1993) does not specify the diagnostic criteria or interview procedures used to identify depressed patients in the reference sample, although the "DSM-III-SADS" [*sic*] is mentioned at one point in *TRACS* (Exner, 1991, p. 23).

Second, and perhaps more important, criterion contamination may have influenced diagnoses of depressed patients in the *TRACS* reference sample. Specifically, *TRACS* (Exner, 1991, 1993) does not indicate that diagnoses of depression were made by judges blind to patients' Rorschach scores. Some patients (particularly those with ambiguous symptoms) may have been classified as depressed on the basis of their Rorschachs, creating a spuriously high correlation between diagnoses and DEPI scores.

Third, the DEPI appears to have fallen victim to an old nemesis of empirically derived indices: shrinkage during cross-validation. The DEPI was developed using actuarial methods (Exner, 1991). That is, variables that discriminated depressed from nondepressed subjects in the Rorschach subject pool were identified empirically. When variables are selected in this manner, they normally show less predictive power when applied to new groups, a phenomenon known as shrinkage (Meier, 1994, p. 116; Wiggins, 1988, pp. 46–49). The proportion of hits may be high in the original group of subjects (the derivation sample) but dismally low in the new group (the cross-validation sample). To the disappointment of many a researcher who has used the actuarial method, variables that appear promising in a derivation sample may have no predictive power in cross-validation groups.

The fact that the DEPI predicts depression among subjects in the Rorschach subject pool is to be expected. The actuarial approach ensures such a result in a derivation sample. The critical question is whether the DEPI can predict depression in new samples. The results of Ball et al. (1991), Meyer (1993), and Sells (1990/1991) suggest that it cannot. The shrinkage of the DEPI in cross-validation may be so great that the scale lacks meaningful predictive power.

The Suicide Constellation

The Suicide Constellation (S-CON) was also developed using actuarial methods and subsequently revised. The Rorschach protocols of patients who had committed suicide and of control subjects were compared (Exner & Wylie, 1977). A cluster of 11 variables, the original S-CON, was found to discriminate between suicides and other patients, with an optimal cutoff score of 8 or above.

During cross-validation, shrinkage in an actuarial scale's predictive power is likely to be substantial if the original sample of subjects was small and the number of variables considered for inclusion in the scale was large. Because the original study of the S-CON (Exner & Wylie, 1977) involved a small sample of subjects (59 suicides) and considered a large number of variables (apparently more than 100), substantial shrinkage was to be expected when the scale was cross-validated on a new sample.

A cross-validation of the S-CON has been reported (Exner, 1986, pp. 411–416; 1993, pp. 342–345), but sur-

prisingly, no shrinkage occurred. In the original sample (Exner & Wylie, 1977), 75% of suicide patients and 0% of nonpatients had S-CON scores of 8 or above. In the cross-validation sample, the corresponding figures were 74% and 0%. Contrary to what might have been expected, sensitivity and specificity were undiminished.

In contrast to the excellent performance of the S-CON as reported in *TRACS* (Exner, 1986, 1993), S.K. Eyman and J.R. Eyman (1987; see also J.R. Eyman & S.K. Eyman, 1992) found that in a sample of 50 patients who committed suicide, only 1 had an S-CON score of 8 or above. They (J.R. Eyman & S.K. Eyman, 1992) concluded that the S-CON was "ineffective in predicting suicidal behavior" (p. 189).

These findings are not definitive: Rorschachs were scored according to the Comprehensive System but administered according to the system of Rapaport, Gill, and Schafer (1945), with inquiry immediately following each card. Furthermore, the length of time between testing and subjects' suicide varied from less than 90 days to more than 2 years. Thus, although the results of S.K. Eyman and J.R. Eyman (1987) raise doubts concerning the validity of the S-CON, further research is needed to resolve the issue. It should be added that a second version of the S-CON has been developed, based on minor modifications of the original scale (Exner, 1993). However, the predictive power of the new version has not yet been demonstrated in a cross-validation study.

