JOURNAL OF PERSONALITY ASSESSMENT, *74*(1), 15–47

ARTICLES

Interobserver Agreement, Intraobserver Reliability, and the Rorschach Comprehensive System

Marvin W. Acklin, Claude J. McDowell II,
Mark S. Verschell, and Darryl Chan
Honolulu, Hawaii

Interrater agreement and reliability for the Rorschach have recently come under increasing scrutiny. This is the second report examining methods of Comprehensive System reliability using principles derived from observational methodology and applied behavioral analysis. This study examined a previous nonpatient sample of 20 protocols (*N* = 412 responses) and also examined a new clinical sample of 20 protocols (*N* = 374 responses) diagnosed with Research Diagnostic Criteria. Reliability was analyzed at multiple levels of Comprehensive System data, including response-level individual codes and coding decisions and ratios, percentages, and derivations from the Structural Summary. With a number of exceptions, most Comprehensive System codes, coding decisions, and summary scores yield acceptable, and in many instances excellent, levels of reliability. Limitations arising from the nature of Rorschach data and Comprehensive System coding criteria are discussed.

In this article, we report on the latest in a series of studies undertaken to examine and improve the performance of the Rorschach Inkblot Test as a clinical and research instrument (Acklin, McDowell, & Ornduff, 1992; McDowell & Acklin, 1996). The Rorschach relies on the application of a complex coding system to samples of verbal behavior to yield data for clinical and research purposes. As such, the Rorschach is less a test than a behavioral observation methodology (McDowell & Acklin, 1996). The Comprehensive System was developed by Exner (1990, 1991, 1993, 1995; Exner & Weiner, 1994) to standardize Rorschach observation (i.e., administration) and coding procedures, and it remains the focus of ongoing investigation and refinement.

The merits of assessing the reliability of diagnostic tests and classification systems are prominent within psychometric theory (Jensen, 1959) and likely date back to the earliest history of psychiatric diagnosis (Matarazzo, 1983). However, the earliest empirical attempts to measure the reliability of behavioral observational systems were not conducted until the early to mid-20th century (i.e., 1930–1965). Most, if not all, of these early studies were flawed in their statistical designs, as they employed reliability measures (e.g., percentage agreement and contingency coefficients) that did not take into account the effects of behavior prevalence and chance agreement or give credit for associated observations that were not in strict agreement (Shrout, Spitzer, & Fleiss, 1987). Cohen's kappa (1960) was introduced to overcome the weaknesses of these measures, and it has since become the standard method for assessing diagnostic agreement in psychiatry and other medical specialties. A standard approach for assessing the reliability of the Rorschach Comprehensive System, however, has yet to be established.

Historically, a variety of statistical techniques have been used in Rorschach reliability research, including phi (ϕ), the product–moment correlation (r), and proportion of agreement (PA). Despite opposing psychometric literature, Weiner's (1991) editorial guidelines established the percentage agreement index as the standard reliability measure for research studies submitted to the *Journal for Personality Assessment.* More recent applied research papers have used Cohen's kappa (κ) and the intraclass correlation coefficient (ICC) for selected Comprehensive System variables (Greco & Cornell, 1992; Netter & Viglione, 1994; Perry & Braff, 1994; Perry, McDougall, & Viglione, 1995; Perry, Sprock, et al., 1995; Perry & Viglione, 1991). Relatively few studies, however, have outlined the conceptual and methodological foundations for the application of these various reliability techniques to the Rorschach Comprehensive System.

In the absence of a standardized methodology for assessing the reliability of Comprehensive System data, technical scrutiny and criticism of Exner's methods have emerged (e.g., Wood, Nezworski, & Stejskal, 1996a). Wood and his colleagues pointed out the conceptual and procedural vagaries inherent in published approaches to Comprehensive System reliability. In particular, they noted the use of inadequate statistical methods, as well as an absence of data documenting the reliability of Comprehensive System summary scores. Working separately, McDowell and Acklin (1996) similarly concluded that Comprehensive System reliability had not been established using appropriate statistical methods. Recent debates published in *Psychological Science* (Exner, 1996; Wood, Nezworski, & Stejskal, 1996a, 1996b) and *Psychological Assessment* (Meyer, 1997a, 1997b; Wood, Nezworski, & Stejskal, 1997) regrettably did not resolve the issues in dispute.

Conceptual inconsistencies and statistical confusion (Suen, 1988) have complicated dialogue concerning the appropriate terms and procedures for assessing the reliability of the Rorschach Comprehensive System (e.g., witness the recent interchange between Exner, 1996, and Wood et al., 1996a, 1996b). In the context of behavioral assessment, Cone (1982) noted that

> The current looseness of terms and their inconsistent usage retards the development of the precision necessary for [behavioral assessment] to advance. … One area particularly in need of [distinctions and clarifications] is direct observation methodology. … Distinguishing accuracy from agreement and reliability and validity illustrates some obvious needs of a lexicon of behavioral assessment. (pp. 2–3)

In an attempt to clarify these issues for the Rorschach Comprehensive System, we propose a reliability framework that is informed by modern behavioral observation methodology.

BEHAVIORAL OBSERVATION PARADIGM

In defining a behavioral observation paradigm, it is necessary to distinguish between the nature and interpretation of data (Suen, 1988). The nature of Rorschach data encompasses both nomothetic trait and idiographic behavior approaches to personality assessment (Weiner, 1998). Closely related to the nature of behavioral observational data are interpretations based on either criterion-referenced or norm-referenced measurement frameworks (Suen, 1988). We propose that Comprehensive System data are interpreted within a criterion-referenced rather than a norm-referenced framework. This distinction is by no means insignificant, as criterion-referenced measurement frameworks tend to yield lower estimates of data reliability due to their inclusion of observers as a source of systematic observation error (McGraw & Wong, 1996; Suen, 1988). Although this viewpoint may sound counterintuitive, the psychometric literature (e.g., Suen & Ary, 1989; Tinsley & Weiss, 1975) seems rather clear:

> Thus, when decisions are based on the *mean* [italics added] of the ratings obtained from a set of observers, or on ratings which have been *adjusted* [italics added] for rater differences (such as ranks or *Z* scores), the interjudge variance should not be regarded as error. On the other hand … if the investigator wishes his results to be generalizable to other samples of judges using the same scale with a similar sample of subjects, the between-raters variance should be included as part of the error term. (Tinsley & Weiss, 1975, p. 363)

With regard to the common measures of observer reliability, percentage agreement and kappa treat systematic observer bias as error, contingency coefficients and the product–moment correlation do not, and ICCs may or may not depending on whether correlation is measured using a consistency or absolute agreement definition of correlation (McGraw & Wong, 1996; Suen, 1988).

LEVELS OF RELIABILITY ANALYSIS

The Comprehensive System yields a plethora of behavioral observation data. Observation procedures produce samples of verbal behavior that correspond to the responses given to the Rorschach inkblots. Individual codes are applied to target behaviors (i.e., key words) within each response, and then are tabulated, summed, and combined to form interpretive indexes that describe a composite Rorschach protocol. A conceptual scheme for organizing response-level verbal behaviors (e.g., Location, Developmental Quality, and Determinant segments) facilitates the processes of both behavioral coding and data interpretation.

Critics of the Rorschach Comprehensive System have decried the fact that there is no published research demonstrating the reliability of aggregate codes and interpretive indexes from the Structural Summary (e.g., Sum *T,* 3*r*+(2)/*R, DEPI*). In particular, Wood et al. (1996a, 1996b, 1997) asserted that acceptable levels of reliability for response-level codes are insufficient to justify the reliability of protocol-level composite scores. In contrast, Meyer (1997a) argued that data representing summary scores should be more reliable than data representing response-level codes because psychometric theory predicts that random errors of measurement will tend to cancel on aggregation. Our review of the relevant literature suggests that the answer to this debate depends on the statistical approach that is used to evaluate data reliability. In most cases, correlation coefficients (e.g., the intraclass correlation and the product–moment correlation) and correlational-like coefficients (e.g., ϕ and κ) will produce response-level reliability estimates that provide a lower bound estimate of total score reliability (e.g., Berk, 1979; Hartmann, 1977). Conversely, estimates of response-level reliability derived via percentage agreement statistics will not demonstrate a formal relation to total score reliability (Hartmann, 1977). Because the relation between response-level reliability and protocol-level reliability is never absolute and is complicated by summary scores that are derived from multiple component codes, the analysis of both levels of Comprehensive System data appears to be warranted.

CONCEPTUAL FOUNDATIONS OF RELIABILITY

Within the context of behavioral assessment, reliability indexes provide a means to evaluate the effectiveness of observer training and the objectivity with which target behaviors may be measured (Berk, 1979): “Conventionally, reliability is most commonly defined as *data* (not observer) consistency. The term reliability refers to the degree to which measurement error is absent from the data. The less measurement error, the more consistent the data” (Suen, 1988, p. 348). Relative to observers as sources of error, data consistency has been assessed through two primary approaches: between-observer reliability (i.e., interobserver agreement) and within-observer reliability (i.e., intraobserver reliability; Suen, 1988, p. 349). Interobserver agreement indexes (e.g., proportion agreement, ϕ, and κ) are usually applied to data at the nominal and ordinal levels of measurement and provide an indication of the degree to which observers are interchangeable (Suen, 1988). Intraobserver reliability indexes (e.g., the ICC) may be applied to data that approximate interval-level measurement and provide an indication of the degree to which observational data are free from measurement error (Suen, 1988). Although indexes of intraobserver reliability tend to provide more rigorous analyses of observer consistency than do indexes of interobserver agreement (Berk, 1979),[1] interobserver agreement indexes represent the only means for assessing the reliability of polychotomous nominal data. Hence, with specific regard to the Rorschach Comprehensive System, it would seem appropriate to apply intraobserver reliability indexes to protocol-level summary scores and response-level data that are coded in dichotomous fashion and to apply interobserver agreement indexes to response-level data that are coded in polychotomous fashion.

STATISTICAL APPROACHES

In a series of important training studies, DeCato (1983, 1984, 1994) was one of the first researchers to methodologically examine the reliability of Rorschach coding using the kappa coefficient. More recently, McDowell and Acklin (1996) operationalized principles from applied behavioral analysis and used Cohen's kappa to examine the reliability of response-level Comprehensive System data in a sample of nonpatient protocols. In addition to demonstrating respectable levels of response-level observer reliability, McDowell and Acklin suggested that *kappa,* a PA coefficient corrected for chance agreement, produced more accurate reliability estimates than the PA index alone, a fact of potential importance in the effort to refine the reliability, validity, and utility of the Comprehensive System. Consequently, and consistent with most expert commentary in the field (e.g., Suen & Ary, 1989), McDowell and Acklin suggested that Cohen's kappa is the method of choice for quantifying response-level reliability in the Rorschach Comprehensive System.

