Field Reliability of Comprehensive System Scoring in an Adolescent Inpatient Sample

Robert E. McGrath, David L. Pogge, John M. Stokes, Ana Cragnolino, Michele Zaccario, Judy Hayman, Teresa Piacentini and Douglas Wayland-Smith

Assessment 2005; 12; 199
DOI: 10.1177/1073191104273384

Field Reliability of Comprehensive System Scoring in an Adolescent Inpatient Sample

Robert E. McGrath
Fairleigh Dickinson University

David L. Pogge
Four Winds Hospital, Katonah, New York

John M. Stokes
Pace University

Ana Cragnolino
Michele Zaccario
Judy Hayman
Teresa Piacentini
Douglas Wayland-Smith
Fairleigh Dickinson University

The extent to which the Comprehensive System for the Rorschach is reliably scored has been a topic of some controversy. Although several studies have concluded it can be scored reliably in research settings, little is known about its reliability in field settings. This study evaluated the reliability of both response-level codes and protocol-level scores among 84 adolescent psychiatric inpatients in a clinical setting. Rorschachs were originally administered and scored for clinical purposes. Among response codes, 87% demonstrated acceptable reliability (> .60), and most coefficients exceeded .80. Results were similar for protocol-level scores, with only one score demonstrating less than adequate reliability. The findings are consistent with previous evidence, indicating reliable scoring is possible even in field settings.

Keywords: Rorschach; interrater reliability; Comprehensive System; field reliability

The interrater reliability of the Comprehensive System for the Rorschach has generated a surprising amount of debate in recent years. The roots of this debate can be traced to the original reliability evidence described by Exner (1993). This consisted of two studies involving the Comprehensive System codes, that is, the categorization of individual responses according to location, contents, and so forth. In both studies, the statistic reported was percentage of agreement with the correct coding, in which the correct coding was determined through consensus among at least

Portions of this article were presented previously at the Midwinter Meeting of the Society for Personality Assessment in San Antonio, Texas, in March 2002. This project was supported in part by a grant from the Fairleigh Dickinson University Grant-in-Aid Program. We gratefully acknowledge Samuel Klagsbrun, Martin Buccolo, Janet Segal, and the administration of Four Winds Hospital, Katonah, New York, for their support in the collection of data for this article. The authors are solely responsible for its content. Correspondence concerning this article should be addressed to Robert McGrath, School of Psychology T-WH1-01, Fairleigh Dickinson University, Teaneck, NJ 07666; e-mail: mcgrath@fdu.edu.

Assessment, Volume 12, No. 2, June 2005 199-209
DOI: 10.1177/1073191104273384

three individuals on the staff of Rorschach Workshops (J. E. Exner, personal communication, May 1, 2002). Exner consistently found percentage of agreement values in excess of .85.

The controversy began when Wood, Nezworski, and Stejskal (1996a, 1996b) and McDowell and Acklin (1996) objected to the use of percentage of agreement as the basis for evaluating interrater reliability, arguing that the generally more conservative chance-corrected agreement (kappa) is a more appropriate statistic. The validity of this objection is universally accepted, even among advocates of the Comprehensive System (e.g., Meyer, 1997). Kappa is the appropriate reliability statistic for Rorschach codes, whether the goal is to evaluate reliability between two or more raters or the reliability of one or more raters with the correct scoring (see Light, 1971).

Wood et al. (1996a, 1996b) also noted that adequate interrater reliability in response-level codes does not necessarily ensure adequate reliability in protocol-level scores, the aggregated outcomes based on the complete record of response codes for a respondent such as WSum6. Because interpretation is based on summary scores, the reliability of the latter can be considered more important than that of the former.

To place this criticism in a broader context, review of the Rorschach reliability literature (particularly Meyer et al., 2002), as well as consideration of the manner in which the Comprehensive System is used clinically, suggests at least four possible classes of targets for interrater reliability studies. Two of these targets are at the response level, whereas two are at the protocol level (see Figure 1). At the level of the individual response, reliability can be evaluated for each coding decision, such as the presence of inanimate movement or the developmental quality of the response. A number of studies since Exner's (1993) original investigations have examined the reliability of individual codes using kappa (e.g., Acklin, McDowell, Verschell, & Chan, 2000; Meyer et al., 2002; Shaffer, Erdberg, & Haroian, 1999).

Some studies have instead or in addition examined the issue of agreement within segments of response coding, such as the overall reliability of determinant or content coding (McDowell & Acklin, 1996; Meyer, 1997; Meyer et al., 2002). Again, kappa is considered the reliability statistic of choice, indicating the chance-corrected rate of exact agreement across raters for the segment.

Meyer (1997) noted one benefit to basing reliability on the segment rather than the code. Where percentage of agreement tends to be inflated by low base rate events, kappa tends to be reduced (Zwick, 1988). Some have argued this is a reasonable feature for a reliability measure: Skew in a variable reduces its variance, so unreliable variability will tend to represent a larger proportion of total variability (Shrout, Spitzer, & Fleiss, 1987). Many Comprehensive System codes occur infrequently, resulting in values for kappa that are less than desirable. In contrast, the distribution of perfect agreement within a segment tends to be less unbalanced, so base rate will have less effect on segment reliability. However, Meyer et al. (2002) recognized segments offer only a global or unfocused analysis of response reliability. On the statistical level, because the segment outcome represents an aggregate of multiple decisions, it could be suggested that a more appropriate reliability statistic would be weighted kappa based on the degree of disagreement between the raters, although this would substantially increase the computational complexity of segment reliability estimates. Finally, Wood, Nezworski, and Stejskal (1997) also noted that as a response-level analysis, evidence of reliability in the segment still did not ensure adequate reliability at the more important protocol level.

Previous studies examining reliability at the protocol level have focused on the dimensional Comprehensive System scores (e.g., Acklin et al., 2000; Meyer et al. 2002; Viglione & Taylor, 2003), for which the optimal reliability statistic is the intraclass correlation coefficient (ICC). A second possible target at the level of the protocol is respondent classification based on cut points or decision rules for the scores. To date, no studies have reported on the reliability of Comprehensive System protocol-based classifications. However, several studies have discussed the potential for unreliability in placement based on errors in the scoring of intelligence tests (e.g., Slate, Jones, Coulter, & Covert, 1992; Whitten, Slate, Jones, & Shine, 1994). If such studies were to be conducted, the optimal reliability statistic would again be the kappa coefficient.

