Viglione D; Blume-Marcovici AC; Miller HL; Giromini L; Meyer G. An Inter-Rater Reliability Study for the Rorschach Performance Assessment System. JOURNAL OF PERSONALITY ASSESSMENT. 94 (6) pp: 607-612.
DOI: 10.1080/00223891.2012.684118

An Inter-Rater Reliability Study

for the Rorschach Performance Assessment System

Donald J. Viglione, Amy C. Blume-Marcovici, Heidi L. Miller

California School of Professional Psychology, San Diego

Luciano Giromini

University of Milano-Bicocca

Gregory Meyer

University of Toledo

Abstract

In response to research demonstrating limitations in Rorschach validity and reliability, Meyer, Viglione, Mihura, Erard and Erdberg (2011) have developed a new Rorschach System, R-PAS. Based on the available research findings, this system attempts to ground the Rorschach in its evidence base, improve its normative foundation, integrate international findings, reduce examiner variability, and increase utility. As this Rorschach system is new, no reliability studies have yet been produced. The present study sought to establish inter-rater reliability for the new R-PAS. 50 Rorschach records were randomly selected from ongoing research projects using R-Optimized administration. Rorschach records were administered by 16 examiners and came from a diverse sample in terms of age, sex, ethnicity, educational background and patient status. Results demonstrated a mean intraclass correlation of .88 and median of .92. Overall, the findings indicate good to excellent inter-rater reliability for the great majority of codes and are consistent with previous findings of strong inter-rater reliability for previously researched Rorschach systems and scores.

Special thanks to those who contributed Rorschach records from their ongoing research: Greg Converse, Ryan Jordan, Vanessa Laughter, Raeanne Moore, and Tonya Oliver.

An Inter-Rater Reliability Study

for the Rorschach Performance Assessment System

In response to research demonstrating limitations in Rorschach validity and reliability, Meyer, Viglione, Mihura, Erard and Erdberg (2011) have developed the new Rorschach Performance Assessment System (R-PAS). Based on the available research findings, this system attempts to ground the Rorschach in its evidence base, improve its normative foundation, integrate international findings, reduce examiner variability, and increase utility. R-PAS variables were selected largely based on four foundations, the most important being empirical support from the published literature. The primary resource for this empirical support is a meta-analytic review of the validity of Comprehensive System (CS) (Exner, 2003) variables by Mihura, Meyer, Dumitrascu, and Bombel which has been submitted for review (2011) and is an integral part of R-PAS. In addition, R-PAS emphasizes the behavioral representation of the variable in the response process. This concept refers to the degree of performance-based support found in the Rorschach as a behavioral assessment or performance task (Foster & Cone, 1995; Viglione & Rivera, 2003). For example, an Aggressive Content (AGC) response ("a violent person") clearly involves an expression of aggression; whereas there is virtually no behavioral or response process foundation for interpreting a white space response such as anger. Utility as rated in surveys of experienced practitioners (Meyer, Hsiao, Viglione, Mihura, & Abraham, in press) parsimony, and simplicity were also considered in R-PAS development.

R-PAS also involves a new administration procedure to curtail variation in the number of responses, uses variables selected from the literature based primarily on their

evidence, provides refined guidelines for completing the inquiry or clarification phase of the test and coding responses, and relies on a new normative reference group with standard score transformations and adjustments for complexity. As this Rorschach system is new, no reliability studies have yet been produced.

Inter-rater reliability addresses the consistency or agreement of scoring across raters. Good inter-rater reliability provides a foundation for various examiners to make the same interpretation from a given Rorschach protocol. Alternatively, poor inter-rater reliability would prevent consistent interpretation across examiners, thus compromising the utility because of variability in scoring and interpretation. A great deal of research with various Rorschach systems and scores has demonstrated strong inter-rater reliability (Acklin, McDowell, Verschell, & Chan, 2000; McDowell & Acklin, 1996; Exner, 1993; Meyer & Archer, 2001; Viglione, 1999; Viglione & Meyer, 2008; Viglione & Taylor, 2003). Inter-rater reliability for the CS (Exner, 2003) has been particularly well-established, with a large-scale meta-analysis of CS inter-rater reliability reporting high inter-rater reliability for the majority of CS scores (specifically, median intraclass correlations for statistically stable scores ranged from .72-.96; Meyer et al., 2002). [1] This range of reliabilities has been considered "good to excellent" by some authors (Cicchetti, 1994; Shrout & Fliess, 1979) and "acceptable to excellent" by others (Hunsley & Mash, 2008). Similar reliabilities have been demonstrated for the Rorschach Prognostic Rating Scale, a scale derived from another method of Rorschach administration (Handler & Clemence, 2005; Meyer, 2004). Overall, these meta-analyses and research findings indicate that reasonably trained raters achieve good reliability, with average Pearson above .85, intraclass correlations (ICCs) for summary scores above .80, and average