This study extended McDowell and Acklin's (1996) previous Comprehensive System reliability analyses by (a) organizing response-level data according to a decision-based coding methodology; (b) examining aggregate codes, ratios, percentages, and derivations from the Structural Summary; (c) examining the relation between response-level reliability, base rate, and quantification method; and (d)

[1] In addition to noting that the Tinsley and Weiss (1975) definition of *reliability* (i.e., proportionality of ratings) is only appropriate for norm-referenced measurement frameworks, we also note that nominal ratings assume interval-level characteristics when coded in dichotomous fashion. Hence, the distinction between *agreement* and *reliability* will hold whether one defines *reliability* as proportionality of ratings or absence of measurement error.

incorporating a new sample of clinical protocols in addition to the original sample of nonpatient protocols. This last facet of this research design helps to establish the discriminant validity of the Comprehensive System across varying diagnostic groups, as it is a well-known fact that the magnitude of reliability estimates varies as a function of the phenomena observed (Haynes, 1978).

Consistent with the recommendations of the previous study, and with the conceptualization of response-level reliability of this study, unweighted kappa (i.e., standard kappa) was applied to all response-level data, with the understanding that this coefficient may be interpreted as an index of intraobserver reliability when data are coded in dichotomous fashion.[2] For response-level data that were coded in polychotomous fashion, kappa was most cautiously interpreted as an index of interobserver agreement.[3]

This study proposes the ICC as the method of choice for assessing the reliability of Comprehensive System summary scores. The intraclass correlation approach offers several methodological advantages over other measures of intraobserver reliability (e.g., ϕ, weighted κ, and the product–moment correlation), including (a) freedom from the classical parallel test assumptions, (b) the ability to provide information on the relative contributions of different sources of measurement error, and (c) the ability to estimate observer reliability within a criterion-referenced measurement framework (Suen, 1988). Specifically, this study uses the ICC that is based on a two-way random effects model of variance and an absolute agreement definition of correlation (McGraw & Wong, 1996; Shrout & Fleiss, 1979). This ICC is appropriate when behavioral observations are performed by raters who may differ in their understanding or application of the relevant behavioral coding system, as it treats systematic observer bias as a source of observation error.

METHOD

Participants

Twenty nonpatient protocols (*N* = 412 responses) were randomly selected from a larger sample of protocols provided by students at a midwestern university during

[2] Unweighted kappa is actually a special case of weighted kappa in which all classification disagreements are treated as equally serious (Soeken & Prescott, 1986). For reasonably large sample sizes and dichotomous data, kappa is equivalent to the ICC that results when a two-way analysis of variance is applied to the data (Fleiss & Cohen, 1973).

[3] To interpret kappa for polychotomous nominal data as a measure of intraobserver reliability, one must be willing to (a) assume the rather untenable classical parallel test assumptions (Bakeman, Quera, McArthur, & Robinson, 1997) or (b) perform an equally indefensible data transformation (i.e., differentially weighting the disagreements among the nominal ratings) that enables kappa to be interpreted as a reliability coefficient (i.e., the intraclass correlation) that is not bound by the classical parallel test assumptions (Fleiss & Cohen, 1973; Suen, 1988).

the years 1987 to 1989. The clinical and demographic characteristics of these participants were described in a previous report (McDowell & Acklin, 1996).

Twenty clinical protocols (*N* = 374 responses) were randomly selected from a larger sample of protocols obtained from psychiatric inpatients at a 56-bed, general hospital over a 4-year period from 1992 through 1995. The average age of the 12 female and 8 male participants was 27.3 years (*SD* = 8.3). The sample was ethnically heterogeneous (30% White, 10% African American, 20% Asian American, 10% Hispanic, and 30% other). All participants received research diagnostic criteria (RDC; Spitzer, Endicott, & Robins, 1989) diagnoses that were based solely on chart review (i.e., independent of clinical information from the Rorschach). RDC diagnoses (and their associated frequencies) were: schizophrenia (2), unspecified functional psychosis (2), bipolar manic (3), intermittent depressive disorder (1), major depression (2), minor depression (3), labile personality (1), drug use disorder (4), antisocial personality (1), and other psychiatric disorder (1). Descriptive statistics for both samples are available on request from Marvin W. Acklin.

Administration and Coding Procedures

Participants from both samples were assessed by graduate clinical psychology students who were trained in Comprehensive System administration and coding procedures (Exner, 1990). Training and supervision were provided by the Marvin W. Acklin. All protocols were valid in terms of Comprehensive System guidelines for response productivity ($R \geq 14$). For the purpose of reliability analysis, the 40 protocols were then independently rescored by Claude J. McDowell II (as per the previous study) and a graduate clinical psychology student with advanced training in the Comprehensive System. Each of these raters possessed a minimum of 3 years of experience in Comprehensive System coding procedures. *A Rorschach Workbook for the Comprehensive System* (Exner, 1990) facilitated scoring decisions.

The Rorschach Interpretive Assistance Program, Version 3.1 (RIAP3; Exner & Ona, 1995), was used to produce structural summaries for all protocols and to export protocol-level data to a statistical database program for reliability analysis. However, limitations of the RIAP3 export facility necessitated that coded responses be manually entered into the statistical database for response-level reliability analysis. Data-checking programs were written by Mark S. Verschell to minimize errors in the data entry process.

Response-Level Coding Scheme

The process of organizing the multitude of Comprehensive System data for reliability analysis is a complex task, and as far as we are aware, it has never been adequately described in the Rorschach literature. In an effort to promote dialogue among Rorschach researchers and, ultimately, the development of a standard approach for evaluating Comprehensive System reliability, we present our data organization scheme in explicit fashion (see also Table 1).

The nine segments of the Comprehensive System may be conceptualized as encompassing 60 specific *coding decisions* that raters must apply to each verbal response. These coding decisions represent a mutually exclusive use of each individual code that comprises the Comprehensive System. That is, no individual code may be applied to more than one coding decision. *Dichotomous* coding decisions involve a choice from among two competing coding possibilities, most often involving the presence or absence of one Comprehensive System code. *Polychotomous* coding decisions involve a choice from among three or more competing coding possibilities and two or more Comprehensive System codes. *Absent* may or may not be a possible coding category, depending on the nature of the specific coding decision. This organizational scheme produces one reliability coefficient for each of the 60 coding decisions and one reliability coefficient for each of their constituent codes.

The Location segment consists of four individual codes that yield two coding decisions: (a) Does the response involve the whole blot (*W*), common detail (*D*), or unusual detail (*Dd*); and (b) does the response involve white space area (*S*) or not?

The Developmental Quality segment consists of three individual codes that yield two coding decisions: (a) Do response percepts involve specific (*o*) or diffuse (*v*) form demand, and (b) Are response percepts involved in a meaningful relationship (+) or not. Within this data organization scheme, (a) the individual code *DQ*[*o*] is applied to all responses that involve at least one object with specific form demand, (b) the individual code *DQ*[*v*] is applied to all responses that do not involve at least one object with specific form demand, and (c) the individual code *DQ*[+] is applied to all responses that involve two or more objects in a meaningful relation, regardless of form demand. This data organization scheme can replicate the Comprehensive System coding criteria using only three individual codes, because the four Developmental Quality codes outlined by the Comprehensive System are in fact composite coding categories. Explicitly, (a) *DQ*[+] represents cooccurrence of specific form demand and percept synthesis (i.e., [*o*] and [+]); (b) *DQ*[*o*] represents specific form demand in the absence of percept synthesis; (c) *DQ*[*v*] represents diffuse form demand in the absence of percept synthesis; and (d) *DQ*[*v*/+] represents the cooccurrence of diffuse form demand and percept synthesis (i.e., [*v*] and [+]).

The Determinants segment consists of 29 individual codes that yield 11 coding decisions. Six of these coding decisions involve determinants (i.e., Color, Achromatic Color, Texture, Vista, Diffuse Shading, and Reflection) with polychotomous categories that reflect varying degrees of form predominance.[4] For example, Does this tex-

[4] We incorporate Color Naming (*CN*) as a category of the color determinant.

TABLE 1
Reliability Coefficients for Response-Level Codes and Coding Decisions by Segment

		Nonpatient Sample[a]				*Clinical Sample*[b]			
Codes	*Level of Analysis*	*PA*	κ	*PC*	*BR*	*PA*	κ	*PC*	*BR*
Location									
W, D, Dd[c]	CD	.963	.896	.415	1.000	.936	.891	.412	1.000
S[c]	CD, IC	.964	.851	.756	.142	.987	.939	.782	.124
W[c]	IC	.956	.913	.500	.498	.963	.925	.501	.521
D[c]	IC	.947	.888	.522	.396	.949	.889	.542	.354
Dd[c]	IC	.976	.873	.809	.107	.960	.816	.782	.124
Developmental Quality									
o, v	CD	.961	.722	.860	1.000	.973	.736	.899	1.000
+[c,d]	CD, IC	.913	.813	.533	.371	.906	.780	.575	.306
o[c,e]	IC	.961	.722	.860	.925	.973	.736	.899	.947
v[c,f]	IC	.961	.722	.860	.075	.973	.736	.899	.053
Determinants									
F[c]	CD, IC	.932	.857	.523	.391	.955	.909	.502	.467
Ma, Mp, Map[c]	CD	.976	.929	.660	.201	.963	.879	.690	.179
Ma[c]	IC	.978	.903	.774	.130	.971	.837	.820	.100
Mp[c]	IC	.983	.872	.867	.072	.971	.798	.855	.079
Map[c]	IC	1.000	—	1.000	.000	1.000	—	1.000	.000
FMa, FMp, FMap[c]	CD	.968	.880	.738	.149	.971	.889	.736	.150
FMa[c]	IC	.976	.880	.798	.114	.979	.885	.813	.104
FMp[c]	IC	.978	.679	.932	.035	.984	.804	.918	.043
FMap[c]	IC	1.000	—	1.000	.000	1.000	—	.995	.003
ma, mp, map[c]	CD	.961	.763	.836	.087	.979	.806	.890	.057
ma[c]	IC	.971	.699	.903	.051	.995	.906	.943	.029
mp[c]	IC	.976	.632	.934	.034	.981	.657	.945	.028
map[c]	IC	1.000	—	.995	.002	1.000	—	1.000	.000
FC, CF, C, CN[c]	CD	.947	.816	.710	.163	.949	.781	.768	.127
FC[c]	IC	.976	.821	.865	.073	.968	.722	.885	.061
CF[c]	IC	.956	.504	.912	.046	.968	.630	.913	.045
C[c]	IC	.981	.768	.916	.044	.989	.709	.963	.019
CN[c]	IC	1.000	—	1.000	.000	.997	—	.997	.001
FC´, C´F, C´[c]	CD	.959	.789	.805	.107	.976	.764	.898	.053
FC´[c]	IC	.966	.782	.844	.085	.981	.778	.916	.044
C´F[c]	IC	.993	.663	.978	.011	.995	—	.989	.005
C´[c]	IC	.998	.888	.978	.011	.997	—	.992	.004
FT, TF, T[c]	CD	.959	.436	.927	.038	.989	.795	.948	.027
FT[c]	IC	.961	.368	.939	.032	.992	.819	.956	.023
TF[c]	IC	.998	—	.998	.001	.997	—	.992	.004
T[c]	IC	1.000	—	.990	.005	1.000	—	1.000	.000
FV, VF, V[c]	CD	.968	.553	.929	.036	.995	—	.984	.008
FV[c]	IC	.976	.571	.943	.029	.995	—	.989	.005