FIGURE 1
Four Possible Classes of Targets for Comprehensive System Reliability Analyses

		Level of Target: Response	Level of Target: Protocol
Type of Target	Original	Class: Codes Examples: m, FM, COP (presence/absence) Reliability statistic: Kappa coefficient	Class: Scores Examples: COP (frequency), 3r + (2)/R Reliability statistic: Intraclass correlation coefficient
	Modified	Class: Segments Examples: Location, Determinants Reliability statistic: Kappa coefficient	Class: Categories Examples: Ambitent, 3r + (2)/R < .33 Reliability statistic: Kappa coefficient

A third issue of continuing debate is the extent to which the existing literature substantiates the interrater reliability of the Rorschach. Guarnaccia, Dill, Sabatino, and Southwick (2001) asked graduate students and practicing clinicians who use the Rorschach to code responses for which the correct coding was available from standard Comprehensive System texts. They found accuracy rates for patient responses were often unacceptable. However, the statistic used to rate accuracy was idiosyncratic, for example, involving subtraction for mismatches. Not only is this not a valid reliability statistic, but the results cannot be directly compared to those of any other study on interrater consistency.

Several studies have now been completed specifically evaluating Comprehensive System coding and scoring using more appropriate reliability statistics. These studies have consistently used a value of approximately .60 as the minimum acceptable level for reliability based on prior recommendations concerning benchmarking (Fleiss, 1981; Landis & Koch, 1977; Shrout, 1998). Acklin et al. (2000) found that 66 of 88 kappa coefficients (75.0%) were ≥ .60 in a sample of nonpatients, whereas 81 of 89 (91.0%) kappa coefficients met this criterion in patients. Out of 34 kappa coefficients reported by Shaffer et al. (1999), 24 (70.6%) met this criterion. Meyer et al. (2002) found that 84 of 108 kappa coefficients (77.8%) surpassed this criterion when one coder was inexperienced, but among experienced raters all kappa values were considered acceptable.

As noted previously, kappa is particularly sensitive to unbalanced base rates. In fact, some of the studies described did not report kappa values when the base rate was < .01 (Meyer et al., 2002; Shaffer et al., 1999). Results are generally more consistent for segment and score analyses. McDowell and Acklin (1996) found eight of nine segments surpassed the criterion, whereas two subsequent studies have reported kappa values in excess of .60 for all major response segments (Meyer, 1997; Meyer et al., 2002). Three reliability studies have examined score reliability for the Comprehensive System as a whole. Acklin et al. (2000) found 84.1% of ICCs in their nonpatient sample and 90.6% in their patient sample were ≥ .60. In contrast, no more than 3.8% of ICCs were < .60 in any of four samples described by Meyer et al. (2002), whereas Viglione and Taylor (2003) found 67 of 68 ICCs met the criterion. Comprehensive System validity studies offer a secondary source for reliability estimates that generally corroborate the positive findings of studies designed to evaluate reliability in general (e.g., Hartmann, Wang, & Berg, 2003; Stokes, Pogge, & Powell-Lunder, 2003). It should be noted, however, that validity studies only provide reliability data for the variables being validated, so poor reliability would probably render the study unpublishable, and focusing on a few variables should improve interrater reliability.

Wood, Nezworski, Lilienfeld, and Garb (2003) have raised two objections to the general conclusion among these studies that the interrater reliability of Comprehensive System coding and scoring is acceptable. First, they proposed .60 is too liberal a standard for acceptable reliability. They recommended the standard of .85 used in Exner's (1993) original studies or the .90 level found in some intelligence tests. However, it is important to remember that Exner used this standard in relation to percentage of agreement, a statistic that is easily inflated compared to true measures of reliability. Although reliability levels of .85 are desirable and are sometimes achieved in practice, it is common for instruments commonly used in clinical settings to demonstrate reliability values lower than .85. For example, although the major components of the test demonstrate reliabilities > .90, 11 of 20 subscales of the Wechsler Intelligence Scale for Children (Wechsler, 2003) demonstrate internal reliabilities less than .85. It is important to remember that a reliability coefficient of .60 would still allow for a validity coefficient as high as .77, another desirable outcome rarely achieved in practice.

Their second objection has to do with some of the samples used in these studies. Meyer et al. (2002) included several samples that were composites from multiple settings. If there were systematic variation across settings in the frequency of codes (either because of differences in rating standards or differences in populations), reliability statistics should overestimate the reliability one would find using cases from a single setting. The objection particularly merits consideration because the subsequent Viglione and Taylor (2003) study also used a composite sample. Future studies with composite samples should consider site and rater pool as potential sources of systematic variability. However, Wood et al. (2003) failed to note that mean and median reliabilities for the Meyer et al. (2002) composite samples were very similar to those from their single-site samples, and the distributions of ICCs were also consistent. In the one study where evidence is available, the results do not support Wood et al.'s (2003) hypothesis.

Wood et al. (1996a, 1996b) raised one more criticism of Comprehensive System reliability that is particularly relevant to the present study. They noted the lack of evidence concerning the reliability of coding and scoring in field settings. Hunsley and Bailey (1999) expanded on this issue. Although acknowledging that what they referred to as field reliability is an issue for any observational instrument, Hunsley and Bailey identified several factors that could work to reduce the field reliability of the Comprehensive System in particular. Among observational mea-

sures, Comprehensive System coding is particularly complex. In addition, a tradition of alternate scoring systems and idiosyncratic scoring may render users of the Rorschach more likely to deviate purposefully from standardized scoring. On the other hand, McGrath (2003) hypothesized in response that the attention focused on Comprehensive System scoring accuracy, including the publication of two manuals devoted to scoring guidelines (Exner, 2001; Viglione, 2002), may have encouraged greater diligence in coding accuracy than is true for users of other behavioral observation procedures.