kappa values for codes assigned to each response above .80. Earlier doubts about CS reliability (Wood, 1996) were based on the criticism that CS inter-rater reliability was determined using percent agreement without correcting for chance. However, as noted, inter-rater reliability has been demonstrated with chance-corrected statistics, including ICC, kappa and Iota coefficients, as the most appropriate and precise statistical methods, proving early criticisms to be unfounded. Indeed, inter-rater reliability for the majority of Rorschach scores compare favorably to other published meta-analyses of inter-rater reliability in psychology, psychiatry, and medicine (Meyer, 2004). Given the wide variety of scores, scales, research projects, and systems from which good reliability has been demonstrated, one must conclude that well-trained coders should achieve acceptable, good, and often excellent inter-rater reliability for the great variety of Rorschach scores.

As demonstrated by Weiner (2003), the Rorschach can be considered to be a method of generating data relevant to personality and information processing. From this perspective, various scores and scoring methods systematize the data produced during Rorschach administration, thus constituting the Rorschach as a test. As shown by the strong reliability data across different types of systems, scores, countries, and languages, this test has produced consistently strong reliability. The R-PAS includes many variables that were also used in the CS, clarifies and specifies coding instructions, and modifies a few (e.g. Sex content) to be more consistent with their interpretation. Thus, the above-reported research findings suggest that inter-rater reliability for these R-PAS variables should be strong.

R-PAS also includes variables not used in the CS (Table 1 includes a list of R-PAS variables), including Complexity, Space Integration (SI) and Space Reversal (SR),

the Rorschach Oral Dependency Scale (ROD) (Bornstein & Masling, 2005; Masling, Rabie & Blondhiem, 1967), now called Oral Dependency Language (ODL), the Mutuality of Autonomy (MA) Scale (Urist, 1977), previously abbreviated as MOA, the Ego Impairment Index (EII; Perry & Viglione, 1991), and Aggressive Content (Gacono & Meloy, 1994). Inter-rater reliability for many of these specific variables has been reported in the literature, with good results. The ODL has been shown to have strong validity and reliability, with Pearson correlation coefficients typically greater than .90 (Bornstein, Rossner, & Hill, 1994; Juni, Masling & Brannon, 1979; O'Neill & Bornstein, 1990) and kappa coefficients greater than .80 (Duberstein & Talbot, 1993; Greenberg & Bornstein, 1989; O'Neill & Bornstein, 1990; O'Neill & Bornstein, 1991) in both clinical and nonclinical samples. Meyer (2004) conducted a meta-analysis of studies examining inter-rater reliability for the ROD utilizing *r*, kappa, and ICC results from 31 studies and 40 samples. They found good inter-rater reliability of this construct with a Pearson's *r* of 0.91 and kappa coefficients and ICCS of 0.91. Inter-rater reliability on MA scales, which assess object relations, was initially reported as relatively lower (Urist, 1977) due to the use of percent agreement rather than correlating items on a dimensional scale. However, after scoring guidelines were improved, a meta-analysis of MA Scale inter-rater reliability found a weighted kappa coefficient of .83, a percent agreement coefficient of .81, an ICC of .94 and Pearson's *r*= .91 (Bombel, 2006 as cited in Bombel, Mihura & Meyer, 2009). The EII, a composite score developed to assess level of ego impairment, was demonstrated to be reliably scored with ICCs of component variables in excess of .90 (Perry & Viglione, 1991). Research has shown excellent reliability for Aggressive Content scores with Kappa coefficients and ICCs ranging from .86 to .94 (Gacono,

Gacono, Meloy, & Baity, 2008). As such, good to excellent reliability findings from multiple methods of test administration and types of coding over many years, from a variety of researchers, in clinical and nonclinical settings, and conducted in different languages, leads to some confidence that the strong inter-rater reliability from the past will carry over to the R-PAS.

However, prior research studies on inter-rater reliability, as well as their implications for R-PAS, are not without limitations. Several studies reveal that reliabilities for low base rate (i.e., rarely occurring) scores are inconsistent across studies (e.g., Acklin, McDowell, Verschell, & Chan, 2000; Meyer et al., 2002; Viglione & Taylor, 2001). Roughly speaking, low base rate variables occur less than once per record (e.g., Reflections, Vista), so that large samples are needed to accurately estimate their reliability.