(Continued)

TABLE 1 (Continued)

		Nonpatient Sample[a]				*Clinical Sample*[b]			
Codes	*Level of Analysis*	*PA*	κ	*PC*	*BR*	*PA*	κ	*PC*	*BR*
VF[c]	IC	.985	—	.986	.007	1.000	—	.995	.003
V[c]	IC	1.000	—	1.000	.000	1.000	—	1.000	.000
FY, YF, Y[c]	CD	.961	.698	.871	.068	.973	.589	.935	.033
FY[c]	IC	.976	.749	.903	.051	.976	.514	.950	.025
YF[c]	IC	.983	.453	.969	.016	.997	—	.992	.004
Y[c]	IC	.998	—	.998	.001	.997	—	.992	.004
FD[c]	CD, IC	.971	.791	.861	.075	.984	.617	.958	.021
Fr, rF[c]	CD	.995	.954	.894	.056	1.000	1.000	.928	.037
Fr[c]	IC	.995	.950	.903	.051	1.000	1.000	.938	.032
rF[c]	IC	1.000	—	.990	.005	1.000	—	.989	.005
Form Quality									
+, *o*, *u*, –, none	CD	.794	.681	.353	1.000	.824	.716	.378	1.000
+[c]	IC	.978	.597	.946	.028	—	—	—	.001
o[c]	IC	.879	.757	.500	.490	.904	.808	.499	.519
u[c]	IC	.850	.521	.686	.194	.869	.585	.684	.197
–[c]	IC	.883	.708	.601	.274	.888	.714	.608	.267
none[c,g]	IC	.998	.908	.974	.013	.989	.661	.968	.016
Pairs									
2[c]	CD, IC	.951	.883	.585	.294	.957	.901	.568	.316
Contents									
H[c]	CD, IC	.971	.893	.728	.163	.984	.937	.745	.150
(*H*)[c]	CD, IC	.971	.724	.895	.056	.995	.964	.852	.080
Hd[c]	CD, IC	.978	.841	.863	.074	.973	.692	.913	.045
(*Hd*)[c]	CD, IC	.998	—	.988	.006	.992	.765	.966	.017
Hx[c]	CD, IC	.993	—	.988	.006	.997	—	.992	.003
A[c]	CD, IC	.966	.929	.523	.393	.960	.920	.501	.475
(*A*)[c]	CD, IC	.990	.773	.957	.022	.995	.831	.968	.016
Ad[c]	CD, IC	.985	.919	.821	.100	.963	.676	.885	.061
(*Ad*)[c]	CD, IC	1.000	—	1.000	.000	.997	—	.992	.004
An[c]	CD, IC	.983	.779	.923	.040	.979	.656	.938	.032
Art[c]	CD, IC	.995	—	.986	.007	.989	.773	.953	.024
AY[c]	CD, IC	.985	.493	.971	.015	.992	.865	.940	.031
Bl[c]	CD, IC	.998	.946	.955	.023	.997	.922	.966	.016
Bt[c]	CD, IC	.983	.836	.897	.055	.992	.919	.901	.052
Cg[c]	CD, IC	.966	.732	.873	.068	.989	.923	.861	.075
Cl[c]	CD, IC	.995	—	.986	.007	.997	.888	.976	.012
Ex[c]	CD, IC	1.000	1.000	.967	.017	1.000	1.000	.979	.011
Fi[c]	CD, IC	.998	.940	.960	.021	.997	.888	.976	.012
Fd[c]	CD, IC	.998	—	.983	.008	.989	.745	.958	.021
Ge[c]	CD, IC	1.000	—	.986	.007	1.000	—	1.000	.000
Hh[c]	CD, IC	.976	.655	.930	.036	.971	.607	.925	.037

(Continued)

TABLE 1 (Continued)

		Nonpatient Sample[a]				*Clinical Sample*[b]			
Codes	*Level of Analysis*	*PA*	κ	*PC*	*BR*	*PA*	κ	*PC*	*BR*
Ls[c]	CD, IC	.995	.963	.869	.070	.989	.883	.908	.048
Na[c]	CD, IC	.973	.408	.955	.023	.984	.617	.958	.021
Sc[c]	CD, IC	.993	.796	.964	.018	.997	.946	.950	.025
Sx[c]	CD, IC	.995	.798	.976	.012	.992	.892	.925	.039
Xy[c]	CD, IC	1.000	—	1.000	.000	1.000	—	.995	.003
Id[c]	CD, IC	.956	.677	.865	.073	.960	.645	.887	.060
Populars									
P[c]	CD, IC	.956	.889	.606	.269	.973	.935	.592	.286
Organizational Quality									
1.0–6.5[c]	CD	.847	.811	.191	.637	.885	.856	.201	.620
1.0[c]	IC	.976	.892	.776	.129	.981	.925	.751	.146
2.0[c]	IC	.985	.793	.930	.036	.981	.811	.901	.051
2.5[c]	IC	.961	.815	.790	.119	.971	.844	.811	.106
3.0[c]	IC	.951	.696	.840	.087	.987	.902	.864	.074
3.5[c]	IC	.985	.494	.971	.015	.992	.819	.956	.023
4.0[c]	IC	.954	.708	.842	.086	.963	.700	.875	.066
4.5[c]	IC	.978	.879	.819	.101	.992	.957	.815	.103
5.0[c]	IC	.990	—	.985	.007	1.000	—	1.000	.000
5.5[c]	IC	.985	.856	.899	.053	.981	.819	.896	.055
6.0[c]	IC	.998	—	.993	.004	.997	—	.997	.001
6.5[c]	IC	1.000	—	1.000	.000	1.000	—	1.000	.000
Special Scores									
DV1, DV2[c]	CD	.947	.349	.918	.042	.960	.530	.915	.044
DV1[c]	IC	.954	.273	.937	.033	.963	.346	.943	.029
DV2[c]	IC	.990	.496	.981	.010	.992	.724	.971	.015
DR1, DR2[c]	CD	.964	.304	.948	.027	.971	.713	.897	.053
DR1[c]	IC	.976	.364	.962	.019	.981	.749	.925	.039
DR2[c]	IC	.985	—	.986	.007	.987	.540	.971	.015
INC1, INC2[c]	CD	.927	.516	.850	.081	.955	.631	.877	.066
INC1[c]	IC	.937	.547	.861	.075	.960	.661	.882	.063
INC2[c]	IC	.988	—	.988	.006	.995	—	.995	.003
FAB1, FAB2[c]	CD	.966	.469	.936	.033	.973	.726	.902	.051
FAB1[c]	IC	.976	.432	.957	.022	.981	.657	.945	.028
FAB2[c]	IC	.988	.440	.978	.011	.992	.820	.956	.023
CON[c]	CD, IC	.998	—	.998	.001	1.000	—	1.000	.000
ALOG[c]	CD, IC	.988	—	.983	.008	.989	.495	.979	.011
PSV[c]	CD, IC	.990	.496	.981	.010	.981	.658	.945	.028
CFB[c]	CD, IC	1.000	—	1.000	.000	1.000	—	.995	.003
AB[c]	CD, IC	.990	—	.986	.007	.995	.797	.974	.013
AG[c]	CD, IC	.990	.862	.930	.036	.989	.795	.948	.027
COP[c]	CD, IC	.971	.712	.899	.053	.984	.849	.894	.056

(Continued)

TABLE 1 (Continued)

		Nonpatient Sample[a]				*Clinical Sample*[b]			
Codes	*Level of Analysis*	*PA*	κ	*PC*	*BR*	*PA*	κ	*PC*	*BR*
MOR[c]	CD, IC	.981	.836	.882	.063	.984	.825	.908	.048
PER[c]	CD, IC	.981	.799	.903	.051	.984	.742	.938	.032
CP[c]	CD, IC	.998	—	.993	.004	1.000	—	1.000	.000

Note. PA = proportion observed agreement; PC = proportion chance agreement; BR = base rate; CD = coding decision statistic; IC = individual code statistic. Mean|median kappa statistics for low BR codes were .482|.530 (nonpatient sample) and .610|.665 (clinical sample). BR (number of occurrences of the code/number of responses) was calculated separately for each rater and then averaged across both raters. Dashes indicate codes that were excluded due to low BR (< .01).

[a] *N* = 412 responses. [b] *N* = 374 responses. [c] Absent coding decision. [d] This coding decision incorporates occurrences of all Developmental Quality synthesis codes (i.e., (+) and (*v*/+)). [e] This individual code incorporates occurrences of Ordinary (*o*) and Ordinary/Synthesis (+) Developmental Quality codes. [f] This individual code incorporates occurrences of Vague (*v*) and Vague/Synthesis (*v*/+) Developmental Quality codes. [g] This “absence” individual code corresponds to agreement for nonoccurrences of Form Quality codes within individual responses.

ture percept incorporate (a) more form features than shading features (*FT*), (b) more shading features than form features (*TF*), or (c) shading features exclusively (*T*). Three coding decisions involve determinants (i.e., human movement, animal movement, and inanimate movement) with polychotomous categories that reflect varying states of percept movement. For example, are these human percepts involved in (a) an active form of movement (M^a), (b) a passive form of movement (M^p), (c) both active and passive forms of movement ($M^{a\text{–}p}$), or (d) no movement? Finally, two coding decisions involve determinants (i.e., form dimension and pure form) with dichotomous categories that reflect simple presence or absence coding. For example, is this percept based solely on the form features of the inkblot (*F*) or not?

The Form Quality segment consists of five individual codes that yield a single[5] coding decision: Do the percepts involve an overelaborated (+), ordinary (*o*), unusual (*u*), arbitrary (–), or no (none) use of the inherent form features within each inkblot?

The Pairs segment consists of one individual code that yields a single present or absent coding decision: Are these percepts based on the symmetry features of the inkblot (2) or not? Similarly, the Content segment consists of 27 individual codes that yield 27 present or absent coding decisions. For example, do the contents of this response involve food (*Fd*) or not. Analogous to the Pairs segment, the Populars segment consists of one individual code that yields a single present or ab-

[5] We suspect that detailed articulation (+) may, in actuality, represent a distinct quality that is independent of ordinary form features and applicable to unusual and minus form features. Hence, it may potentially represent a separate coding decision.

sent coding decision: Does this response involve the most commonly perceived inkblot percepts (*P*) or not?