The evidence is insufficient to draw a conclusion either way on this issue. The coding of a rater who is aware the data are being collected for research purposes does not necessarily provide an acceptable analog for coding in clinical settings, because the rater is probably aware that reliability will be assessed. Research on the accuracy of intelligence testing field scoring consistently demonstrates substantially poorer outcomes than would be assumed given the reliabilities generated during scale development (e.g., Slate et al., 1992; Whitten et al., 1994). Ideally, one would like to find settings where two clinicians complete the coding independently under the impression they are doing so for purely clinical purposes, but this ideal is unlikely to occur for economic reasons. The next best alternative would compare an original scoring completed for clinical reasons with a second scoring completed for the evaluation of interrater reliability.

To date, only one such study of field reliability has been published. Meyer et al. (2002) presented reliability statistics for 69 adult protocols in which the original scoring was completed during the course of normal clinical activities (labeled Sample 4).[1] They reported a mean kappa for a subset of codes of .89, whereas the mean ICC for scores based on those codes was .92. Furthermore, this is not one of Meyer et al.'s composite samples. Although this is good evidence for field reliability, it warrants replication. The current study was conducted to evaluate whether the positive findings reported by Meyer et al. occur in other settings as well. To our knowledge, it is the first study to examine the field reliability of Comprehensive System scoring for adolescents in a clinical setting. It also has the advantage over some previous studies of restricting participation to patients from a single site.

METHOD

Participants

Approximately two thirds of adolescents admitted to Four Winds Hospital, a private psychiatric facility in the New York metropolitan area, undergo intensive psychological evaluation for purposes of differential diagnosis, risk assessment, treatment planning, and/or discharge planning. The Rorschach is a standard part of the battery. During a period of slightly less than 30 months from 1996 to 1999, approximately 1,100 adolescents between the ages of 13 and 17 completed the Rorschach.[2] Of these, approximately 10% were not considered for inclusion in the present study, either because the adolescent generated less than 14 responses or rejected at least one card or because the inquiry was judged unacceptable (discussed further below). This left a pool of 998 Rorschachs. From these cases, 84 were chosen at random as the basis for an interrater reliability study. Demographic data for the participants are summarized in Table 1. More than half of adolescent hospitalizations at the facility are funded by Medicaid, with the remainder reimbursed through private insurance.

TABLE 1
Demographic Characteristics of the Sample

	M	SD	n	%
Gender				
Males			40	47.6
Females			44	52.4
Ethnicity				
White			54	64.3
Black			15	17.9
Hispanic			11	13.1
Other			4	4.8
Admission diagnoses				
Psychosis			10	10.0
Conduct disorder			26	31.0
Depression			49	58.3
Discharge diagnoses				
Psychosis			6	7.1
Conduct disorder			16	19.0
Depression			40	47.6
Learning disabled			12	14.3
Age	14.7	1.3	84	
Grade in school	9.1	1.4	80	
Admission axis *V*	32.3	7.0	80	

NOTE: Values for diagnosis are not mutually exclusive.

Table 2 is modeled on the tables of descriptive statistics for the Comprehensive System scores that Exner (2001) provided for various samples. The mean number of responses is somewhat lower than those reported by Exner in his samples, and the mean lambda is substantially higher. These findings are consistent with greater resistance to testing than in Exner's samples or a more simplistic approach to synthesizing information. Either or both hypotheses would be consistent with expectations for inpatient adolescents, especially given the high rate of economically disadvantaged youths in this population. Unfortunately, there are no preexisting analyses of inpatient adolescents to compare with these results, so future re-