In addition, some codes have demonstrated lower reliability and thus appear to be more challenging to code accurately. Viglione and Meyer (2008) summarized the relevant literature and identified some response level codes and distinctions that are subject to such variable inter-rater reliability. These include Developmental Quality vague and vague/synthesis, Form versus Color, various shading subtypes, Form Quality particularly Unusual, Cognitive Special Score subtypes, and Cognitive Score Level 1 versus Level 2. Special care must be taken to develop coding expertise for these variables. It should be pointed out that the procedures and guidelines for these variables in the R-PAS manual are more extensive than the guidelines previously available and, it is hoped, may result in improved reliability for these distinctions.

The present study sought to establish inter-rater reliability for the R-PAS. This is the first full report of R-PAS inter-rater reliability.

Method

Rorschach records were selected from ongoing Rorschach research and archived clinical files. Specifically, records were drawn from five dissertation projects, homework assignments in a personality assessment graduate course, and from the clinical files of the primary investigator. Records were selected randomly. The only selection criteria were a legible verbatim record and a location sheet. All records were administered according to the new R-Optimized administration procedures outlined in the draft versions of the R-PAS manual. R-Optimized administration closely resembles CS administration but provides extra guidance to limit the variability of the number of responses, R. Examiners advise respondents when introducing the test to give "two,…or maybe three" responses per card. If only one response is given to any card, the examiner encourages the respondent to give more (Prompt). If four responses are produced to any card, the examiner politely requests the respondent to return the card (Pull), so that there is a maximum of four responses per card, and reminds the respondent to give "two,…or maybe three" responses.

A total of 50 Rorschach protocols were collected for the present study. There were six children (12% of the sample), all White, five girls and one boy. Age ranged from 8 to 13 years (M = 11.2; SD = 1.9). The children were non-patient, community children who volunteered for practice evaluations conducted by six graduate students in a performance assessment course, which included the Rorschach test.

Among the 44 adults, 29 (58% of the sample) were collected for research purposes (11 non-patients, 7 college students, 5 outpatients with diagnosis of schizophrenia, and 6 outpatients with other diagnoses), 8 (16% of the sample) were forensic cases referred for evaluation (among these, 7 were referred for sex offender evaluation), and 7 (14% of the sample) were clinical patients referred for evaluation. Age ranged from 18 to 67, the mean age being 35.1 (SD = 14.1). In terms of ethnicity of the sample, 23 (52%) were White, 10 (23%) Latino/a, 7 (16%) Asian/Pacific American, and 4 (9%) African American. Finally, the 44 adults in the sample had a range of educational backgrounds, characterized by 3 individuals with some high school, 18 individuals with a high school degree, 11 individuals with some college, 7 with a college degree, 2 with some graduate school, 2 with a graduate degree, and one unknown. The adult records were collected by 16 different advanced graduate students conducting dissertation research or working in clinical contexts. Among them, 9 examiners were female. Each examiner administered 1 to 10 records (M=2.8). Thus, this adult and child sample was diverse in terms of age, race, educational background, non-patient and patient status, source, and examiners; characteristics that encompass the many applications of the test in practice and increase generalizability of the results. The rationale for including such a diverse sample was to mirror the diverse populations with whom the Rorschach is used in practice.

Blume-Marcovici and Miller, two graduate students, each independently recoded these records using draft versions of the coding chapters in the R-PAS manual (Meyer et al., 2011). These versions were nearly identical to the final manual particularly in terms of critical guidelines and definitions of codes; however, the final manual was “emerging”

in a series of drafts as the study was being conducted. This required the coders to re-code or code anew several variables during the study to utilize the most up-to-date coding guidelines available. It should also be pointed out that the FQ coding used in this study was not the final R-PAS version, yet nearly identical to R-PAS in terms of the entries and organization found in the table. Blume-Marcovici and Miller coded each record independently and were blind to each other's coding. In some cases, the original coding of the records was available to the first coder but not the other. Both coders were trained in the R-PAS method by Donald Viglione, and served as teaching assistants in Viglione's Performance Assessment graduate-level course, where they received supervision for R-PAS scoring in over 50 records.

Results and Discussion

The inter-rater reliability results of this study are presented in Tables 1, 2 and 3 for the 60 interpreted variables in R-PAS. Two-way random effects model single measures intraclass correlation coefficient (ICC) was used. Table 1 is presented in the order used in the interpretive output for R-PAS, such that the R-PAS Summary Page 1 variables appear first and Summary Page 2 variables afterward. Page 1 variables are strong in terms of research support and response process or behavioral foundation. Page 2 variables have less support. As found in the R-PAS Summary, variables are grouped under five domains within each Summary Page: Administration Behaviors and Observations, Engagement and Cognitive Processing, Perception and Thinking Problems, Stress and Distress, and Self and Other Representation. Reliability for variables in each domain are reported.