The Organizational Quality segment consists of four individual codes that would normally yield one coding decision: Does this response (a) integrate form features of the whole inkblot (*ZW*), (b) integrate concrete form features and white space areas of the inkblot (*ZS*), (c) integrate adjacent form features of the inkblot (*ZA*), (d) integrate nonadjacent (i.e., distant) form features of the inkblot (*ZS*), or (e) not involve meaningful integration of the form features of the inkblot? However, we were unable to model this coding decision for reliability analysis because the RIAP3 sequence of scores does not list the Organizational Quality codes that were originally entered by the raters. Although RIAP3 does provide the numerical values that are associated with raters' original Organizational Quality coding decisions, these values are not mutually exclusive for any one Organizational Quality code (e.g., Card II: *ZW* = *ZS* = 4.5; Card VI: *ZW* = *ZA* = 2.5; Exner, 1995). Consequently, it was necessary to base our data organization scheme solely on these numerical values. Although this arrangement does not provide an estimate of rater agreement for Organizational Quality codes, the resulting response-level reliability estimates are still valid in the sense that the numerical values reflect the actual data on which interpretive indexes (e.g., *Zd*) are derived.

The Special Scores segment consists of 18 individual codes that yield 14 coding decisions. Four of these coding decisions involve special scores (i.e., Incongruous Combination, Fabulized Combination, Deviant Verbalization, and Deviant Response) with polychotomous categories that reflect varying degrees of cognitive slippage. For example, does this percept involve (a) appropriate attributes, (b) a mildly unrealistic combination of attributes (*Inc1*), or (c) a moderately or severely unrealistic combination of attributes (*Inc2*)? The remaining 14 coding decisions involve special scores with dichotomous categories that reflect simple presence or absence coding. For example, does this response involve morbid (*MOR*) content or not?

Calculating Response-Level Reliability

Response-level reliability estimates were derived using PA and Cohen's unweighted kappa (κ), according to the following formulas:

$$PA = \frac{\text{No. of Occurrence and Nonoccurrence Agreements}}{\text{No. of Total Responses}} \qquad (1)$$

$$\kappa = \frac{PA - PC}{1 - PC} \qquad (2)$$

where proportion of chance agreement (PC) equals the product of the values of raters' marginal proportions for each coding category summed over all categories (Zwick, 1988).

Additionally, we calculated base rate of occurrence (BR) for each individual code and coding decision as the proportion of occurrences of each individual or composite code averaged across both raters.

Protocol-Level Data

Protocol-level data include aggregate codes, ratios, percentages, and derivations from the Comprehensive System Structural Summary. These data are most easily organized for reliability analysis by using the RIAP3 export facility, and the Pcanalx syntax file that reads the resulting ASCII text files into an SPSS (SPSS, Inc., 1998) database.[6] In most cases, the resulting Structural Summary data are produced in an appropriate format for immediate reliability analysis. One anomaly, however, seems to occur for Structural Summary ratios that involve divisions between nonzero numerators and zero denominators. RIAP3 will report the values of these undefined ratios as the numerator values alone, yet we could find no explanation in any of Exner's works or in the RIAP3 manual as to why this should be so. Therefore, we performed a manual check of our Structural Summary ratio data and excluded all data points that corresponded to undefined ratios.

Calculating Protocol-Level Reliability

Reliability estimates for all Structural Summary data were derived using the ICC that is based on a two-way random effects model of variance and an absolute agreement definition of correlation (McGraw & Wong, 1996; Shrout & Fleiss, 1979). This ICC formula is calculated as follows:

$$ICC(\mathrm{A},1) = \frac{MS_{\mathrm{R}} - MS_{\mathrm{E}}}{MS_{\mathrm{R}} + (k-1)MS_{\mathrm{E}} + \frac{k}{n}(MS_{\mathrm{c}} - MS_{\mathrm{E}})} \tag{3}$$

where MS_{R} is the mean square for rows (i.e., scores), MS_{C} is the mean square for columns (i.e., raters), MS_{E} is the mean square error, k is the number of raters, and n is the number of protocols.

[6] An updated syntax file that associates variable names with the Structural Summary data is available from Mark S. Verschell.

RESULTS

Table 1 presents PA, Cohen's kappa (κ), PC, and BR data for response-level coding decisions and individual codes. Data are organized according to the nine segments of the Comprehensive System, and are identified as representing a coding decision (CD), an individual code (IC), or in many cases, both. Due to the difficulty of making distinctions under conditions of highly restricted true score variability (e.g., Shrout et al., 1987), kappa coefficients were excluded when the base rates of their underlying codes fell below .01 (i.e., < 1 occurrence per 100 responses; G. Meyer, personal communication, August 20, 1998). Excluded coefficients are indicated in Table 1 by a dash. Of the 116 ICs and CDs that were examined, 27 (23%) were excluded from the nonpatient group (*M*|*Mdn* κ = .482|.530), and 28 (24%) were excluded from the clinical group (*M*|*Mdn* κ = .610|.665).

In the nonpatient group, kappa values ranged from .273 on Deviant Verbalization (*DV1*) to 1.000 on Explosion (*Ex*), with mean and median kappa values of .726 and .776, respectively. In the clinical group, kappa values ranged from .346 for Deviant Verbalization (*DV1*) to 1.000 for *Ex*plosion (*Ex*), with mean and median kappa values of .784 and .798, respectively.

Considering Determinant coding decisions and individual codes exclusively, nonpatient kappa values ranged from .368 for Form–Texture (*FT*) to .954 for Form Reflection (*Fr*), with mean and median kappa values of .737 and .775, respectively. Kappa values for Determinants in the clinical group ranged from .589 for Shading-Diffuse to 1.000 for Form Reflection, with mean and median kappa values of .786 and .798, respectively.

Considering Content coding decisions and individual codes exclusively, nonpatient kappa values ranged from .408 for Nature (*Na*) to 1.000 for *Ex*plosion (*Ex*), with mean and median kappa values of .795 and .798, respectively. Kappa values for Contents in the clinical group ranged from .607 for Household (*Hh*) to 1.000 for Blood (*Bl*), with mean and median kappa values of .827 and .883, respectively.

Considering Special Score coding decisions and individual codes exclusively, nonpatient kappa values ranged from .273 for Deviant Verbalization (*DV1*) to .862 for Aggressive Movement (*AG*), with mean and median kappa values of .526 and .496, respectively. Kappa values for Special Scores in the clinical group ranged from .346 for Deviant Verbalization (*DV1*) to .849 for Cooperative Movement (*COP*), with mean and median kappa values of .681 and .719, respectively.

Table 2 presents ICCs for protocol-level aggregate codes, ratios, percentages, and derivations from the Structural Summary. Data are organized according to Exner's (1991) framework of interpretive clusters. ICCs were excluded when

TABLE 2
ICCs for Structural Summary Variables

	Nonpatient Sample[a]				*Clinical Sample*[a]			
Cluster	*ICC*	*Low*	*Up*	*n*	*ICC*	*Low*	*Up*	*n*
Controls and Stress								
Adjusted *D*	.533**	.151	.781	20	.678***	.342	.859	20
Coping Deficit Index	.732***	.433	.885	20	.648***	.301	.844	20
EA	.963***	.910	.985	20	.923***	.815	.969	20
M	.981***	.948	.993	20	.951***	.880	.980	20
Weighted Sum *C*	.960***	.904	.984	20	.913***	.792	.965	20
Sum *C*	.984***	.961	.994	20	.944***	.865	.977	20
Adjusted *es*	.852***	.584	.944	20	.933***	.839	.973	20
es	.948***	.797	.982	20	.911***	.790	.964	20
FM	.941***	.857	.976	20	.947***	.870	.979	20
Sum *C´*	.804***	.426	.928	20	.873***	.709	.947	20
Sum *V*	.576**	.180	.809	20	†	†	†	†
Sum *T*	.377*	–.023	.687	20	.856***	.671	.941	20
D	.826***	.586	.929	20	.549**	.146	.794	20
m	.911***	.736	.967	20	.827***	.612	.928	20
Sum *Y*	.680***	.357	.859	20	.711***	.399	.875	20
Affect								
Depression Index	.643***	.298	.841	20	.914***	.793	.965	20
EB Pervasive	.998***	.993	.999	11	.931***	.772	.981	11
FC	.913***	.752	.967	20	.777***	.524	.905	20
CF	.738***	.414	.891	20	.861***	.686	.942	20
C	.796***	.550	.914	20	.651***	.297	.847	20
CF + *C*	.870***	.686	.948	20	.856***	.674	.940	20
FC – *CFC*	.706***	.358	.876	20	.746***	.469	.890	20
FC: CF + *C*	.543**	.071	.823	14	.165	–.504	.685	11
Affective Ratio	1.000***	1.000	1.000	20	1.000***	1.000	1.00	20
CP	†	†	†	†	†	†	†	†
Space	.865***	.696	.944	20	.927***	.812	.971	20
Blends: *R*	.954***	.888	.982	20	.907***	.781	.962	20
Color Shading Blends	.684***	.267	.871	20	.835***	.633	.931	20
Self-Perception								
Fr + *rF*	.979***	.947	.991	20	1.000	††	††	20
Egocentricity Index	.984***	.959	.993	20	.984***	.960	.994	20
Form Dimension	.809***	.585	.919	20	.678***	.358	.857	20
H	.952***	.883	.981	20	.949***	.876	.980	20
(*H*)	.818***	.597	.924	20	.982***	.956	.993	20
Hd	.870***	.706	.946	20	.757***	.481	.896	20
(*Hd*)	†	†	†	†	.862***	.686	.943	20
(*H*) + *Hd* + (*Hd*)	.933***	.840	.973	20	.856***	.674	.940	20
H – ((*H*) + *Hd* + (*Hd*))	.900***	.766	.959	20	.848***	.657	.937	20
Hx	†	†	†	†	†	†	†	†

(Continued)

TABLE 2 (Continued)

