Psychoanalytic Psychology

Rorschach Assessment of Ego Involvement in Affect Regulation: Interrater Reliability of Form Dominance in Shading and Achromatic Color Responses

Anthony D. Bram, Donald J. Viglione, Oren Lee-Parritz, Kiley A. Gottschalk, Jed A. Yalof, James H. Kleiger, Ksera Dyette, and Ali Khadivi

Online First Publication, April 13, 2023. https://dx.doi.org/10.1037/pap0000458

Psychoanalytic Psychology

ISSN: 0736-9735

https://doi.org/10.1037/pap0000458

Rorschach Assessment of Ego Involvement in Affect Regulation: Interrater Reliability of Form Dominance in Shading and Achromatic Color Responses

Anthony D. Bram, PhD[1, 2], Donald J. Viglione, PhD[3], Oren Lee-Parritz, PsyD[4], Kiley A. Gottschalk, PsyD[5], Jed A. Yalof, PsyD[6], James H. Kleiger, PsyD[7], Ksera Dyette, PsyD[8], and Ali Khadivi, PhD[9]

[1] Department of Psychiatry, Cambridge Health Alliance, Harvard Medical School
[2] Boston Psychoanalytic Society and Institute, Newton, Massachusetts, United States
[3] Department of Psychology, Alliant International University, San Diego
[4] LifeStance Health, Boston, Massachusetts, United States
[5] Massachusetts Institute for Psychoanalysis, Brookline, Massachusetts, United States
[6] Austen Riggs Center, Stockbridge, Massachusetts, United States
[7] Private Practice, Bethesda, Maryland, United States
[8] Court Clinic Operations Team, William James College
[9] Department of Psychiatry, Albert Einstein College of Medicine

Although emphasis on psychodynamic thinking has waned in assessment training, the ascendant Rorschach Performance Assessment System (R-PAS; Meyer et al., 2011) has reintegrated psychoanalytic concepts into empirical Rorschach assessment: R-PAS adds scores involving object relations, implicit dependency, aggressive ideation, and ego impairment. R-PAS has, however, excluded the psychodynamic framework for assessing ego involvement in the regulation of anxiety/dysphoria by eliminating the coding of Form Dominance in Shading and Achromatic Color (FDSHAC) that has been part of the Comprehensive System (Exner, 2003). This decision was based in part on concerns about efficiently and reliably coding distinctions among Form Dominant, Form Secondary, and Formless levels of FDSHAC. To establish that such distinctions can be coded reliably, we applied supplemental guidelines (Viglione, 2010) to evaluate reliability among four experienced assessors who coded determinants for 155 Rorschach responses, 115 of which required FDSHAC determination. Applying Gwet's AC2′s to ordinal scales, interrater reliabilities were good to excellent. Reliabilities were strongest for Form Dominance in Texture and Achromatic Color, modestly so for Form Dominance in Diffuse Shading, and problematic for Form Dominance of Vista. Among levels of Form involvement across FDSHAC variables, raters had the most difficulty distinguishing Form Secondary. We discuss considerations for clinical coding, psychodynamic configurational analyses for interpretation, and construct validation research.

Keywords: rorschach, ego involvement, affect regulation, form dominance

Supplemental materials: https://doi.org/10.1037/pap0000458.supp

Although the past 2 decades have seen training both in psychodynamic thought and in depth-oriented assessment diminish in the graduate training of psychologists (Bram et al., 2018; Piotrowski, 2015), recently there have been encouraging developments enlivening a reintegration of psychoanalytic concepts into empirically based Rorschach assessment (Bram & Yalof, 2018). Notably, the latest approach to Rorschach administration, scoring, and interpretation—the Rorschach Performance Assessment System (R-PAS; Meyer et al., 2011)—has added four scores based on psychodynamic concepts and research. These new empirically supported scores were/are not part of methodology that had previously dominated Rorschach assessment for the previous 40-plus years, the Rorschach Comprehensive System (CS; Exner, 2003). These psychodynamically relevant scores involve (a) implicit dependency (Oral Dependent Language [ODL]), (b) object relations (Mutuality of Autonomy Health or Pathology [MAH/MAHP]), (c) aggressive preoccupations (Aggressive Content); and (d) vulnerability in ego functioning (Ego Impairment Index [EII-3]). As R-PAS gains popularity among clinical practitioners and in training programs that still value depth-oriented assessment (Viglione et al., 2022; Villanueva van den Hurk et al., 2021), the inclusion of these new scores means that these constructs are now more routinely considered.

Anthony D. Bram, PhD https://orcid.org/0000-0002-5178-1666

The authors thank the American Psychoanalytic Association's Fund for Psychoanalytic Research for funding our project (Grant to Anthony D. Bram, PhD). They also express appreciation to Chris Segrin, PhD, for statistical consultation, Linda Helmig Bram, PhD, for assistance in preparing the article, and Edward DeVos, EdD, for facilitating the IRB review.