The Influence of Response Frequency

As has long been recognized, many Rorschach scores are correlated with *R*, the number of responses made by the subject (Fiske & Baughman, 1953; Meyer, 1992). Because *R* is influenced by intelligence, educational level, and social class, its influence on other scores is problematic (see discussion in Anastasi, 1988). Some commentators (e.g., Groth-Marnat, 1984) believe that the Comprehensive System has eliminated this problem by adjusting for *R* or using ratios.

In fact, however, many of the clinical scores and indices of the system are unadjusted. For example, the S-CON, Schizophrenia Index (SCZI), DEPI, Coping Deficit Index (CDI), Hypervigilance Index (HVI), and Obsessive Style Index (OBS) are either unadjusted or only partially adjusted for *R*.

The most thorough investigations of this topic are those of Meyer (1992, 1993; but see Exner, 1992a). Meyer (1993) found that among psychiatric patients, *R* was significantly correlated with the S-CON, SCZI, DEPI, CDI, HVI, and OBS. The size of the correlations ranged from .25 for the S-CON to .60 for the HVI. In addition, Meyer concluded that the indices might be valid for some values of *R* but not others. For example, the DEPI correlated significantly with depression scales of the MMPI-2 among patients with high *R*, but not among those with average or low *R*, or among the patient sample as a whole.

Interpretations Based on a Single Response

In a recent refinement (Exner, 1991), the Comprehensive System provides interpretive statements for particular test scores or test score combinations. In several instances, these statements are based on a single test response:

- If even a single reflection response appears in a Rorschach protocol, an interpretive statement of the Comprehensive System indicates that "a nuclear element in the subject's self-image is a narcissistic-like feature that includes a marked tendency to overvalue personal worth" (Exner, 1991, p. 173).
- A single Human Experience response ("two people who are deeply in love, gazing longingly at each other") is interpreted to mean that the subject "attempts to deal with issues of self-image and/or self-value in an overly intellectualized manner that tends to ignore reality" and is likely to have "ideational impulse control problems" (Exner, 1991, p. 176).
- A single Food response ("a Thanksgiving turkey already eaten") is interpreted to mean that the subject "can be expected to manifest many more dependency behaviors than usually expected. . . ." If the subject is also "passive," then "it is reasonable to conclude that a passive-dependent feature is an important core component in the personality structure of the subject" (Exner, 1991, p. 184).

There are two reasons to question Comprehensive System interpretations that identify a "core component" of personality on the basis of a single test response. First, any single-sign approach to personality assessment is limited by that sign's reliability. Because a single behavioral indicator or test response is unlikely to have high reliability, most diagnostic and assessment approaches employ multiple indicators or signs. For example, in the fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV; American Psychiatric Association, 1994), a diagnosis of narcissistic personality disorder or dependent personality disorder requires the presence of five criteria. The DSM-IV implicitly recognizes the unreliability of any individual sign, and therefore requires that diagnoses be based on multiple signs. By contrast, the Comprehensive System identifies narcissism as a "nuclear element" or dependency as a "core component" of personality on the basis of a single reflection or Food response.

Second, if clinical decisions about a subject's personality are to be based on a single test response, that response should have high validity. However, the single-sign interpretations of the Comprehensive System often lack well-documented validity:

- As already discussed, we (Nezworski & Wood, 1995) concluded that the reflection response is probably unrelated to self-esteem or self-focus.
- *TRACS* (Exner, 1991, 1993) provides no empirical evidence that a single Human Experience response indicates an "overly intellectualized" personality style or "ideational impulse control problems."
- The most recent edition of *TRACS* (Exner, 1993, pp. 438–439) reports three studies that have found a relationship between Food responses and dependency: One is an unpublished study of the Rorschach Workshops, and the remaining two are described without scholarly citation. The discussion in *TRACS* is brief: Two of the studies are each summarized in a single paragraph, and the third is summarized in a single sentence. No statistical tests are reported. An earlier edition of *TRACS* (Exner, 1974, p. 303) discussed the relationship between Food responses and dependency, cited two studies with contradictory findings, and concluded that the evidence was "at best, limited."