Table 2 provides summary statistics for the ICC reliability coefficients. The variables' Mean ICC is .88 and median is .92. Overall, the findings indicate good to excellent (Cicchetti, 1994; Shrout & Fliess, 1979) or acceptable to excellent (Hunsley & Mash, 2008) inter-rater reliability for the great majority of codes. These inter-reliabilities are similar to the adequate and good reliabilities produced consistently with the CS (Acklin, McDowell, Verschell, & Chan, 2000; Meyer et al., 2002; Viglione & Taylor, 2001). A total of 32% of the ICCs were exceedingly high at .95 or above. As predicted, strong evidence of reliability for Rorschach's administered in different methods generalizes to R-PAS.

Table 3 presents descriptive statistics and base rates for Rorschach scores in the present sample. The tables distinguish between rare variables (base rates of less than one occurrence per protocol), infrequent (between one and two occurrences per protocol), and common base rates (two or more). The relationship between lower base rates and lower ICCs is not as strong as has been in previous inter-rater reliability studies. However, the two lowest ICCs, for Vista (ICC = .44) and Vague (ICC = .54), accompany low base rate variables, as do five of the eight lowest ICCs. Indeed, Vista, which has the lowest ICC in the study also has the lowest base rate (.14). As noted by Acklin, McDowell II, Verschell, and Chan (2000), reliability significantly decreases in value with low base rate variables. Vista and Vague, as well as FQu%, the third lowest ICC, have been found to produce variable interrater reliabilities in other studies (Viglione & Meyer, 2008). Confidence in the reliability of these variables awaits further study. Thus, patterns found in previous studies recur here.

Two explanations, one certain and statistical and another speculative and procedural, might explain lower reliabilities with infrequent codes. Infrequent codes, by definition, produce few occurrences or observations. Statistically, this results in more random error relevant to the effect being studied, so that reliability statistics are less accurate. Larger samples are needed to reduce measurement error so as to provide more accurate and stable estimates of the reliability of infrequent codes. A second possible experiential or procedural explanation is that coders spend less time practicing the coding of infrequent variables which may result in less proficiency with coding these variables. In addition, coders might occasionally overlook such variables if they are not showing up frequently.

There are limitations associated with this study. First of all, the coders were both students from the same lab supervised by one of the R-PAS developers. Thus, their level of agreement might be greater than that derived in across-site comparisons or between coders with different training and mentors. From that perspective, the ICCs might be greater than one would find in the field. On the other hand, these coders and all the examiners learned R-PAS as it emerged. The R-PAS manual had not been finalized so they used draft forms of the guidelines and instructions. As described previously, with the exception of FQ which were nearly identical to that in the final manual, the coding guidelines used for the present study are the same as those in the final manuscript; however, they were “emerging” in a series of drafts as the study was being conducted, adding additional clarification to the coding of some variables as the present study progressed.

It should also be noted that utilizing two coders may limit the generalizability of these reliability findings. However, using only two coders is a common procedure with many psychological tests (e.g., Sivik & Hösterey, 1992; Woloszyn, Murphy, Wetzel & Fisher, 1993) including the Rorschach (e.g., Meyer, Hilsenroth, Baxter, Exner, Fowler, Piers & Resnick, 2002; Viglione & Taylor, 2003) as well as other scientific fields, such as medicine (e.g., Hao, Wong & Kwan, 2011; Fischer, Haley, Saarinen & Chretien, 2011; see also recommendations of Walter, Eliasziw, & Donner, 1998). Moreover, Giraudeau and Mary (2001) demonstrated that two or three replicates (or coders) per respondent is of no influence on precision of ICC estimates.

Another study limitation is that six of the study protocols were administered by students in a graduate course on personality assessment. Thus, their inquiry was likely less refined than those of clinicians using the Rorschach in practice. Finally, although 50 records might be ample for high base rate variables with relatively normal distributions (e.g., MC, FQ-%, D), it might be insufficient to produce stable or accurate estimates of low base rate variables with many zero values. Indeed, those variables with the lowest reliabilities in the present study (e.g., V, Vg, m, MAP, CBlend, Dd%) are low base rate variables. The one exception is FQu%, a variable that has demonstrated weak reliability in the past. For these and other reasons, additional research with larger samples by others not associated with the R-PAS systemetizers is needed to confirm its inter-rater reliability.

References

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

Bombel, G. A. (2006). *A meta-analysis of interrater scoring reliability for the Rorschach Mutuality of Autonomy (MOA) Scale*. Unpublished master's thesis, University of Toledo, Toledo, Ohio.

Bombel, G. A., Mihura, J. L., & Meyer, G. J. (2009). An examination of the construct reliability of the Rorschach Mutuality of Autonomy (MOA) Scale. *Journal of Personality Assessment, 91*(3), 227-237.