TABLE 2
Descriptive Statistics for Rorschach Scores

Variable	M	SD	*Minimum*	*Maximum*	*Mdn*	*SK*	*KU*
R	18.90	5.29	14.00	39.00	17.00	1.55	2.27
W	8.33	3.70	0.00	17.50	9.00	0.15	–0.02
D	7.15	4.51	0.00	22.00	6.75	1.49	2.61
Dd	3.27	2.89	0.00	12.50	2.25	1.36	1.48
S	1.89	1.39	0.00	5.50	2.00	0.58	–0.16
DQ+	4.27	2.74	0.00	14.50	4.00	0.97	1.44
DQo	12.98	5.14	3.50	27.00	12.00	0.77	0.21
DQv	1.30	1.55	0.00	7.00	1.00	1.63	2.53
DQv/+	0.20	0.48	0.00	2.50	0.00	2.79	8.04
FQx+	0.00	0.00	0.00	0.00	0.00	—	—
FQxo	7.21	2.61	1.00	13.50	6.75	0.34	–0.05
FQxu	6.07	2.87	1.00	14.50	5.50	0.72	0.24
FQx–	5.01	2.50	0.50	15.50	4.50	1.09	2.60
FQxNone	0.46	0.91	0.00	5.50	0.00	2.87	10.98
MQ+	0.00	0.00	0.00	0.00	0.00	—	—
MQo	1.14	1.14	0.00	5.00	1.00	1.30	1.55
MQu	0.74	0.81	0.00	4.00	0.50	1.24	2.13
MQ–	0.68	0.83	0.00	3.00	0.25	1.04	0.15
MQNone	0.01	0.11	0.00	1.00	0.00	9.17	84.00
SQual–	0.90	0.97	0.00	4.50	1.00	1.46	2.75
M	2.57	2.03	0.00	11.00	2.50	1.35	3.12
FM	1.92	1.66	0.00	8.00	1.50	1.41	2.64
m	0.86	1.05	0.00	5.00	0.75	1.53	2.47
FM + m	2.77	1.92	0.00	8.50	2.50	0.69	0.34
FC	0.77	0.90	0.00	4.00	0.50	1.25	1.49
CF	1.17	1.17	0.00	4.00	1.00	1.02	0.13
C	0.54	0.88	0.00	4.00	0.00	2.15	4.92
Cn	0.01	0.05	0.00	0.50	0.00	9.17	84.00
Sum Color	2.49	1.75	0.00	7.00	2.00	0.55	–0.58
WSumC	2.37	1.87	0.00	7.50	2.00	0.87	0.10
Sum C’	1.54	1.56	0.00	6.50	1.00	1.31	1.44
Sum T	0.17	0.51	0.00	3.00	0.00	3.62	14.50
Sum V	0.35	0.69	0.00	4.00	0.00	2.75	9.66
Sum Y	1.04	1.26	0.00	7.00	1.00	1.88	5.59
Sum Shd	3.10	2.63	0.00	11.00	2.50	1.03	0.79
Fr + rF	0.18	0.48	0.00	2.00	0.00	2.64	6.19
FD	0.76	0.80	0.00	3.00	1.00	0.68	–0.49
F	9.69	4.76	1.00	23.00	9.50	0.64	0.25
(2)	6.53	3.82	0.50	21.00	6.00	1.34	2.25
3r + (2)/R	0.37	0.15	0.04	0.71	0.37	–0.15	–0.71
Lambda	1.60	1.93	0.07	14.00	1.11	3.98	21.23
EA	4.94	2.88	0.00	13.50	4.50	0.57	0.04
es	5.87	3.36	1.00	15.00	5.00	0.66	–0.12
D Score	–0.26	0.96	–3.00	2.00	0.00	–0.91	1.89
AdjD	–0.04	0.79	–3.00	2.00	0.00	–0.77	3.30
Active	3.19	2.48	0.00	15.00	3.00	1.59	5.03
Passive	2.15	1.57	0.00	8.50	2.00	0.97	2.05
Ma	1.41	1.68	0.00	11.00	1.00	2.72	12.08
Mp	1.16	1.15	0.00	6.50	1.00	1.55	4.45
Intellect	0.53	0.82	0.00	4.00	0.00	1.89	3.86
Zf	10.23	3.84	2.00	23.50	10.00	0.65	1.53
Zd	–1.40	3.91	–12.00	9.50	–1.00	–0.11	0.59
Blends	2.32	1.82	0.00	8.00	2.00	0.77	0.11
Blends/R	0.13	0.10	0.00	0.37	0.11	0.80	–0.27
C–Shd Blnds	0.46	0.69	0.00	3.00	0.00	1.47	1.60
Afr	0.48	0.19	0.18	1.22	0.45	1.29	2.57
Populars	4.05	1.49	1.00	9.00	4.00	0.37	0.57
XA%	0.71	0.11	0.41	0.96	0.72	–0.26	–0.07
WDA%	0.81	0.13	0.40	1.00	0.84	–1.00	0.80
X+%	0.39	0.12	0.05	0.67	0.39	0.05	0.27
X–%	0.27	0.10	0.04	0.50	0.27	0.20	–0.43
Xu%	0.32	0.11	0.07	0.58	0.33	–0.03	–0.33
Isolate/R	0.17	0.15	0.00	0.65	0.14	1.11	0.81
H	1.98	1.67	0.00	9.00	2.00	1.54	3.76
(H)	1.05	1.13	0.00	5.00	1.00	1.29	1.89
Hd	1.13	1.24	0.00	5.00	1.00	1.07	0.48
(Hd)	0.57	0.78	0.00	3.50	0.00	1.45	2.09
Hx	0.02	0.13	0.00	1.00	0.00	6.11	39.65
All H Cont	4.75	2.74	0.00	13.50	4.00	0.83	0.93
A	8.21	3.39	2.00	18.00	7.75	0.71	0.40
(A)	0.74	0.88	0.00	3.50	0.50	1.14	0.58
Ad	1.77	1.63	0.00	9.00	1.50	1.58	4.23
(Ad)	0.11	0.31	0.00	1.50	0.00	2.79	6.92
An	0.69	0.95	0.00	4.00	0.00	1.49	1.85
Art	0.32	0.65	0.00	4.00	0.00	3.02	12.38
Ay	0.15	0.35	0.00	2.00	0.00	3.05	10.83
Bl	0.25	0.49	0.00	2.00	0.00	1.79	2.47
Bt	1.04	1.10	0.00	5.00	1.00	1.23	1.54
Cg	0.94	1.19	0.00	4.50	0.50	1.27	0.87
Cl	0.11	0.41	0.00	3.00	0.00	5.04	30.34
Ex	0.12	0.42	0.00	2.00	0.00	3.79	14.06
Fi	0.37	0.66	0.00	3.00	0.00	1.95	3.59
Food	0.20	0.38	0.00	1.50	0.00	1.89	2.64
Ge	0.11	0.35	0.00	2.00	0.00	3.32	11.57
Hh	0.46	0.77	0.00	3.00	0.00	2.00	3.85
Ls	0.42	0.63	0.00	3.00	0.00	1.92	3.91
Na	0.69	1.02	0.00	5.50	0.50	2.25	6.40
Sx	0.03	0.16	0.00	1.00	0.00	5.61	31.22
Xy	0.02	0.22	0.00	2.00	0.00	9.17	84.00
Idio	1.04	1.02	0.00	4.00	1.00	0.69	–0.32
DV	0.57	0.68	0.00	3.00	0.50	1.43	1.88
INCOM	0.65	0.82	0.00	3.50	0.50	1.36	1.44
DR	0.25	0.48	0.00	1.50	0.00	1.66	1.24
FABCOM	0.40	0.69	0.00	3.50	0.00	2.26	6.09
DV2	0.31	0.38	0.00	1.50	0.25	1.37	1.44
INC2	0.26	0.57	0.00	3.50	0.00	3.19	12.91
DR2	0.14	0.28	0.00	1.25	0.00	2.23	4.86
FAB2	0.27	0.55	0.00	3.00	0.00	2.75	9.14
ALOG	0.23	0.56	0.00	4.00	0.00	4.34	25.36
CONTAM	0.02	0.13	0.00	1.00	0.00	6.11	39.65
RSum6	2.77	2.47	0.00	15.00	2.50	1.99	6.62
Level 2	0.66	1.10	0.00	7.00	0.00	3.16	13.75
WSum6	8.96	10.38	0.00	65.00	6.25	2.70	10.48
AB	0.03	0.16	0.00	1.00	0.00	5.61	31.22
AG	0.27	0.55	0.00	3.00	0.00	2.73	9.04
COP	0.64	0.86	0.00	5.00	0.25	2.03	6.80
CP	0.00	0.00	0.00	0.00	0.00	—	—
GHR	2.71	1.92	0.00	9.50	2.25	0.95	1.17
PHR	2.30	1.72	0.00	8.00	2.00	0.75	0.35
MOR	0.86	1.05	0.00	5.50	0.50	1.69	4.10
PER	0.40	0.95	0.00	5.00	0.00	3.26	12.17
PSV	0.29	0.47	0.00	1.50	0.00	1.24	–0.03