Incremental Validity

Although the present discussion has focused on evidence of validity, the issue of incremental validity merits comment (Sechrest, 1963; Wiggins, 1988, pp. 250–251). Specifically, do Comprehensive System scores contribute information relevant to clinical decision making beyond what can be gained from a diagnostic interview and consideration of MMPI-2 scores?

For example, cross-validation studies (Archer & Gordon, 1988; Meyer, 1993) have confirmed that scores on the SCZI of the Comprehensive System (Exner, 1993) are related to diagnoses of schizophrenia among adult and adolescent psychiatric patients. However, Archer and Gordon (1988, p. 285) found that when optimal cutoff points were used, Scale 8 (Schizophrenia) of the MMPI classified schizophrenic and nonschizophrenic adolescents more accurately than did the SCZI. Furthermore, classifications based on Scale 8 and the SCZI combined were not significantly more accurate than classifications based on Scale 8 alone.

The study by Archer and Gordon (1988) is not cited here as proof that the SCZI lacks incremental validity. More research is necessary before the issue can be resolved one way or the other (Archer & Krishnamurthy, 1993). However, the example illustrates the point that even valid indicators, such as the SCZI, may sometimes add very little to existing sources of information. In the future, research may more thoroughly examine the incremental validity of Comprehensive System scores (Archer & Krishnamurthy, 1993).

RESEARCH BASE OF THE COMPREHENSIVE SYSTEM

The various editions of *TRACS* (Exner 1974, 1978, 1986, 1991, 1993; Exner & Weiner, 1982) provide numerous citations to unpublished studies of the Rorschach Workshops. For example, the various editions of *TRACS* cite 156 of Exner's works. Twenty-seven of these (17%) have been published in peer-reviewed journals, whereas ninety-nine (63%) are unpublished studies of the Rorschach Workshops. It may be said that the Workshops Studies constitute the broad empirical foundation of the Comprehensive System.

In response to queries, the Rorschach Workshops has informed us that more than 1,000 Workshops Studies were undertaken from 1968 to 1990. Thus, the studies cited in *TRACS* constitute less than 10% of the total. It is worth noting that 1,000 studies, each examining a single outcome variable, would yield 50 statistically significant results by chance alone.

Many readers of *TRACS* are probably under the impression that the Workshops Studies are actual documents that can be examined by other scholars. However, this impression is often mistaken. In preparation for writing this article, we requested 23 of the Workshops Studies cited in *TRACS*. Letters from the Rorschach Workshops informed us that some of the Workshops Studies were not in their files. The methods and results of the remaining studies either had not been formally written up or could not be released. We were informed that the Rorschach Workshops could provide raw data related to specific questions, but that we might have to pay for computer costs.

Based on the response from the Rorschach Workshops, we arrived at three conclusions: First, most of the Workshops Studies cited in *TRACS* are apparently research projects, not written documents. Second, the methods and results of many Workshops Studies are unavailable for public examination, except for summaries in articles and books. Third, scholars who seek to examine the studies often cannot obtain reports or quantitative analyses. Raw data may be obtained, but the expense of doing so may be substantial.

DISCUSSION

Current acceptance of the Comprehensive System seems to be based on several implicit and explicit assumptions:

1. In contrast to earlier Rorschach systems, the Comprehensive System has demonstrated a high level of interrater reliability.
2. The clinical interpretations generated by the Comprehensive System are consistent with research findings and have been well validated.
3. The various indices of the Comprehensive System have performed well in cross-validation samples.
4. The research base of the Comprehensive System is well documented and has been scrutinized and confirmed by independent scholars.