	Nonpatient Sample[a]				*Clinical Sample*[a]			
Cluster	*ICC*	*Low*	*Up*	*n*	*ICC*	*Low*	*Up*	*n*
An	.812***	.584	.921	20	.746***	.460	.891	20
Xy	†	†	†	†	†	†	†	†
An + *Xy*	.812***	.584	.921	20	.744***	.457	.890	20
Morbid	.882***	.730	.951	20	.810***	.543	.924	20
Interpersonal Perception								
Hypervigilance Index	.810***	.565	.921	20	.872***	.638	.951	20
Active	.951***	.882	.980	20	.876***	.714	.949	20
Passive	.793***	.243	.932	20	.712***	.398	.876	20
a–p	.784***	.441	.916	20	.584**	.193	.813	20
Fd	†	†	†	†	.678***	.355	.858	20
H + (*H*) + *Hd* + (*Hd*)	.974***	.922	.990	20	.921***	.812	.968	20
PER	.888***	.743	.954	20	.954***	.889	.981	20
COP	.800***	.563	.916	20	.825***	.593	.928	20
AG	.871***	.707	.946	20	.937***	.848	.974	20
Isolation Index	.767***	.506	.900	20	.769***	.505	.901	20
Processing								
Lambda	.932***	.815	.974	20	.983***	.958	.993	20
OBS Index	†††			†††				
Zf	.977***	.944	.991	20	.938***	.849	.975	20
W	.985***	.963	.994	20	.956***	.893	.982	20
D	.969***	.924	.988	20	.966***	.902	.987	20
Dd	.941***	.858	.976	20	.958***	.820	.986	20
D + *Dd*	.984***	.959	.994	20	.985***	.964	.994	20
W:M	.936***	.842	.974	20	.946***	.862	.980	17
DQ+	.909***	.787	.963	20	.718***	.422	.877	20
DQv	.844***	.653	.935	20	.880***	.723	.951	20
DQvp	.513**	.084	.778	20	†	†	†	†
Zd	.665***	.113	.875	20	.762***	.494	.898	20
PSV	†	†	†	†	.826***	.596	.929	20
Zsum	.960***	.904	.984	20	.860***	.680	.942	20
Mediation								
Popular	.719***	.417	.879	20	.934***	.841	.973	20
FQx+	.893***	.749	.956	20	†	†	†	†
X + %	.745***	.461	.890	20	.815***	.440	.933	20
F + %	.292	–.109	.630	20	.823***	.607	.926	20
Xu%	.156	–.245	.534	20	.483*	.056	.759	20
X – %	.621***	.272	.829	20	.656***	.324	.847	20
S – %	.269	–.208	.634	20	.837***	.631	.932	20
Ideation								
Schizophrenia Index	.452*	.049	.735	20	.560**	.184	.797	20
a:p	.679***	.218	.881	16	.796***	.527	.920	17
Ma	.954***	.888	.982	20	.870***	.701	.947	20

(Continued)

TABLE 2 (Continued)

	Nonpatient Sample[a]				*Clinical Sample*[a]			
Cluster	*ICC*	*Low*	*Up*	*n*	*ICC*	*Low*	*Up*	*n*
Mp	.747***	.470	.891	20	.751***	.468	.894	20
Ma – Mp	.824***	.607	.926	20	.724***	.420	.881	20
Intellect Index	.889***	.744	.954	20	.899***	.745	.960	20
Sum6	.830***	.623	.929	20	.880***	.714	.951	20
WSum6	.738***	.457	.887	20	.801***	.557	.917	20
Level 2	.296	–.173	.650	20	.678***	.355	.858	20
M–	.632***	.283	.835	20	.887***	.739	.954	20
M none	†	†	†	†	†	†	†	†
DV1	.156	–.293	.551	20	.368**	–.069	.689	20
DV2	†	†	†	†	.523**	.134	.776	20
INC1	.689***	.366	.864	20	.614***	.250	.826	20
INC2	†	†	†	†	†	†	†	†
DR1	.685***	.333	.865	20	.814***	.590	.922	20
DR2	†	†	†	†	.257	–.180	.617	20
FAB1	.894***	.756	.956	20	.665***	.336	.851	20
FAB2	.504**	.113	.765	20	.816***	.597	.922	20
ALOG	†	†	†	†	.816***	.589	.923	20
CONTAM	†	†	†	†	†	†	†	†
Suicide Constellation	.549**	.152	.793	20	.674***	.351	.855	20

Note. Low base-rate mean|median ICC values were .443|.444 (nonpatient sample) and .337|.040 (clinical sample). ICC = intraclass correlation coefficient, two-way random effects model with absolute agreement definition of correlation; Up and Low = upper and lower bounds of the 95% confidence interval for the single rater estimate; *n* = number of protocols with valid occurrences of the respective variable; † = response-level base-rate < .01 averaged across both raters; †† = *F* ratio undefined; ††† = protocol-level base rate < .05 averaged across both raters.

[a] *N* = 20 protocols.

**p* < .05. ***p* < .01. ****p* < .001.

the base rates of their underlying variables fell below .05 (i.e., < 1 occurrence in 20 protocols), or when the base rates of their constituent response-level codes fell below a base rate of .01 (i.e., < 1 occurrence per 100 responses; G. Meyer, personal communication, August 23, 1998). Excluded reliability coefficients are indicated in Table 2 by a dagger. Of the 95 variables examined, 13 (14%) were excluded from the nonpatient sample (*M*|*Mdn* ICC = .443|.444), and 10 (11%) were excluded from the clinical sample (*M*|*Mdn* ICC = .337|.040).

In the nonpatient group, ICC values ranged from .156 for Deviant Verbalization (*DV1*) and *Xu*% to 1.00 for Affective Ratio (*Afr*), with mean and median ICC values of .780 and .767, respectively. In the clinical group, ICC values ranged from .165 for *FC:CF+C* to 1.00 for Affective Ratio (*Afr*) and *Fr+rF*, with mean and median kappa values of .803 and .810, respectively.

DISCUSSION

This study examined the reliability of Comprehensive System data that were collected from samples of nonpatient and clinical protocols. Response-level reliability was evaluated using the kappa coefficient, and protocol-level reliability was evaluated using the ICC. This study also examined the relation among response-level reliability, quantification method, and base rate.

Interpreting Reliability Coefficients

Kappa is interpreted as the proportion of possible agreement that is achieved by raters beyond chance agreement. Kappa values may range from –1.00 to 1.00. Kappa values greater than zero indicate that raters agree more frequently than would be predicted by chance. A kappa value of zero indicates that raters' observed agreement is no better than chance. Kappa values of less than zero indicate that raters agree less frequently than would be predicted by chance.

Kappa is directly related to the proportion of observed agreement. For example, the Location coding decision in the clinical sample indicates that raters achieved a PC equal to .412 and reliability estimates of .891 and .936 for kappa and PA, respectively. The possible range of agreement beyond chance agreement was equal to .588 (1 – .412), and the kappa value of .891 indicates that the raters achieved 89% of this range. Thus, the proportion of raters' nonchance agreement was equal to .524 (.891 × .588 = .524). Adding the proportion of nonchance agreement to the PC yielded the PA (PA = .412 + .524 = .936).

The ICC is interpreted as the proportion of total variance in observers' ratings that is attributable to true variation among target behaviors (i.e., verbal responses; Bartko, 1991). ICC values will usually range from 0 to 1.0, with 1.0 indicating perfect observer agreement when there is at least some degree of variation among the target behaviors. Negative ICC values are interpreted as representing zero reliability (Bartko, 1976).

For a given amount of true score variance, the magnitude of reliability coefficients will vary as a function of the different factors that contribute to error variance. These factors include (a) the adequacy of the behavioral observation coding criteria, (b) the occasion on which the behavioral observation data were collected, (c) the nature (i.e., signal) of the actual behaviors under investigation, (d) the training and experience of the observers, and (e) the setting or context in which the actual behaviors are observed (Bartko, 1991). Although estimates of observer reliability should always be interpreted with these factors in mind, several general and reasonably similar guidelines (e.g., Fleiss, 1981; Gelfand & Hartmann, 1975; Landis & Koch, 1977) have been proposed for interpreting the magnitude of the kappa coefficient. For the purpose of analyzing response-level data within this study, we interpret kappa values (a) less than .61 as representing unacceptable levels of reliability, (b) greater than or equal to .61 and less than .81 as representing substantial and acceptable levels of reliability, and (c) greater than or equal to .81 as representing excellent levels of reliability.

Guidelines for interpreting the magnitude of ICCs appear relatively infrequently in the research literature and are more varied in their recommendations than are the corresponding guidelines for interpreting the kappa coefficient. For example, Mitchell (1979) suggested that ICC values of .50 and .60 should not be considered as low. Shrout and Fleiss (1979) suggested that ICC values of .75 or .80 are minimally acceptable in substantive studies. Berk (1979) suggested that ICC values of .80 and above are indicative of a “high degree of agreement” (p. 467). To a large extent, these differences represent the differing generalizations (i.e., generalizability theory; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) that may be derived from behavioral observation data and differing forms of the ICC.

ICCs that estimate the reliability of a single observer’s scores are expected to be lower in value than ICCs that estimate the average reliability of multiple observers’ scores (Shrout & Fleiss, 1979). Although the reliability of Comprehensive System summary scores could be evaluated using the latter approach (i.e., average score ICCs; McGraw & Wong, 1996), the former approach is more conservative and provides a lower bound estimate of the true dependability of protocol-level data. In addition, this approach produces reliability estimates that seem to be appropriately characterized by the same interpretive framework that has been postulated for the kappa coefficient. Therefore, this study applied the previously described interpretive framework to both response-level and protocol-level data.

Response-Level Reliability

The results of this study indicate that well-trained and experienced raters can apply the majority of Comprehensive System codes in a consistent fashion. Of the 88 individual codes and coding decisions that met our base-rate inclusion criteria in the nonpatient sample, 36 (41%) demonstrated kappa values in the range of excellent reliability ($\kappa \geq .81$), 30 (36%) demonstrated kappa values in the range of substantial reliability ($.61 \leq \kappa < .81$), and 22 (25%) demonstrated kappa values in the range of unacceptable reliability ($\kappa < .61$). Individual codes and coding decisions that fell within the unacceptable range of reliability include *FQ+, FV,* Shading-Vista, *INC1, FQu,* Incongruous Combination, *CF, DV2, PSV, Z* = 3.5, *Ay,* Fabulized Combination, *YF, FAB2, FAB1,* Shading-Texture, *Na, FT, DR1,* Deviant Verbalization, *DV1,* and Deviant Response (Table 1).

Mean and median kappa values in the nonpatient sample were in the upper range of substantial reliability across all codes combined (.726|.776), across Determinant codes exclusively (.737|.775), and across Content codes exclusively (.795|.798). Special Scores (*M*|*Mdn* κ = .526|.496) and other low-frequency codes appeared to be the most problematic.

Of the 89 individual codes and coding decisions that met our base-rate inclusion criteria in the clinical sample, 42 (47%) demonstrated kappa values in the range of excellent reliability (κ ≥ .81), 39 (44%) demonstrated kappa values in the range of substantial reliability (.61 ≤ κ < .81), and 8 (9%) demonstrated kappa values in the range of unacceptable reliability (κ < .61). Individual codes and coding decisions that fell within the unacceptable range of reliability include *ALOG,* Deviant Verbalization, *DR2, DV1, FQu, Hh, FY,* and Shading-Diffuse.

In the clinical sample, mean and median kappa values were in the upper range of substantial reliability across all codes combined (.784|.798) and across Determinant codes exclusively (.786|.798). Mean and median kappa values were in the middle range of substantial reliability across Special Score codes exclusively (.681|.719). Mean and median kappa values were in the range of excellent reliability across Content codes exclusively (.827|.883).

Protocol-Level Reliability

The results of this study also indicate that well-trained and experienced raters produce consistent Comprehensive System protocols across the majority of aggregate codes, percentages, ratios, and derivations from the Structural Summary. Of the 82 variables that met our base-rate inclusion criteria in the nonpatient sample, 45 (55%) demonstrated ICC values in the range of excellent reliability (ICC ≥ .81), 24 (29%) demonstrated ICC values in the range of substantial reliability (.61 ≤ ICC < .81), and 13 (16%) demonstrated ICC values in the range of unacceptable reliability (ICC < .61). Variables that fell within the unacceptable range of reliability include *AdjD, DQv*/+, *F*+%, *DV1, FAB2, FC:CF+C,* Level 2, *S*-%, *SCON, SCZI,* Sum *T,* Sum *V,* and *Xu*%. Mean ICC values in the nonpatient sample fell within the upper range of substantial reliability (.780), and median ICC values fell within the range of excellent reliability (.825).