(continued)

TABLE 2 (continued)

NOTE: SK = skew; KU = kurtosis. For further description of the variables, see Exner (1993).

search will be needed to evaluate whether the present findings are unusual. It is worth noting that the mean lambda in this sample is actually lower than that reported by Hamel, Shaffer, and Erdberg (2000) for non-patient children, although so is the mean number of responses.

Procedure

Data gathering. During the period of the initial data collection for the archival study, four licensed psychologists and 39 students were involved in the administration and scoring of Rorschachs at the hospital. The psychologists all received training in the Comprehensive System in graduate school or through Rorschach workshops; two taught the Comprehensive System at the graduate level. The students included doctoral interns and doctoral-level psychology students. Cases were assigned to testers on a rotating basis. Given the high volume of testing at the site and the nature of the testing files, it was impossible to determine the administrator in individual cases.

All students received a 2-day introduction to the Comprehensive System prior to their first testing, including instruction in the administration of the Rorschach. They also observed at least two administrations by a more experienced administrator. Although students were responsible for administration of the Rorschach, response coding was always reviewed with one of the four licensed psychologists. Initially, this review involved the reading aloud of each response and its inquiry and coding of the response by the psychologist. If the psychologist deemed the inquiry unacceptable, it was considered invalid and not scored. Such cases were rare, although it would be very difficult to determine the precise frequency of this outcome. After achieving a certain level of expertise (frequently requiring 6 months or more), students were also expected to code responses prior to the supervision session. However, the psychologist always remained responsible for the final coding. The Rorschach Interpretive Assistance Program, Version 3 (RIAP-3; Exner, Cohen, & McGuire, 1990), was then used to generate scores. The handwritten transcript of the administration became part of the adolescent's testing chart.

For purposes of evaluating the reliability of field-based coding and scoring, a second coding of each protocol was completed in 2001. The transcript was provided to one of two doctoral students in clinical psychology who had completed at least 2 years of Rorschach administration and coding under the system described above, but no additional training was provided in preparation for this study. One student recoded 37 of the files; the other recoded 47. Protocols were randomly assigned to raters depending on their relative availability. Consistent with the practical obstacles to true field studies raised in the introductory paragraphs of this article, these second judges were aware they were recoding protocols that were several years old for research purposes. The second coding was not supervised in any way and was completed without knowledge of the first coding. RIAP-3 was again used to generate the scores. Newer scores such as GHR were generated from initial response codes with an SPSS script. A program was also developed using Visual Basic Version 6.0, which allowed extraction of response code variables from RIAP-3 files.

Analytic decisions. The decision was made to focus on the individual codes rather than the segments at the response level, because response segment reliability is not particularly relevant to the interpretive process. At the protocol level, it is true that clinical judgments are based on categorical placement. However, given the goals of the study, it was thought important to generate results that allowed comparison to previous findings. Furthermore, there is little research concerning the appropriate cut scores for adolescent Rorschachs. Accordingly, the protocol-level analyses focused on dimensional scores, although the reliability of categorical placement will also be addressed briefly. Analyses at the response level were based on 1,588 responses, whereas protocol-level analyses were based on the corresponding 84 structural summaries.

Acklin et al. (2000) noted that at the response level, one can compute a kappa coefficient for each code (e.g., W present or absent) or for each coding decision (e.g., W vs. D vs. Dd). The latter seems more desirable, because a single coding decision can determine the outcome for multiple codes. For example, the decision whether W is present or absent is ipsatively related to deciding whether D or Dd is present or absent. As a result, reliability statistics based on individual codes demonstrate dependencies that complicate the computation of accurate descriptive statistics, such as the mean. It should be noted that in many cases, coding decisions are equivalent to decisions about individual codes, as is the case when deciding whether white space was used.

Some coding files contained *Z* values, whereas others contained *Z* codes such as *ZW*. The *Z* code cannot always be recovered from the *Z* value, because some codes have the same value, but the value can always be determined from the code. Furthermore, a decision between two *Z* codes with the same value has no interpretive significance in the Comprehensive System (Exner, 2000). For these reasons, the evaluation of *Z*-score reliability was based on values rather than codes.

At the protocol level, the issue of dependencies also deserved consideration. Many of the summary scores reported in Table 2 are by definition correlated because of overlapping content. On the other hand, reducing the num-