Bornstein, R. F., & Masling, J. M. (2005). The Rorschach Oral Dependency scale. In R. F. Bornstein & J. M. Masling (Eds.), Scoring the Rorschach: Seven validated systems (pp. 135-157). Mahwah, NJ: Erlbaum.

Bornstein, R. F, Rossner, S. C., & Hill, E. R. (1994). Retest reliability of scores on objective and projective measures of dependency: Relationship to life events and interest interval. *Journal of Personality Assessment, 62,* 398-415.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284-290.

Duberstein, P. R., & Talbot, N. L. (1993). Rorschach oral imagery, attachment style, and interpersonal relatedness. *Journal of Personality Assessment, 61*, 294-310.

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

Exner, J.E. (2003). *The Rorschach: A Comprehensive System: Volume 1: Basic foundations and principles of interpretation* (4th ed.). Hoboken, NJ: John Wiley & Sons, Inc..

Fischer, M., A., Haley, H., Saarinen, C., L., & Chretien, K., C. (2011). Comparison of blogged and written reflections in two medicine clerkships. Medical Education, 45(2), 166-175.

Foster, S. L., & Cone, J. D. (1995). Validity issues in clinical assessment. *Psychological Assessment, 7,* 248-260.

Gacono; C. B., Gacono, L. A., Meloy, J. R., Baity, M.R. (2008). The Rorschach assessment of aggression: The Rorschach extended aggression scores. In C.B Gacono & F.B. Evans with N. Kaser-Boyd (Eds.) *Handbook of forensic Rorschach psychology* (pp. 22-54). Matwah, NJ: Lawrence Erlbaum Associates.

Gacono, C. B., & Meloy, J. R. (1994). The Rorschach assessment of aggressive and psychopathic personalities. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Giraudeau, B., & Mary, J., Y. (2001). Planning a reproducible study: How many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient. Statistics in Medicine, 20, 3205-3214.

Greenberg, R. P., & Bornstein, R. F. (1989). Length of psychiatric hospitalization and oral dependency. *Journal of Personality Disorders, 3,* 199-204.

Handler, L., & Clemence, A. (2005) Klopfer's Rorschach Prognostic Rating Scale: Reliability and Validity. In J. Masling and R. Bornstein, (Eds.), *Scoring the Rorschach: Seven validated systems* (pp. 25-54). Washington, DC: APA.

Hao, X., Wong, I., S., M., & Kwan, P. (2011). Interrater reliability of the international consensus definition of drug-resistant epilepsy: A pilot study. Epilepsy & Behavior, 22(2), 388-390.

Hunsley, J., & Mash, E.J. (2008). *A guide to assessments that work*. New York, New York: Oxford University Press.

Juni, S., Masling, J., & Brannon, R. (1979). Interpersonal touch and orality. *Journal of Personality Assessment, 45,* 235-237.

Masling, J.M., Rabie, L., & Blondheim, S.H. (1967). Obesity, level of aspiration, and Rorschach and TAT measures of oral dependence. *Journal of Consulting Psychology, 31,* 233–239.

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.

Meyer, G. J. (2004). The reliability and validity of the Rorschach and TAT compared to other psychological and medical procedures: An analysis of systematically gathered evidence. In M. Hilsenroth & D. Segal (Eds.), Personality assessment. Volume 2 in M. Hersen (Ed.-in-Chief), *Comprehensive handbook of psychological assessment* (pp. 315-342). Hoboken, NJ: John Wiley & Sons.

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

Meyer, G.J., & Archer, R.P. (2001). The hard science of Rorschach research: What do we know and where do we go? *Psychological Assessment, 13,* 486-502.

Meyer, G. J., Hilsenroth, M. J., Baxter, D., Exner, J. E., Jr., Fowler, J. C., Piers, C. C., & Resnick, J. (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., Mihura, J.L., Erard, & Erdberg (2011). A Manual for the Rorschach Performance Assessment System. Toledo, OH: R-PAS.

Meyer, Hsiao, Viglione, Mihura, & Abraham. (in press) Survey of clinical experience with the Rorschach. *Journal of Personality Assessment.*

Mihura, J. L., Meyer, G. J., Dumitrascu, N., & Bombel, G. (2011). A systematic review and meta-analysis of the Rorschach Comprehensive System validity literature. Submitted for review for publication.

O'Neill, R. M., & Bornstein, R. F. (1990). Oral dependence and gender: Factors in help-seeking response set and self-reported psychopathology in psychiatric inpatients. *Journal of Personality Assessment, 55,* 28-40.

O'Neill, R. M., & Bornstein, R. F. (1991). Orality and depression in psychiatric inpatients. *Journal of Personality Disorders, 5,* 1-7.