Of the 85 variables that met our base-rate inclusion criteria in the clinical sample, 53 (62%) demonstrated ICC values in the range of excellent reliability (ICC ≥ .81), 24 (28%) demonstrated ICC values in the range of substantial reliability (.61 ≤ ICC < .81), and 8 (9%) demonstrated ICC values in the range of unacceptable reliability (ICC < .61; rounding error of 1% noted). Variables that fell within the unacceptable range of reliability include *a–p, D* score, *DR2, DV1, DV2, FC:CF+C, SCZI,* and *Xu*%. Mean and median ICC values in the clinical sample fell within in the range of excellent reliability (.803|.837).

Sensitivity, Specificity, Base Rate, and Reliability

True reliability coefficients (e.g.,κ[7] and the ICC) indicate the proportion of total measurement variance that is attributable to true score differences among a set of rated behaviors ($\sigma^2_T / [\sigma^2_T + \sigma^2_E]$; Shrout et al., 1987). Given fixed levels of diagnostic sensitivity and specificity (i.e., the accuracy of fallible observers' ratings in relation to an absolute criterion), reliability estimates tend to decrease in value as the base rates of target behaviors approach 0 or 100% (Bartko, 1991; Grove, Andreasen, McDonald-Scott, Keller, & Shapiro, 1981; Kraemer, 1979; Shrout et al., 1987). Conversely, reliability estimates tend to be maximized when observers are presented with the most varied number of behaviors to rate (i.e., maximum true score variance at a base rate of 50%; Bartko, 1991).

Reliability coefficients are also affected by changes in the sensitivity and specificity of observational procedures across populations with differing base rates (Langerbucher, Labouvie, & Morgenstern, 1996). Under certain situations of decreasing base rates, increases in diagnostic specificity (i.e., the rate of true negative classifications) are not offset by decreases in diagnostic sensitivity (i.e., the rate of true positive classifications). Hence, reliability estimates may actually be higher in homogenous populations than in heterogeneous populations. These situations correspond to increases in positive predictive value, fewer diagnostic classification errors, and lower error variance (σ^2_E).

Considerable debate has centered on the appropriateness of calculating reliability estimates in the context of low base-rate phenomena. Although Carey and Gottesman (1978) coined the phrase *the base-rate problem,* their intent was to reference the inappropriateness of generalizing reliability estimates across observational phenomena with differing base rates, as opposed to assessing reliability in low base-rate situations. Grove et al. (1981) were perhaps the first to suggest that reliability estimates not be reported for low base-rate phenomena (i.e., < 5%), for concerns that low reliability estimates might lead researchers to reject valid diagnostic procedures (i.e., those with acceptable levels of sensitivity and specificity). Spitznagel and Helzer (1985) further recommended that Yule's *Y* be substituted for kappa when assessing reliability in low base-rate situations, as this statistic does not share the same base-rate dependence (i.e., problem) as the kappa statistic. These recommendations, however, have since been criticized by several researchers (e.g., Bartko, 1991; Langerbucher et al., 1996; Shrout et al., 1987) who have pointed out that (a) Yule's *Y* is in fact a measure of association and may only be interpreted as an index of reliability under rare and restrictive conditions and (b) Yule's *Y* produces biased estimates of observational reliability and may mislead researchers into believing that measurement error is not a significant problem,

[7] Variance components for the kappa coefficient can be conceptualized as arising from the differing weights that are assigned to the various cross-classifications of observers' ratings.

when, in fact, it really is. Furthermore, these researchers have explained that (a) the low base-rate problem is not a function of reliability statistics per se but rather represents the difficulty of making distinctions in increasingly homogenous populations and (b) the kappa statistic demonstrates stable psychometric properties in properly designed reliability studies.

Despite these clarifications, developing standards within Comprehensive System reliability research still call for the exclusion of “unstable” response-level codes and protocol-level variables that occur at extremely low base rates. To further clarify these issues, we reexamined our response-level data, including codes and coding decisions that were previously excluded on the basis of their low base rates. For the purpose of these analyses, we contrasted reliability estimates above and below the 5% low base-rate cutoff recommended by Grove et al. (1981).

Table 3 presents response-level data for both samples sorted by descending base rate. In the nonpatient sample, 53% of the low base-rate codes and coding decisions fell within the unacceptable reliability range (i.e., $\kappa < .61$), with an average kappa value of .564. In contrast, only 5% of the higher base-rate codes and coding decisions (i.e., > 5%) fell within the unacceptable reliability range, with an average kappa value of .803. In the clinical sample, 21% of the low base-rate codes and coding decisions fell within the unacceptable reliability range, with an average kappa value of .710. In contrast, only 2% of the higher base-rate codes and coding decisions fell within the unacceptable reliability range, with an average kappa value of .817. These results highlight the difficulty of making distinctions in increasingly homogenous populations, as lower base-rate Comprehensive System codes and coding decisions were indeed associated with lower estimates of observer reliability.

To shed further light on these issues, we plotted kappa, PA, and PC in relation to increasing base rates of response-level codes and coding decisions. Figures 1 (nonpatient) and 2 (clinical) clearly illustrate why PA is an inappropriate measure of reliability (e.g., Spitzer, Cohen, Fleiss, & Endicott, 1967), as it produces reliability estimates that are insensitive to variations in the base rate of behavioral occurrence and, correspondingly, to the degree of chance agreement. Kappa, on the other hand, is much more sensitive to variations in the rate of behavioral occurrence and the degree of chance agreement. Although our data confirm Shrout et al.’s (1987) assertions that there is no mathematical necessity for small kappa values to be associated with low base-rate phenomena, Figures 1 and 2 demonstrate that the variability of kappa values increased as codes and coding decisions occurred at increasingly lower base rates and as chance agreement correspondingly increased in value. Formally, kappa values for codes and coding decisions that occurred at a base rate of 5% or less demonstrated significantly greater group variability (i.e., as indicated by the average of their squared deviation scores about the sample mean κ value) than kappa values for codes and coding decisions that occurred at higher base rates: nonpatient sample $F(1, 106) = 26.935$, $p = .000$, and clinical sample $F(1, 105) = 7.250$, $p = .008$. Furthermore, our data demonstrate the well-established relation (e.g., Hanley, 1987) be-

TABLE 3
Reliability Coefficients for Response-Level Data by Descending Base Rate

Nonpatient Sample[a]			*Clinical Sample*[b]		
Codes and Coding Decisions	κ	*Base Rate*	*Codes and Coding Decisions*	κ	*Base Rate*
FQ:p, o, u, –, none	.681	1.000	*DQ:o, v*	.736	1.000
DQ:o, v	.722	1.000	*FQ:p, o, u,* –, none	.716	1.000
W, D, Dd	.896	1.000	*W, D, Dd*	.891	1.000
DQo[c]	.722	.925	*DQo*[c]	.736	.940
Z = 1.0–6.5[c]	.811	.637	*Z* = 1.0–6.5[c]	.856	.620
FQo[c]	.757	.518	*W*[c]	.925	.521
W[c]	.913	.498	*FQo*[c]	.808	.520
D[c]	.888	.396	*A*[c]	.920	.475
A[c]	.929	.393	*F*[c]	.909	.467
F[c]	.857	.391	*D*[c]	.889	.354
DQ+[c]	.813	.371	(2)[c]	.901	.316
(2)[c]	.883	.294	*DQ+*[c]	.780	.306
FQ–[c]	.708	.274	*P*[c]	.935	.286
P[c]	.889	.269	*FQ–*[c]	.714	.267
Ma, Mp, Map[c]	.929	.201	*FQu*[c]	.585	.197
FQu[c]	.521	.194	*Ma, Mp, Map*[c]	.879	.179
FC, CF, C, CN[c]	.816	.163	*FMa, FMp, FMap*[c]	.889	.150
H[c]	.893	.163	*H*[c]	.937	.150
FMa, FMp, FMap[c]	.880	.149	*Z* = 1.0[c]	.925	.146
S[c]	.851	.142	*FC, CF, C, CN*[c]	.781	.127
Ma[c]	.903	.130	*Dd*[c]	.816	.124
Z = 1.0[c]	.892	.129	*S*[c]	.939	.124
Z = 2.5[c]	.815	.119	*Z* = 2.5[c]	.844	.106
FMA[c]	.880	.114	*FMA*[c]	.885	.104
FC´, C´F, C´[c]	.789	.107	*Z* = 4.5[c]	.957	.103
Dd[c]	.873	.107	*Ma*[c]	.837	.100
Z = 4.5[c]	.879	.101	(*H*)[c]	.964	.080
Ad[c]	.919	.100	*Mp*[c]	.798	.079
Z = 3.0[c]	.696	.087	*Cg*[c]	.923	.075
ma, mp, map[c]	.763	.087	*Z* = 3.0[c]	.902	.074
Z = 4.0[c]	.708	.086	*INC1, INC2*[c]	.631	.066
FC´[c]	.782	.085	*Z* = 4.0[c]	.700	.066
INC1, INC2[c]	.516	.081	*INC1*[c]	.661	.063
INC1[c]	.547	.075	*Ad*[c]	.676	.061
DQv[c]	.722	.075	*FC*[c]	.722	.061
FD[c]	.791	.075	*Id*[c]	.645	.060
Hd[c]	.841	.074	*ma, mp, map*[c]	.806	.057
Id[c]	.677	.073	*COP*[c]	.849	.056
FC[c]	.821	.073	*Z* = 5.5[c]	.819	.055
Mp[c]	.872	.072	*FC´, C´F, C´*[c]	.764	.053

(Continued)

TABLE 3 (Continued)