TABLE 3
Kappa Coefficients for Coding Variables

Code	*Kappa*	*Base Rate*	*Code*	*Kappa*	*Base Rate*	*Code*	*Kappa*	*Base Rate*
Loc (W, D, Dd)	.94		(Hd)	.81	0.03	Sx	.80	<0.01
S	.90	0.10	Hx	.00	<0.01	Xy	1.00	<0.01
DQ (+, o, v, v/+)	.88		A	.94	0.44	Id	.67	0.06
M	.91	0.14	(A)	.77	0.04	Pop	.93	0.22
FM	.90	0.10	Ad	.85	0.09	Z Score	.90	0.45
m	.90	0.05	(Ad)	.35	0.01	DV (Lv1, Lv2)	.56	0.03
ap (a, p, a-p)	.91	0.28	An	.93	0.04	INCOM (Lv1, Lv2)	.62	0.05
Map (a, p, a-p)	.88	0.14	Art	.83	0.02	DR (Lv1, Lv2)	.57	0.02
C (C, CF, FC, Cn)[a]	.86	0.13	Ay	.48	0.01	FABCOM (Lv1, Lv2)	.75	0.04
C' (C', C'F, FC')	.79	0.08	Bl	1.00	0.01	ALOG	.49	0.01
T (T, TF, FT)	.73	0.01	Bt	.91	0.06	CONTAM	.67	<.01
V (V, VF, FV)	.66	0.02	Cg	.86	0.05	AB	.33	<.01
Y (Y, YF, FY)	.69	0.06	Cl	1.00	0.01	AG	.76	0.01
r (Fr, rF)	.97	0.01	Ex	1.00	0.01	COP	.76	0.03
FD	.79	0.04	Food	.64	0.01	CP	—[b]	0.00
F	.95	0.48	Fi	.98	0.02	GHR	.88	0.14
FQ (+, o, u, –)	.80	0.02	Ge	.95	0.01	PHR	.85	0.12
Pair	.91	0.35	Hh	.82	0.03	MOR	.89	0.05
H	.95	0.11	Ls	.52	0.02	PER	.73	0.02
(H)	.90	0.06	Na	.67	0.04	PSV	.84	0.01
Hd	.82	0.06	Sc	.78	0.03			

NOTE: $N = 1{,}588$ responses coded twice. All symbols are taken from Exner (1993). All coding decisions also had an absent option, except those for which the base rate is missing.
a. Because of the interpretive importance of color, reliability was also computed separately for the presence-absence of FC (.79), CF (.78), and C (.79).
b. Could not be computed; never coded.

ber of scores to minimize redundancy would have meant eliminating some scores that are interpretively important.

As a compromise, Exner's (2000) interpretive manual was reviewed to identify scores that clearly affect interpretation. For example, FM and m affect interpretation by their contribution to es, D, Adj es, and Adj D, and the left side of eb. All of these are considered important elements of the interpretive process. In contrast, the consideration of FM and m as separate scores is thought to add additional richness to the interpretation, but their role is secondary, and so to avoid further redundancy in scores, these were not examined. F% was added to the standard set of Comprehensive System scores because of recent evidence supporting its statistical superiority to Lambda (Meyer, Viglione, & Exner, 2001), and because it allowed an evaluation of the impact the absence of a ceiling has on the reliability of Lambda.

RESULTS

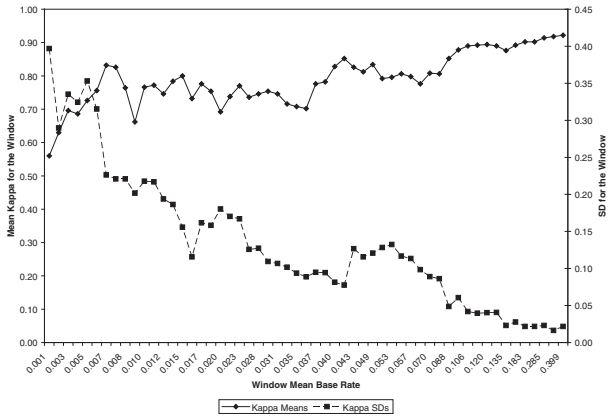
Analyses of Response Codes

Most values for kappa were generated using SYSTAT Version 10.2. In five cases where the raters used different sets of categories (e.g., the original coders never used a code that was used by the later coders), the software could not generate a value for kappa. These were computed using an Excel spreadsheet instead. As can be seen from Table 3, 3 out of 61 kappa coefficients (5%) were < .40, which Fleiss (1981) considered poor reliability. Another 5 (8%) fell in the range of .40 to .59, which Fleiss considered fair. Nine (15%) fell in the range of .60 to .74, considered good, whereas 44 (72%) were > .74. Fleiss considered this evidence of excellent reliability. In fact, 35 (57%) exceeded .80, which Landis and Koch (1977) classified as nearly perfect reliability and Shrout (1998) called substantial. The mean kappa was .79 (median = .84).

The table also provides the base rate for each code. To generate this value, the code was dichotomized as present-absent. The Base Rate column indicates the proportion of responses in which the code was present or absent, whichever was smaller. In all but three cases, the smaller value was the proportion of cases in which the code was present. The smaller this base rate value, the more skewed the distribution of the code. Most of the distributions are quite skewed, with several codes occurring in 1% of responses or less.

Figure 2 uses these base rates to provide insight into the relationship between skew and kappa. Kappa values were sorted according to the base rates provided in Table 3 from lowest to highest. Windows of five consecutive kappa values were created such that each window shared four kappa values with each of its immediate neighbors. Within each

FIGURE 2
Relationship Between Distribution Skew (Base Rate) and the Stability and Value of Kappa



NOTE: As base rate increases toward .50, the mean value increases from .56 in the first window to .92 in the last. The standard deviation decreases, from .40 in the first window to .02 in the last.

window, the mean and standard deviation of the kappa values were computed, and the figure presents these statistics as a function of increasing base rate. Values on the *X* axis reflect the mean base rate for the window. The boxes represent the standard deviations for each window, and the diamonds reflect the mean kappa values.

As the base rate increased, kappa values became more stable, as indicated by the declining standard deviation. This occurred because for extremely infrequent codes, a shift in the coding of a single response could dramatically alter the outcome. This relationship between base rate and stability led Meyer et al. (2002) to suggest that scores should not be considered statistically stable unless the base rate exceeded .01. In the present case, stability continued to improve until the base rate reached approximately .10. The standard deviation continued to decline even beyond the base rate of .10, although more gradually.

The means increased as a function of base rate, a finding that is consistent with expectation and with previous research (e.g., Acklin et al., 2000). Interestingly, the means did not consistently surpass .80 until about the same base rate (.10) at which relative stability was achieved. However, the increase in means is not as dramatic as the decrease in standard deviations, suggesting that even at base rates where kappa is relatively unreliable, the average field reliability for codes was reasonable.