Perry, W., Viglione, D.J. (1991). The ego impairment index as a predictor of outcome in melancholic depressed patients treated with tricyclic antidepressants. *Journal of Personality Assessment, 56(3),* 487-501.

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

Sivik, T. M., & Hösterey, U. (1992). The Thematic Apperception Test as an aid in understanding the psychodynamics of development of chronic idiopathic pain syndrome. Psychotherapy and Psychosomatics, 57(1-2), 57-60.

Urist, J. (1977). The Rorschach Test and the assessment of object relations. *Journal of Personality Assessment, 41,* 3–9.

Viglione, D. J. (1999). A review of recent research addressing the utility of the Rorschach. *Psychological Assessment, 11,* 251-265.

Viglione, D.J., & Meyer, G.J. (2008). An Overview of Rorschach psychometrics for forensic practice. In C. B. Gacono & F. B. Evans with N. Kaser-Boyd & L. A. Gacono (Eds.), *Handbook of forensic Rorschach psychology* (pp. 21-53). Mahwah, NJ: Lawrence Erlbaum Associates.

Viglione, D.J., & Rivera, B. (2003). Assessing personality and psychopathology with projective tests. In J. R. Graham & J. A. Naglieri (Eds.), *Comprehensive Handbook of Psychology: Assessment Psychology* (Vol. 10, pp. 531-553). New York: Wiley.

Viglione, D.J., & Taylor, N. (2003). Empirical support for interrater reliability of the Rorschach Comprehensive System coding. *Journal of Clinical Psychology, 59(1),* 111-121.

Walter, S., D., Eliasziw, M., Donner, A. (1998). Sample size and optimal designs for reliability studies. Statistics in Medicine, 17, 101-110.

Weiner, I. B. (2003). *Principles of Rorschach Interpretation* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum and Associates.

Woloszyn, D., B., Murphy, S., G., Wetzel, L. & Fisher, W. (1993). Interrater agreement on the Wechsler Memory Scale-Revised in a mixed clinical population. Clinical Neuropsychologist, 7(4), 467-471.

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

Table 1
Inter-Rater Reliabilities for R-PAS Summary Variables on Page 1 and Page 2

Variable	ICC	Description	Base Rate[a]
		Summary Page 1	
		Administration Behaviors & Observations	
Pr	0.95	Prompts	Rare
Pu	0.94	Pulls	Rare
CT	1.00	Card Turns; total number of responses in which the card was turned, regardless of final orientation for the response	Rare
		Engagement & Cognitive Processing	
Complexity	0.99	A composite that quantifies the amount of differentiation and integration in a protocol	Common
R	1.00	Number of Responses	Common
F%	0.98	Form Percent, F/R	Common
Blend	0.92	Blend response; two or more determinants	Common
Sy	0.96	Synthesis response	Common
MC	0.97	Sum of Human Movement (M) and the weighted sum of Color responses (WSumC = (0.5*FC) + CF + (1.5*C))	Common
MC - PPD	0.86	Subtract Potentially Problematic Determinants	Common

		(PPD = shading + achromatic color + inanimate movement) from MC	
M	0.97	Human Movement	Common
M/MC	0.92	M divided by the sum of M and WSumC	Common
(CF+C)/SumC	0.72	CF+C divided by FC+CF+C	Infrequent
		Perception & Thinking Problems	
EII-3	0.94	Ego Impairment Index-3rd Revised for R-PAS	Common
TP-Comp	0.91	Thought and Perception Composite (a dimensional version of the CS PTI)	Common
WSumCog	0.98	Weighted Sum of Cognitive Codes	Common
SevCog	0.93	Sum of Severe Cognitive Codes (Level 2 codes + Peculiar/Inappropriate Logic + Contaminations)	Rare
FQ-%	0.81	% of all responses that are distorted	Common
WD-%	0.81	% of W and D responses with FQ-	Common
FQo%	0.84	% of all responses that are common and accurate.	Common
P	0.89	Popular response	Common
		Stress & Distress	
m	0.69	Inanimate Movement	Infrequent
Y	0.86	Diffuse Shading	Rare
MOR	0.93	Morbid Content	Infrequent
SC-Comp	0.83	Suicide Concern Composite (a dimensional version of the CS Suicide Constellation)	Common
		Self & Other Representation	