As the present article indicates, these assumptions are mistaken. Basic issues regarding the reliability and validity of the Comprehensive System have not been resolved. Clinical interpretations generated by the system are often either inadequately supported or inconsistent with research findings. In addition, the empirical underpinnings of the Comprehensive System are themselves in doubt. In particular, the unpublished Rorschach Workshops Studies, which constitute the main empirical support for the system, are often unavailable for examination and review.

Although the present discussion has focused on a limited number of issues, the problems identified are so fundamental as to raise questions about the Comprehensive System as a whole. We suggest that psychologists thoughtfully review the relevant scientific literature before relying on a particular Comprehensive System score or index for clinical assessment. Ethical and practical considerations (American Psychological Association, 1985) discourage reliance on test scores whose reliability or validity has not been established, particularly if the resulting decisions may have a major impact on clients' lives (e.g., in clinical or forensic settings).

In our opinion, there is need for examination of the Comprehensive System by researchers. Fundamental issues of reliability and validity are yet to be resolved.

- The interrater reliability of the various Comprehensive System scores needs to be studied under both ideal and field conditions. Also needed are field studies of the administration, scoring, and recording practices of psychologists who use the Comprehensive System.
- In the future, the validity of scores in the Comprehensive System should be tested rigorously rather than assumed. The present article has focused on clinical scores. Similar considerations apply to the validation of nonclinical scores. Consistent with our earlier conclusions (Nezworski & Wood, 1995), we suggest that clinical validation studies (a) use well-defined and rigorous diagnostic criteria (e.g., from structured interviews based on the DSM-III-R or DSM-IV), (b) ensure that diagnosticians are blinded to Rorschach results, (c) use other tests in addition to the Rorschach to measure relevant constructs, (d) examine the relationship between Rorschach scores and ecologically valid, real-world behaviors, and (e) report findings thoroughly and completely, including measures of diagnostic performance (see Kessel & Zimmerman, 1993).

Acknowledgments—We gratefully acknowledge Richard Bootzin, Jane Bretschger, Russell Clark, Robyn Dawes, Lynette Heslet, Pamela McCauley, Lee Sechrest, Leonore Simon, and Randy Whitworth for their assistance and comments on this article.

REFERENCES

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

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

American Psychological Association. (1985). *Standards for educational and psychological testing*. Washington, DC: Author.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Annas, G.J. (1992). Setting standards for the use of DNA-typing results in the courtroom: The state of the art. *New England Journal of Medicine, 326*, 1641, 1644.

Archer, R.P., & Gordon, R.A. (1988). MMPI and Rorschach indices of schizophrenic and depressive diagnoses among adolescent inpatients. *Journal of Personality Assessment, 52*, 276–287.

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

Ball, J.D., Archer, R.P., Gordon, R.A., & French, J. (1991). Rorschach depression indices with children and adolescents: Concurrent validity findings. *Journal of Personality Assessment, 57*, 465–476.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213–220.

Erdberg, P. (1985). The Rorschach. In C.S. Newmark (Ed.), *Major psychological assessment instruments* (pp. 65–68). Boston: Allyn and Bacon.

Exner, J.E. (1969). *The Rorschach systems*. New York: Grune and Stratton.

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

Exner, J.E. (1978). *The Rorschach: A comprehensive system: Vol. 2. Current research and advanced interpretation*. New York: Wiley.

Exner, J.E. (1986). *The Rorschach: A comprehensive system: Vol. 1. Basic foundations* (2nd ed.). New York: Wiley.

Exner, J.E. (1988). Scoring issues. *1988 Alumni Newsletter*, pp. 4–8.

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

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

Exner, J.E. (1992b). Some comments on "A conceptual critique of the EA:es comparison in the Comprehensive Rorschach System." *Psychological Assessment, 4*, 297–300.