Nonpatient Sample[a]			*Clinical Sample*[b]		
Codes and Coding Decisions	κ	*Base Rate*	*Codes and Coding Decisions*	κ	*Base Rate*
Ls[c]	.963	.070	*DR1, DR2*[c]	.713	.053
FY, YF, Y[c]	.698	.068	*DQv*[c]	.736	.053
Cg[c]	.732	.068	*Bt*[c]	.919	.052
MOR[c]	.836	.063	*FAB1, FAB2*[c]	.726	.051
(*H*)[c]	.724	.056	*Z* = 2.0[c]	.811	.051
Fr, rF[c]	.954	.056	*Ls*[c]	.883	.048
Bt[c]	.836	.055	*MOR*[c]	.825	.048
COP[c]	.712	.053	*CF*[c]	.630	.045
Z = 5.5[c]	.856	.053	*Hd*[c]	.692	.045
ma[c]	.699	.051	*DV1, DV2*[c]	.530	.044
FY[c]	.749	.051	*FC´*[c]	.778	.044
PER[c]	.799	.051	*FMp*[c]	.804	.043
Fr[c]	.950	.051	*DR1*[c]	.749	.039
CF[c]	.504	.046	*Sx*[c]	.892	.039
C[c]	.768	.044	*Hh*[c]	.607	.037
DV1, DV2[c]	.349	.042	*Fr, rF*[c]	1.000	.037
An[c]	.779	.040	*FY, YF, Y*[c]	.589	.033
FT, TF, T[c]	.436	.038	*An*[c]	.656	.032
FV, VF, V[c]	.553	.036	*Fr*[c]	1.000	.032
Hh[c]	.655	.036	*PER*[c]	.742	.032
Z = 2.0[c]	.793	.036	*Ay*[c]	.865	.031
AG[c]	.862	.036	*DV1*[c]	.346	.029
FMp[c]	.679	.035	*ma*[c]	.906	.029
mp[c]	.632	.034	*FAB1*[c]	.657	.028
DV1[c]	.273	.033	*mp*[c]	.657	.028
FAB1, FAB2[c]	.469	.033	*PSV*[c]	.658	.028
FT[c]	.368	.032	*AG*[c]	.795	.027
FV[c]	.571	.029	*FT, TF, T*[c]	.795	.027
FQp[c]	.597	.028	*FY*[c]	.514	.025
DR1, DR2[c]	.304	.027	*Sc*[c]	.946	.025
Na[c]	.408	.023	*Art*[c]	.773	.024
Bl[c]	.946	.023	*FAB2*[c]	.820	.023
FAB1[c]	.432	.022	*FT*[c]	.819	.023
(*A*)[c]	.773	.022	*Z* = 3.5[c]	.819	.023
Fi[c]	.940	.021	*Food*[c]	.745	.021
DR1[c]	.364	.019	*FD*[c]	.617	.021
Sc[c]	.796	.018	*Na*[c]	.617	.021
Ex[c]	1.000	.017	*C*[c]	.709	.019
YF[c]	.453	.016	(*Hd*)[c]	.765	.017
Ay[c]	.493	.015	(*A*)[c]	.831	.016
Z = 3.5[c]	.494	.015	*Bl*[c]	1.000	.016

(Continued)

TABLE 3 (Continued)

Nonpatient Sample[a]			*Clinical Sample*[b]		
Codes and Coding Decisions	κ	*Base Rate*	*Codes and Coding Decisions*	κ	*Base Rate*
FQnone[c]	.908	.013	*FQnone*[c]	.661	.016
Sx[c]	.798	.012	*DR2*[c]	.540	.015
FAB2[c]	.440	.011	*DV2*[c]	.724	.015
C´F[c]	.663	.011	*AB*[c]	.797	.013
C´[c]	.888	.011	*Cl*[c]	.888	.012
DV2[c]	.496	.010	*Fi*[c]	.888	.012
PSV[c]	.496	.010	*ALOG*[c]	.495	.011
ALOG[c]	.281	.008	*Ex*[c]	1.000	.011
Food[c]	.856	.008	*FV, VF, V*[c]	.665	.008
DR2[c]	–.007	.007	*C´F*[c]	.497	.005
VF[c]	–.007	.007	*FV*[c]	.497	.005
AB[c]	.328	.007	*rF*[c]	1.000	.005
Z = 5.0[c]	.331	.007	(*Ad*)[c]	.665	.004
Art[c]	.664	.007	*C´*[c]	.665	.004
Cl[c]	.664	.007	*TF*[c]	.665	.004
Ge[c]	1.000	.007	*Y*[c]	.665	.004
INC2[c]	.000	.006	*YF*[c]	.665	.004
Hx[c]	.396	.006	*CONFAB*[c]	1.000	.003
(*Hd*)[c]	.799	.006	*FMap*[c]	1.000	.003
rF[c]	1.000	.005	*Hx*[c]	1.000	.003
T[c]	1.000	.005	*INC2*[c]	.000	.003
CP[c]	.666	.004	*VF*[c]	1.000	.003
Z = 6.0[c]	.666	.004	*Xy*[c]	1.000	.003
map[c]	1.000	.002	*CN*[c]	.000	.001
CONTAM[c]	.000	.001	*FQp*[c]	.000	.001
TF[c]	.000	.001	*Z* = 6.0[c]	.000	.001
Y[c]	.000	.001	*CP*[c]	—	.000
(*Ad*)[c]	—	.000	*CONTAM*[c]	—	.000
CN[c]	—	.000	*Ge*[c]	—	.000
CONFAB[c]	—	.000	*map*[c]	—	.000
FMap[c]	—	.000	*Map*[c]	—	.000
Map	—	.000	*Z* = 5.0[c]	—	.000
Z = 6.5[c]	—	.000	*Z* = 6.5[c]	—	.000
V[c]	—	.000	*T*[c]	—	.000
Xy[c]	—	.000	*V*[c]	—	.000

Note. Base rate (number of occurrences of the code/number of responses) was calculated separately for each rater and then averaged across both raters. Dashes indicate mathematically undefined statistics.

[a]*N* = 412 responses. [b]*N* = 374 responses. [c]Absent coding decision.

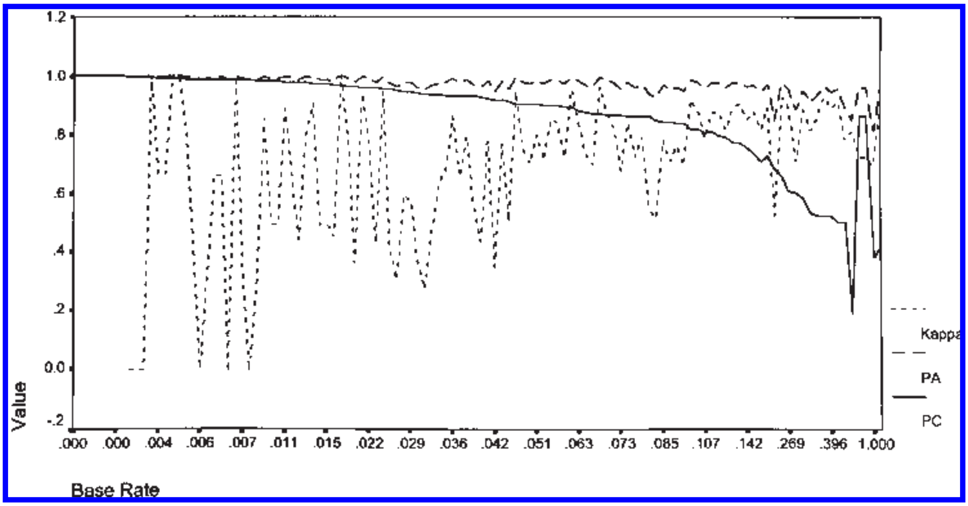

FIGURE 1 Coefficients of interrater agreement in relation to increasing base rates of Comprehensive System codes: Nonpatient sample. PA = proportion of agreement; PC = proportion of chance agreement.

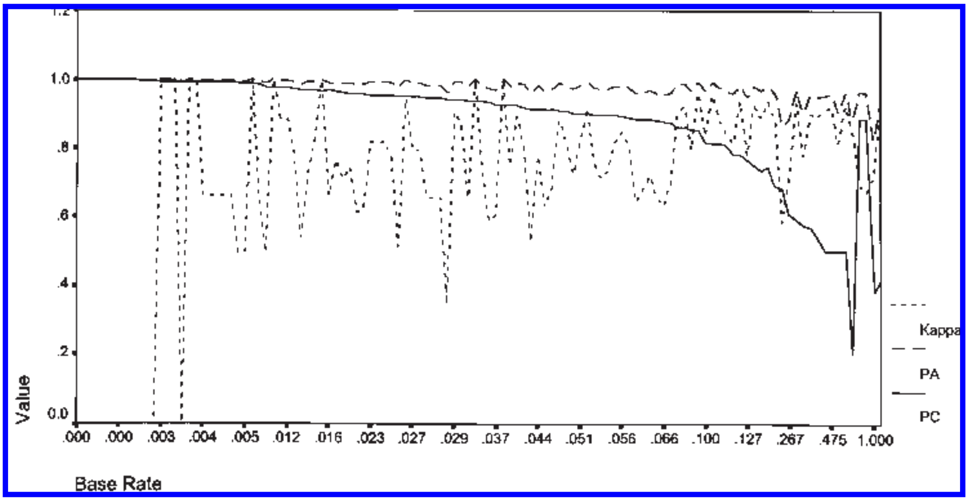

FIGURE 2 Coefficients of interrater agreement in relation to increasing base rates of Comprehensive System codes: Clinical sample. PA = proportion of agreement; PC = proportion of chance agreement.

tween the decreasing precision of the kappa statistic and the decreasing base rate of behavioral observation phenomena. Formally, kappa values for codes and coding decisions that occurred at a base rate of 5% or less demonstrated significantly lower stability than kappa values for codes and coding decisions that occurred at higher base rates: average nonpatient *SEM* = .130 versus .047, *F*(1, 106) = 57.834, *p* = .000, and average clinical *SEM* = .118 versus .049, *F*(1, 105) = 27.204, *p* = .000. These results highlight the exacting nature of reliability measurements that are made in homogenous environments (e.g., Bartko, 1991). That is, when true score variability is low, observers have a smaller range beyond chance agreement in which to agree or disagree, and classification mistakes have a much greater impact on reliability estimates than would the same number of mistakes when made under conditions of greater true score variability.

Finally, our data demonstrate that reliability estimates do not uniformly decrease in value when codes and coding decisions are assessed in increasingly homogenous environments. Specifically, of the 99 codes and coding decisions that occurred at least once in both samples (i.e., base rate > 0), 36 (36%) demonstrated higher kappa values in the lower base-rate sample. In addition, this result was equally likely to occur for codes and coding decisions with an average base rate less than or equal to 5% as it was for codes and coding decisions with an average base rate greater than 5%, suggesting that low base-rate instability was not a contributing factor. We hypothesize that increased specificity (i.e., fewer false positive diagnostic classification errors) is a likely factor contributing to the higher reliability of certain Comprehensive System codes in more homogenous samples (e.g., Langerbucher et al., 1996).

In sum, our data suggest that generalized reliability interpretation frameworks (e.g., Landis & Koch, 1977) may not be appropriate for Comprehensive System data that occur at base rates less than or equal to 5%. In these situations, it may be more appropriate to interpret the magnitude of reliability estimates in relation to the demonstrated levels of precision (i.e., the standard errors of measurement) and the maximum levels of reliability that may be achieved given the base rate, sensitivity, and specificity of Comprehensive System data (e.g., Grove et al., 1981). Furthermore, although it is true that low base-rate environments place restrictions on the adequacy of the variance estimates that are used to calculate the reliability of observational data (Bartko, 1991), no single study will be able to determine the base rate that represents the most appropriate cutoff between stringent and unstable estimates of reliability for Comprehensive System data. We suggest that the true reliability of Comprehensive System codes and coding decisions will most appropriately be established via replication and the eventual implementation of meta-analytic techniques.