Analyses of Protocol Scores

Shrout and Fleiss's (1979) ICC(1, 1) formula was used to evaluate the reliability of Comprehensive System scores. Although this statistic is most consistent with a model in which no rater is involved in coding more than one protocol, it is generally considered the most appropriate of the available intraclass correlations for circumstances where different targets are rated by different judges (Meyer et al., 2002; P. Shrout, personal communication, October 10, 2004), a circumstance that is almost inevitable in field reliability research. Of 71 interpretively distinct scores from the structural summary, two were never assigned (CP and OBS positive), and one score occurred in only one protocol (MQNone). In the latter case, both raters agreed on its presence, resulting in an artificially elevated value for the ICC.

All ICC (1, 1) values were generated using SPSS. As demonstrated in Table 4, the reliability of summary scores was better than that found for codes. None of the 69 ICCs computed fell in Fleiss's (1981) range indicating poor reliability, and only 1 (1%) was fair. Nine (13%) fell in the good range, 59 (86%) were excellent, and 49 (71%) met Landis and Koch's (1977) criterion for nearly perfect reliability. The mean ICC was .86 (median = .89) with or without MQNone considered in the computation. Consistent with evidence provided by Meyer et al. (2002), even the ICCs for variables based on codes with lower values for kappa, such as WSum6, were adequate. If, as Wood et al. (1996a, 1996b) correctly concluded, the reliability of summary scores is more important than the reliability of codes, then the results indicate acceptable and in many cases excellent field reliability for the Comprehensive System.

The reliability estimates for Lambda and F% were similar. Contrary to our preliminary expectation, the reliability estimate for Lambda actually exceeded that for F% by a small margin. The standard deviation of Lambda (1.93 averaged across the two raters) was 10 times that for F% (.19). Because the two raters achieved a very high degree of consistency in their coding of pure *F* responses (kappa = .95), unreliable variability represented a higher proportion of the total variability for F% than was true for Lambda. However, it is still reasonable to hypothesize that the reliability estimate for F% will demonstrate greater consistency across samples than is true of Lambda.

Finally, one analysis was generated to demonstrate the degree to which score-based categories could affect interrater reliability. Adolescents were dichotomized three times, into those with and without at least one Morbid response, those with or without at least two Morbid responses, and those with or without at least three Morbids. As indicated in Table 5, the more extreme the cut rule and the smaller the resulting base rate of those with the requisite number of Morbid responses, the more kappa declined. Based on this finding, it seems likely that dichotomization based on summary scores would tend to

TABLE 4
Intraclass Correlation Coefficient (ICC) Values for Score-Level Variables

Variable	*ICC*	*Variable*	*ICC*	*Variable*	*ICC*	*Variable*	*ICC*
W	.96	(2)	.95	Intellect	.69	Zd	.90
D	.91	H	.96	MOR	.91	PSV	.72
Dd	.91	(H) + Hd + (Hd)	.92	RSum6	.78	COP	.77
S	.83	Animal Cont	.94	Level 2	.63	AG	.66
DQo + DQ+	.99	C-Shd Blends	.89	WSum6	.78	GHR	.92
DQv + DQv+	.89	Lambda	.99	MQ–	.83	PHR	.91
M	.93	F%	.96	MQNone	1.00	Food	.58
FM + m	.89	WSumC	.94	Afr	1.00	PER	.89
FC	.79	EA	.94	Blends	.84	Isolate/R	.90
CF + C	.91	EBPer	.90	CP	—[a]	3r + (2)/R	.93
C	.79	es	.91	XA%	.75	An + Xy	.95
Sum C’	.89	Adjes	.92	WDA%	.91	PTI	.75
Sum T	.85	D	.78	X-%	.74	DEPI	.77
Sum V	.84	AdjD	.81	S-	.82	CDI	.87
Sum Y	.76	Active	.94	Populars	.82	SCON	.88
Fr + rF	.98	Passive	.73	X+%	.88	HVI	.85
FD	.73	Ma	.91	Xu%	.61	OBS	—[b]
F	.98	Mp	.74	Zf	.95		

NOTE: *N* = 100 protocols scored twice. All symbols are taken from Exner (2001). ICC = intraclass correlation coefficient.
a. No variability.
b. Dichotomized; no variability.

TABLE 5
An Example of Category Reliability

		Category	
	ICC	*Base Rate*	*Kappa*
MOR (number)	.91		
MOR > 0		.52	.86
MOR > 1		.23	.83
MOR > 2		.05	.74

NOTE: *N* = 100 protocols scored twice. MOR = morbid response.

be associated with lower levels of reliability if the goal is to predict relatively low base rate outcomes.

This is a worthwhile topic for future study. In clinical practice, test data are often used for the purpose of drawing conclusions about binary clinical judgments. Such judgments might include a decision about the presence or absence of a diagnosis or a clinical state, or a treatment decision. If the reliability of dichotomizations based on the test scores is generally lower than the reliability of the dimensional score, then reliability coefficients generated on the basis of the dimensional score may overestimate the reliability of tests as they are actually used. It is important to note this issue potentially applies to all clinical measures used as the basis for clinical judgments, not just the Rorschach.

DISCUSSION

It is increasingly recognized that reliability is not a characteristic of a scale but instead varies as a function of the context of the measurement (Streiner, 2003). Similarly, interrater reliability is not inherent to a scoring system but should vary across populations of raters. Even within the population of field raters, there are likely to be important moderators of reliability, particularly the degree of diligence demonstrated in ensuring accurate scoring. Rather than talking about the field reliability of an instrument, it is probably better to explicate the conditions under which an adequate level of field reliability can be achieved.

The present results can be compared to the two prior studies that examined both response-level and protocol-level reliability. The mean kappa value reported here (.79) is somewhat lower than that reported by Meyer et al. (2002; .89), but very similar to that reported by Acklin et al. (2000; .78). Meyer et al’s (2002) omission of kappa values when the base rate was low could have contributed to this difference. Even so, in most cases the results were generally supportive of the conclusion that the field reliability is acceptable for most codes, although not necessarily optimal. The mean ICC from the present sample (.86) was exactly midway between those reported in these two prior studies (.78 for Acklin et al., 2000; .92 for Meyer et al., 2002).