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

Exner, J.E., & Bryant, E. (1975). *Pursuit motor performance and the Ea-ep relation* (Workshops Study No. 222). Unpublished manuscript, Rorschach Workshops, Asheville, NC.

Exner, J.E., & Bryant, E. (1976). *Mirror star tracing as related to different Rorschach variables* (Workshops Study No. 240). Unpublished manuscript, Rorschach Workshops, Asheville, NC.

Exner, J.E., & Weiner, I.B. (1982). *The Rorschach: A comprehensive system: Vol. 3. Assessment of children and adolescents*. New York: Wiley.

Exner, J.E., & Wylie, J. (1977). Some Rorschach data concerning suicide. *Journal of Personality Assessment, 41*, 339–348.

Eyman, J.R., & Eyman, S.K. (1992). Personality assessment in suicide prediction. In R.W. Waris, A.L. Berman, J.T. Maltsberger, & R.I. Yufit (Eds.),

Assessment and prediction of suicide (pp. 183–201). New York: Guilford Press.

Eyman, S.K., & Eyman, J.R. (1987, August). *An investigation of Exner's Suicide Constellation.* Paper presented at the meeting of the American Psychological Association, New York. (Available from J.R. Eyman, Menninger Clinic, Box 829, Topeka, KS 66601-0829)

Eysenck, H.J. (1959). The Rorschach Inkblot Test. In O.K. Buros (Ed.), *The fifth mental measurements yearbook* (pp. 276–278). Highland Park, NJ: Gryphon Press.

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

Fleiss, J.L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley & Sons.

Groth-Marnat, G. (1984). *Handbook of psychological assessment.* New York: Van Nostrand Reinhold.

Groth-Marnat, G. (1990). *Handbook of psychological assessment* (2nd ed.). New York: Wiley & Sons.

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

Kaplan, R.M., & Saccuzzo, D.P. (1993). *Psychological testing* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Kessel, J.B., & Zimmerman, M. (1993). Reporting errors in studies of the diagnostic performance of self-administered questionnaires: Extent of the problem, recommendations for standardized presentation of results, and implications for the peer review process. *Psychological Assessment, 5,* 395–399.

Kleiger, J.H. (1992). A conceptual critique of the EA:es comparison in the Comprehensive Rorschach System. *Psychological Assessment, 4,* 288–296.

Light, R.J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin, 76,* 365–377.

Lord, E. (1950). Experimentally induced variation in Rorschach performance. *Psychological Monographs, 64*(Whole No. 316).

Masling, J. (1960). The influence of situational and interpersonal variables in projective testing. *Psychological Bulletin, 57,* 65–85.

Meier, S.T. (1994). *The chronic crisis in psychological measurement and assessment.* San Diego: Academic Press.

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

Meyer, G.J. (1993). The impact of response frequency on the Rorschach constellation indices and on their validity with diagnostic and MMPI-2 criteria. *Journal of Personality Assessment, 60,* 153–180.

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

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Rapaport, D., Gill, M., & Schafer, R. (1945). *Diagnostic psychological testing.* Chicago: Yearbook Publishers.

Sechrest, L. (1963). Incremental validity: A recommendation. *Educational and Psychological Measurement, 23,* 153–158.

Sells, J.E. (1991). A validity study of the DEPI index: The Rorschach Comprehensive System (Doctoral dissertation, University of Utah, 1990). *Dissertation Abstracts International, 51,* 5590B.

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

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

Wiener-Levy, D., & Exner, J.E. (1981). The Rorschach EA-ep variable as related to persistence in a task frustration situation under feedback conditions. *Journal of Personality Assessment, 45,* 118–124.

Wiggins, J.S. (1988). *Personality and prediction: Principles of personality assessment.* Malabar, FL: Krieger.

Zubin, J., Eron, L.D., & Schumer, F. (1965). *An experimental approach to projective techniques.* New York: Wiley.

(RECEIVED 8/10/94; ACCEPTED 12/11/94)