ODL%	0.94	Oral Dependency Language	Common
SR	0.91	Space Reversal; the focal object is seen in the white space so that figure and background are perceptually reversed	Rare
MAP/MAHP	0.90	The Mutuality of Autonomy Pathology (Levels 5, 6 and 7) as a proportion of the sum of MAP and Mutuality of Autonomy Health (Level 1)	Rare
PHR/GPHR	0.93	Poor Human Representation as a proportion of all human representational responses	Common
M-	0.92	Human Movement with distorted form	Rare
AGC	0.79	Aggressive Content	Common
V-Comp	0.97	Vigilance Composite (a dimensional version of the CS Hypervigilance Index)	Common
H	0.96	Whole Human content	Common
COP	0.94	Cooperative Movement	Infrequent
MAH	0.86	Mutuality of Autonomy-Health (Level 1)	Infrequent

Summary Page 2

Engagement & Cognitive Processing

W%	0.99	% of all responses that are whole location	Common
Dd%	0.88	% of all responses that are detail location	Common
SI	0.86	Space Integration response; space and ink are included in location	Common

IntCont	0.94	Intellectualized Content	Infrequent
Vg%	0.54	% of all responses that are coded Vague	Rare
V	0.44	Vista, where shading creates a sense of dimensionality	Rare
FD	0.66	Form Dimension , where form creates a sense of dimensionality	Rare
R8910%	0.98	% of all responses which occur on Cards VIII, IX, & X	Common
WSumC	0.93	Weighted Sum of Color determinants	Common
C	0.78	Pure Color	Rare
Mp/(Ma+Mp)	0.88	Passive Human Movement (Mp) divided by sum of Mp and Active Human Movement (Ma)	Infrequent
		Perception & Thinking Problems	
FQu%	0.64	Percentage of all responses that are of intermediate accuracy and frequency.	Common
		Stress & Distress	
PPD	0.92	Potentially Problematic Determinants, FM+m+Y+T+V+C'	Common
YTVC'	0.95	Total number of shading (Y, T, V) and achromatic color (C') determinants)	Common
CBlend	0.76	Color (FC, CF, C) occurs with shading (Y, T, V) or achromatic color (C’)	Rare
C’	0.95	Achromatic color or white	Infrequent

CritCont%	0.90	Critical Contents divided by R	Common
		Self & Other Representation	
SumH	0.97	Sum of all the Human content codes, H+(H)+Hd+(Hd)	Common
NPH/SumH	0.94	The Non-Pure Human divided by Sum H	Common
r	0.97	Reflection determinant	Rare
p/(a+p)	0.83	The Passive Movement (p) divided by the sum of active movement (a) and p	Common
AGM	0.89	Aggressive Movement	Rare
T	0.86	Texture determinant	Rare
PER	0.90	Personal Knowledge Justification, the use of personal experience to explain or justify a response.	Rare
An	0.92	Anatomy content.	Infrequent

[a] Base rates have been included to demonstrate the frequency with which each variable appeared in the records used for this study. When the mean frequency of the variable is lower than 1 the base rate is considered “Rare”, when it is between 1 and 2 the base rate is considered “Infrequent”, and when it is greater than 2 the base rate is considered “Common”. For proportion scores the mean frequency of the numerator variable was used; for differences or composite scores the mean frequency of the sum of the variables was used. Actual numerical values for these base rates are presented in Table 3.

Table 2

Summary of ICC Inter-Rater Reliability Results for 60 R-PAS variables (N = 50)

Mean	0.88
SD	0.11
Minimum	0.44
25th percentile	0.86
Median	0.92
75th percentile	0.95
Maximum	1.00
# of Poor ICCs $< .40$[a]	0
# of Fair ICC .40–.59	2 (3%)
# of Good ICC .60–.74	4 (7%)
# of Excellent ICC $\geq .75$	54 (90%)
Mean ICC for 9 low base rate, rare variables	.87
Mean ICC for 17 moderately low base rate, infrequent variables	.84
Mean ICC for 34 common base rate variables	.91

[a] The characterization of the ranges of the reliability coefficients is derived from Cicchetti (1994) and Shrout and Fliess (1979).

Table 3

Descriptive Statistics and Base Rates

Summary Page 1

Administration Behaviors & Observations

	Rater 1			*Rater 2*			
Variable	N	M	SD	N	M	SD	Base Rate[a]
Pr	50	.90	1.39	50	.90	1.54	.90
Pu	50	.32	.82	50	.34	.98	.33
CT	50	.88	1.32	50	.88	1.32	.88

Engagement & Cognitive Processing

	Rater 1			*Rater 2*			
Variable	N	M	SD	N	M	SD	Base Rate[a]
Complexity	50	70.68	22.79	50	70.86	23.04	> 25
R	50	24.58	5.37	50	24.58	5.37	24.58
F%	50	49.06	19.13	50	48.56	19.38	12.06
Blend	50	2.76	2.15	50	3.16	2.53	2.96
Sy	50	6.48	3.92	50	6.30	4.10	6.39
MC	50	6.21	3.53	50	6.41	3.68	6.31
MC - PPD	50	-1.21	3.34	50	-1.23	3.58	13.84
M	50	3.68	2.65	50	3.60	2.77	3.64
M/MC	50	.59	.29	50	.56	.29	3.64
(CF+C)/SumC	42	.48	.33	43	.57	.36	1.70