Level of oversight may also be an important situational moderator of reliability. It was suggested earlier that the reliability of coding for research purposes might be enhanced by the raters' awareness that the accuracy of the coding will be evaluated. Similarly, in the setting where the present study was completed, the students were aware that one of the psychologists would review and correct any errors they found in the coding. One might question whether the present results would generalize to settings where there is less supervision. On the other hand, coding by individuals who are neither experienced professionals nor closely supervised is an ethically questionable assessment practice.

Finally, all codes were converted to scores using software. The reliability of scores might not have been so high if the conversion were accomplished by hand. Given the complexity of the Comprehensive System, errors are almost inevitable if summary scores are computed without the aid of a computer. The use of software for this purpose should be encouraged.

The present results, taken in combination with previous findings, suggest that Rorschach coding can in most cases meet desirable standards for interrater reliability (statistics in the range of .80 to 1.00). At the same time, studies consistently identify a subset of codes that fail to meet even the minimum acceptable level of reliability (.60), although this phenomenon seems to have more to do with base rate issues than with an inherent limitation in the coding criteria.

As Wood et al. (1997) noted, score reliability is a more important issue than code reliability, and the present findings concerning score reliability are more consistently positive. Only one score (number of food responses) did not meet the minimum standard for reliability, and most fell in the range of .80 to 1.00. Scoring of the Comprehensive System purely for clinical purposes can be completed in a manner that achieves desirable levels of reliability so long as there is appropriate supervision or experience with the measure. Perhaps the findings of this and other recent studies (e.g., Acklin et al., 2000; Meyer et al., 2002) can allow assessors to move on from issues of the potential for reliable scoring of the Rorschach in clinical settings to more important issues of ensuring the competent use of all tests in clinical settings.

NOTES

1. Several other studies evaluating the validity and/or actuarial use of Comprehensive System scores have examined the reliability of data originally collected for clinical purposes and therefore have reported field reliability statistics. In fact, some were conducted in the same setting as the current study (e.g., Stokes, Pogge, & Grosso, 2001; Stokes, Pogge, & Powell-Lunder, 2003). However, because interrater reliability was a secondary concern, these studies only reported statistics for those variables used in the main analyses. Meyer et al.'s (2002) analyses represent the only full-scale study of field reliability to be published.

2. Because this is a clinical setting with a high volume of clinical assessments, exact numbers are unavailable.

REFERENCES

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

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

Exner, J. E., Jr. (2000). *A primer for Rorschach interpretation*. Asheville, NC: Rorschach Workshops.

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

Exner, J. E., Jr., Cohen, J., & McGuire, H. (1990). *RIAP–Rorschach Interpretation Assistance Program, Version 3*. Asheville NC: Rorschach Workshops.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. New York: Wiley.

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

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

Hartmann, E., Wang, C. E., & Berg, M. (2003). Depression and vulnerability as assessed by the Rorschach method. *Journal of Personality Assessment*, *81*, 242-255.

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

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

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

McDowell, C. J., & Acklin, M. W. (1996). Standardizing procedures for calculating Rorschach interrater reliability: Conceptual and empirical foundations. *Journal of Personality Assessment*, *66*, 308-320.

McGrath, R. E. (2003). Achieving accuracy in testing procedures: The Comprehensive System as a case example. *Journal of Personality Assessment*, *81*, 104-110.

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

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

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

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

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, *7*, 301-317.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420-428.

Shrout, P. E., Spitzer, R. L., & Fleiss, J. L. (1987). Quantification of agreement in psychiatric diagnosis revisited. *Archives of General Psychiatry*, *44*, 172-177.

Slate, J. R., Jones, C. H., Coulter, C., & Covert, T. L. (1992). Practitioners' administration and scoring of the WISC-R: Evidence that we do err. *Journal of School Psychology*, *30*, 77-82.

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

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

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

Viglione, D. J. (2002). *Rorschach coding solutions: A reference guide for the Comprehensive System*. San Diego, CA: Author.

Viglione, D. J., & Taylor, N. (2003). Empirical support for interrater reliability of Rorschach Comprehensive System coding. *Journal of Clinical Psychology*, *59*, 111-121.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children manual* (4th ed.). San Antonio: Psychological Corporation.

Whitten, J., Slate, J. R., Jones, C. H., & Shine, A. E. (1994). Examiner errors in administering and scoring the WPPSI-R. *Journal of Psychoeducational Assessment*, *12*, 49-54.

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

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

Wood, J. M., Nezworski, M. T., & Stejskal, W. J. (1996b). Thinking critically about the Comprehensive System for the Rorschach: A reply to Exner. *Psychological Science*, *7*, 14-17.

Wood, J. M., Nezworski, M. T., & Stejskal, W. J. (1997). The reliability of the Comprehensive System: A comment on Meyer (1997). *Psychological Assessment*, *9*, 490-494.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin*, *103*, 374-378.

Robert E. McGrath, Ph.D., is professor of psychology at Fairleigh Dickinson University. His primary research interests are in the areas of assessment and professional issues in psychology.

David L. Pogge, Ph.D., is director of psychology at Four Winds Hospital in Katonah, New York; senior clinical lecturer at Fairleigh Dickinson University; and a visiting assistant professor of psychology in psychiatry at the Albert Einstein College of Medicine in the Bronx, New York.

John M. Stokes, Ph.D., is a professor of psychology in the Psy.D. Program in School-Clinical Child Psychology at Pace University.

Ana Cragnolino, Ph.D., is presently in private practice in the Washington, D.C., area. Her clinical and research interests are in assessment.

Michele Zaccario, Ph.D., is a pediatric psychologist on the rehabilitation and neonatal intensive care units at New York University Medical Center and an adjunct professor at Pace University. Current research includes outcome studies involving premature infants and children with traumatic brain injuries.

Judy Hayman, Ph.D., is presently working at Psychological HealthCare in Syracuse, New York, providing neuropsychological evaluations and treatment for adults with traumatic brain injury.

Teresa Piacentini, Ph.D., currently works in the Department of Child and Adolescent Psychiatry at Columbia University.

Douglas Wayland-Smith, Ph.D., works as a staff psychologist at Hall-Brooke Hospital and maintains a private psychotherapy and testing practice in Danbury, Connecticut.