Limitations of Comprehensive System Reliability

A number of Comprehensive System codes and aggregate variables demonstrate less than acceptable levels of reliability. Consistent with the behavioral assessment literature, we suggest three specific factors that may depress Comprehensive System reliability estimates. First, low prevalence will naturally depress the reliability of the majority of Comprehensive System data. For example, the average base rates of response-level codes and coding decisions in this study were 10.4% (nonpatient sample) and 10.8% (clinical sample), respectively. Only 13% (nonpatient) and 12% (clinical) of these codes and coding decisions occurred at a base rate of 20% or greater.

Second, the coding of Rorschach responses is a demanding procedure that has two mutually interdependent vulnerabilities: the adequacy of raters' administration procedures and the application of coding criteria to individual responses. Errors arising in either of these procedures will tend to lower Comprehensive System reliability estimates. Poor inquiry is a significant source of administration error that impedes subsequent coding procedures and hence lowers reliability estimates. Coding decisions are also subject to many direct sources of error, including conservative and liberal biases, errors of commission and omission, and problems with intraobserver consistency.

Ambiguous coding criteria are a third factor that tends to depress Comprehensive System reliability estimates. Texture is perhaps a good example. Texture can be scored without specific mention of a shading component if a somatosensory experience can be identified by the rater (e.g., the examinee rubs the card or states, "It looks like it would feel fluffy"). Improved coding criteria await the development of a standardized methodology for assessing Comprehensive System reliability. Although this study represents only an initial step in this endeavor, valid and increasingly potent criticisms of the Comprehensive System necessitate that Rorschach researchers no longer sacrifice statistical rigor for computational ease. Kappa and the ICC offer the greatest potential for elucidating the various sources of Comprehensive System administration and coding error and, ultimately, for enhancing the Rorschach's validity as a clinical and research measure.

Conclusions, Caveats, and Directions for Future Research

The findings reported here represent a stringent and extensive demonstration of interobserver and intraobserver reliability for the Rorschach Comprehensive System. By clearly defining our framework for conceptualizing, organizing, and interpreting Comprehensive System data, we hope that this contribution will advance future discussion and research. We believe that this study provides strong evidence for the reliability of the Rorschach Inkblot Test across multiple levels of Comprehensive System data. Furthermore, these results are consistent with conclusions drawn from the majority of previously reported reliability studies for the Rorschach Comprehensive System (e.g., Meyer, 1997b).

We note that conclusions drawn from this study must be tempered by the relatively small sample sizes that were used to evaluate the reliability of our protocol-level data. In addition, the reliability of a number of response-level codes and protocol-level variables remains problematic. Parsing the errors arising from Comprehensive System administration and coding procedures from the base-rate problem that inevitably affects the magnitude of reliability estimates is a yet uncompleted task. Finally, we provide a word of caution for reviewers who might draw conclusions from the low reliability estimates that are associated with certain Structural Summary derivations and interpretive indexes, as the majority of protocol-level data are interpreted in terms of categorical, rather than absolute, agreement. Hence, although the ICC used in this study accurately describes the reliability of derived Comprehensive System scores, the interpretive consistency of Comprehensive System protocols may need to be evaluated using an alternative assessment framework. As far as we know, the reliability of Exner's interpretive clusters (i.e., Exner, 1991) has yet to be examined.

One further and relatively unexplored area of Comprehensive System research involves the question of *scorer accuracy*, or the degree of correspondence between a rater and a criterion measure that is considered to be relatively incontrovertible (Suen, 1988). Standards of rater accuracy (i.e., master protocols) may be established through the use of videotaped administrations and consensus ratings among Comprehensive System clinicians and researchers who are deemed to be experts. Subsequent research designs may then evaluate rater accuracy by employing statistics that are relevant for assessing behavioral observation data within a validity framework, namely sensitivity, specificity, positive predictive value, and negative predictive value (Shrout et al., 1987). In addition to facilitating rater training and the development of refined coding criteria, these "gold standard" studies will help to establish acceptable levels of reliability for low base-rate Comprehensive System data, thereby enabling Rorschach researchers to confidently address the issue of field reliability that has been raised by Wood et al. (1996a, 1996b, 1997). Hence, scorer accuracy is a component of Comprehensive System validity that represents an important step toward ensuring the continued clinical and research utility of the Rorschach Inkblot Test.

ACKNOWLEDGMENTS

This research was the collaborative effort of members of the Hawaii Rorschach Research Group (Marvin W. Acklin, Claude J. McDowell II, and Mark S. Verschell). We share equal authorial credit for this article. Greg Meyer played a significant role in the development of our thinking about the nature and analysis of Rorschach data and gave generously of his expertise. Harald Janson made helpful comments concerning coding schemes and data analysis. Special thanks to Tracy Eells, who provided early inspiration and direction. Roger Bakeman, James Wood, Stephen Haynes, and Gyoen Soggin were helpful with technical assistance.

REFERENCES

Acklin, M. W., McDowell, C. M., & Ornduff, S. (1992). Statistical power and the Rorschach: 1975–1991. *Journal of Personality Assessment, 59,* 366–379.

Bakeman, R., Quera, V., McArthur, D., & Robinson, B. F. (1997). Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods, 7,* 357–370.

Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. *Psychological Bulletin, 83,* 762–765.

Bartko, J. J. (1991). Measurement and reliability: Statistical thinking considerations. *Schizophrenia Bulletin, 17,* 483–489.

Berk, R. A. (1979). Generalizability of behavioral observations: A clarification of interobserver agreement and interobserver reliability. *American Journal of Mental Disability, 83,* 460–472.

Carey, G., & Gottesman, H. (1978). Reliability and validity in binary ratings: Areas of common misunderstanding in diagnosis and symptom ratings. *Archives of General Psychiatry, 35,* 1454–1459.

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

Cone, J. D. (1982). Validity of direct observation procedures. *New Directions for Methodology of Social and Behavioral Science, 12,* 67–69.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

DeCato, C. M. (1983). Rorschach reliability: Cross-validation. *Perceptual & Motor Skills, 56,* 11–14.

DeCato, C. M. (1984). Rorschach reliability: Toward a training model for interscorer agreement. *Journal of Personality Assessment, 48,* 58–64.

DeCato, C. M. (1994). Toward a training model for scoring revisited: A follow-up on a training system for interscorer agreement. *Perceptual & Motor Skills, 78*(1), 3–10.

Exner, J. E. (1990). *A Rorschach workbook for the Comprehensive System* (3rd ed.). Asheville, NC: Rorschach Workshops.

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

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

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

Exner, J. E. (1996). A comment on "The Comprehensive System for the Rorschach: A critical examination." *Psychological Science, 7*(1), 11–13.

Exner, J. E., & Ona, N. (1995). Rorschach interpretation assistance program (Version 3.1) [Computer software]. Odessa, FL: Psychological Assessment Resources.

Exner, J. E., Jr., & Weiner, I. B. (1994). *The Rorschach: A comprehensive system: Vol. 3. Assessment of children and adolescents.* New York: Wiley.

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

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33,* 613–619.

Gelfand, S., & Hartmann, D. (1975). *Child behavior analysis and therapy.* New York: Pergamon.

Greco, C. M., & Cornell, D. G. (1992). Rorschach object relations of adolescents who committed homicide. *Journal of Personality Assessment, 59,* 574–583.

Grove, W. M., Andreasen, N. C., McDonald-Scott, P., Keller, M. B., & Shapiro, R. W. (1981). Reliability studies of psychiatric diagnosis. *Archives of General Psychiatry, 38,* 408–413.

Hanley, J. A. (1987). Standard error of the kappa statistic. *Psychological Bulletin, 102,* 315–321.

Hartmann, D. P. (1977). Considerations in the choice of reliability estimates. *Journal of Applied Behavior Analysis, 10,* 103–116.

Haynes, S. N. (1978). *Principles of behavioral assessment.* New York: Gardner.

Jensen, A. R. (1959). The reliability of projective techniques: Review of the literature. *Acta Psychologica, 16,* 108–136.

Kraemer, H. C. (1979). Ramifications of a population model for *k* as a coefficient of reliability. *Psychometrika, 44,* 461–472.

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

Langerbucher, J., Labouvie, E., & Morgenstern, J. (1996). Measuring diagnostic agreement. *Journal of Clinical and Consulting Psychology, 64,* 1285–1289.

Matarazzo, J. D. (1983). The reliability of psychiatric and psychological diagnosis. *Clinical Psychology Review, 3,* 103–145.

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.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1,* 30–46.

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

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

Mitchell, S. K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin, 86,* 376–390.

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

Perry, W., & Braff, D. L. (1994). Information-processing deficits and thought disorder in schizophrenia. *American Journal of Psychiatry, 151,* 363–368.

Perry, W., McDougall, A., & Viglione, D. J. (1995). A five-year follow-up on the temporal stability of the Ego Impairment Index. *Journal of Personality Assessment, 64,* 112–118.

Perry, W., Sprock, J., Schaible, D., McDougall, A., Minassian, A., Jenkins, M., & Braff, D. (1995). Amphetamine on Rorschach measures in normal subjects. *Journal of Personality Assessment, 64,* 456–465.

Perry, W., & Viglione, D. J. (1991). The Ego Impairment Index as predictor of outcome in melancholic depressed patients treated with tricyclic antidepressants. *Journal of Personality Assessment, 56,* 487–501.

Shrout, P. E., & Fleiss, J. (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.

Soeken, K. L., & Prescott, P. A. (1986). Issues in the use of kappa to estimate reliability. *Medical Care, 24,* 733–741.

Spitzer, R. L., Cohen, J., Fleiss, J. L., & Endicott, J. (1967). Quantification of agreement in psychiatric diagnosis: A new approach. *Archives of General Psychiatry, 17,* 83–87.

Spitzer, R. L., Endicott, J. E., & Robins, E. (1989). *Research diagnostic criteria (RDC) for a selected group of functional disorders.* New York: New York State Psychiatric Institute, Biometric Research.

Spitznagel, E. L., & Helzer, J. E. (1985). A proposed solution to the base rate problem in the kappa statistic. *Archives of General Psychiatry, 42,* 725–729.

SPSS, Inc. (1998). SPSS for Windows (Version 8.2) [Computer software]. Chicago: Author.

Suen, H. K. (1988). Agreement, reliability, accuracy, and validity: Toward a clarification. *Behavioral Assessment, 10,* 343–366.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Tinsley, H. E., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology, 22,* 358–376.

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

Weiner, I. B. (1998). *Principles of Rorschach interpretation.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Wood, J. M., Nezworski, M. T., & Stejskal, W. (1996b). Thinking critically about the Comprehensive System for the Rorschach: A reply to Exner. *Psychological Science, 7*(1), 14–17.

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

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

Marvin W. Acklin
850 West Hind Drive
Suite 203
Honolulu, HI 96821

Received April 5, 1999
Revised July 9, 1999