Perception & Thinking Problems

	Rater 1			*Rater 2*			
Variable	N	M	SD	N	M	SD	Base Rate[a]
EII-3	50	-.03	1.07	50	-.04	1.1	> 25
TP-Comp	50	.72	1.15	50	.71	1.08	> 25
WSumCog	50	8.66	16.73	50	8.06	15.28	8.36
SevCog	50	.50	1.61	50	.50	2.04	.50
FQ-%	50	21.57	10.14	50	21.72	9.66	5.39
WD-%	50	17.52	10.67	50	17.57	9.19	3.83
FQo%	50	48.90	11.75	50	49.14	13.15	11.92
P	50	5.50	2.22	50	5.52	2.03	5.51

Stress & Distress

	Rater 1			*Rater 2*			
Variable	N	M	SD	N	M	SD	Base Rate[a]
m	50	1.26	1.27	50	1.28	1.34	1.27
Y	50	.90	1.28	50	.98	1.33	.94
MOR	50	1.18	1.19	50	1.18	1.21	1.18
SC-Comp	50	4.24	1.03	50	4.45	1.06	> 25

Self & Other Representation

	Rater 1			*Rater 2*			
Variable	N	M	SD	N	M	SD	Base Rate[a]
ODL%	50	10.46	9.23	50	10.14	9.42	2.46

SR	50	.94	1.19	50	.84	1.04	.89
MAP/MAHP	42	.33	.37	37	.27	.33	.64
PHR/GPHR	50	.48	.21	50	.48	.22	3.33
M-	50	.80	1.03	50	.74	1.07	.77
AGC	50	3.12	1.93	50	2.80	2.12	2.96
V-Comp	50	3.42	1.38	50	3.35	1.37	> 25
H	50	2.00	1.53	50	2.08	1.66	2.04
COP	50	1.34	1.35	50	1.38	1.41	1.36
MAH	50	1.40	1.32	50	1.22	1.22	1.31

Summary Page 2

Engagement & Cognitive Processing

	Rater 1			***Rater 2***			
Variable	N	M	SD	N	M	SD	Base Rate[a]
W%	50	45.48	20.57	50	45.04	20.95	10.82
Dd%	50	11.74	8.48	50	12.35	8.98	3.10
SI	50	2.08	1.55	50	2.30	1.62	2.19
IntCont	50	1.70	2.71	50	2.02	2.87	1.86
Vg%	50	3.02	4.31	50	5.20	6.89	.97
V	50	.14	.35	50	.14	.40	.14
FD	50	.64	.94	50	.88	1.24	.76
R8910%	50	30.36	3.88	50	30.26	3.98	7.49
WSumC	50	2.53	2.06	50	2.81	2.12	2.67
C	50	.28	.61	50	.32	.59	.30

Mp/(Ma+Mp)	47	.52	.35	46	.51	.35	1.78

Perception & Thinking Problems

	Rater 1			***Rater 2***			
Variable	N	M	SD	N	M	SD	Base Rate[a]
FQu%	50	28.50	10.53	50	27.86	10.38	7.01

Stress & Distress

	Rater 1			***Rater 2***			
Variable	N	M	SD	N	M	SD	Base Rate[a]
PPD	50	7.42	3.62	50	7.64	3.71	7.53
YTVC'	50	2.84	2.23	50	3.08	2.50	2.96
CBlend	50	.36	.66	50	.38	.70	.37
C’	50	1.36	1.64	50	1.48	1.79	1.42
CritCont%	50	17.16	11.32	50	18.53	12.56	4.34

Self & Other Representation

	Rater 1			***Rater 2***			
Variable	N	M	SD	N	M	SD	Base Rate[a]
SumH	50	6.06	3.11	50	6.06	3.03	6.06
NPH/SumH	50	67.32	20.57	50	66.45	21.77	4.02
r	50	.66	1.12	50	.70	1.11	.68
p/(a+p)	50	.48	.27	50	.45	.26	3.66
AGM	50	.62	1.07	50	.68	1.10	.65
T	50	.44	.76	50	.48	.74	.46

PER	50	.98	1.36	50	.88	1.26	.93
An	50	1.16	1.23	50	1.28	1.21	1.22

[a] Base rates were computed as the mean frequency of the variable. For proportion scores the mean frequency of the numerator variable was used; for difference or composite scores the mean frequency of the sum of the variables was used.