Donald J. Viglione, PhD is a member of a company that sells the Rorschach Performance Assessment System manual and associated products. He is also an author and self-publisher of *Rorschach Coding Solutions*.

Correspondence concerning this article should be addressed to Anthony D. Bram, PhD, Department of Psychiatry, Cambridge Health Alliance, Harvard Medical School, 329 Massachusetts Avenue, No. 2, Lexington, MA 02420, United States. Email: Anthony_Bram@hms.harvard.edu

Despite these encouraging developments within R-PAS, there has been one change from the CS that does represent a loss to a psychoanalytic approach to Rorschach interpretation (Bram & Yalof, 2018): The elimination of coding Form Dominance in Shading and Achromatic Color (FDSHAC) determinants. Form Dominance refers to whether "Form," that is, shape as an indicator of cognitive organization/structure, is primary, secondary, or absent in containing an affect-laden "determinant" (Color, Shading, or Achromatic color). Shading determinants include Diffuse Shading (Y, interpreted in terms of stress-related anxiety), Shading Texture (T, anxiety associated with attachment needs), and Shading Vista (V, distress associated with self-critical introspection). Achromatic Color (C′) is an implicit indicator of dysphoric and irritated affect.[1]

In the CS, Form Dominance has been scored for responses with Color, Achromatic, and Shading determinants, but in R-PAS Form Dominance is only scored for responses with a Color determinant. The R-PAS authors (Meyer et al., 2011) made this decision to not code FDSHAC because nuanced distinctions among different degrees of Form Dominance (e.g., Form Dominance over Diffuse Shading [FY] vs. Form Secondary to Diffuse Shading [YF] vs. Formless Y) for Shading and Achromatic Color are more challenging for new learners (Viglione et al., 2017), do not yet have research support, and had not been interpreted formally in the CS structural summary. Also, because some specific FDSHAC determinants have low base rates, interrater reliabilities have been unstable across samples (Meyer et al., 2014).

In the psychoanalytic tradition of assessment, coding FDSHAC (as well as coding Form Dominance for Color [FDC]) is a crucial indicator of ego involvement in the regulation of various kinds of affects (Bram & Peebles, 2014; Kleiger, 1997; Peebles-Kleiger, 2002; Schafer, 1954). Greater integration of Form with Shading and Achromatic Color (SHAC) and Color determinants has been understood conceptually to indicate more cognitive control, frontal-lobe-based, organizing, containing activity in experiencing, modulating, and expressing emotions. For instance, a pure C′ (Achromatic Color) response ("It's a horrible storm, all of the darkness"; Formless) is less modulated and contained than a C′F ("A storm cloud. The darkness of it, and it's also kind of shaped like a cloud"; Form Secondary), which itself is less than a Form Dominance over Achromatic Color (FC′; "Cloud. Shaped like cloud, and it's dark, almost like it might rain"; Form Dominant). As Form plays a greater role, there is more indication of ego involvement in containing and modulating the dysphoric quality of experience. The absence of Form evokes a certain raw and uncontained experience. But with the current R-PAS approach to scoring, these kinds of psychodynamic nuances are lost, hindering inference-making with configurational and sequence analyses (Bram & Peebles, 2014; Schafer, 1954; Weiner, 1998)[2] and making further research less likely.

Schafer (1954) posited Form Dominance—which he referred to as "the relative emphasis on specific, articulated form" (p. 175)—to be one of six criteria to assess the degree of psychological adaptation versus instability within a response and across a protocol.[3] Schafer wrote that the absence or subordinate role of Form "indicates more or less failure to take hold of the problem situations represented by the ten Rorschach cards and impose articulate, meaningful structure on them" (p. 175). Schafer was speaking to the role of secondary process thinking in the regulatory function of the ego. Jettisoning FDSHAC limits the applicability of this psychoanalytic interpretive strategy when it would be valuable to understand with nuance how a person is impacted by and manages anxiety and dysphoria. Also, deemphasizing Form Dominance sacrifices teaching opportunities to bridge psychoanalytic theory with Rorschach practice. To make an evidence-based case that coding FDSHAC should be retained and disseminated in clinical assessment literature and training, we aimed to establish that these distinctions can be coded reliably. Previous studies have made use of coding guidelines from the CS text and workbook (Exner, 2003; Exner et al., 2001). Our study applied expanded scoring criteria, examples, training materials, and coding decision tree based on Viglione's (2010) rigorous companion to the CS workbook. Clinically, we hoped that demonstrating interrater reliability (IRR) will encourage assessors to code FDSHAC, at least adjunctively if using R-PAS. Because reliability sets the upper limit for validity (Murphy & Davidshofer, 2005), we also hoped this also would be a step toward developing valid Rorschach implicit measures of ego involvement in the regulation of anxiety and other affects.

There are three complementary vantage points to evaluate coding consistency of Rorschach variables, each addressing different questions (Meyer et al., 2002): (a) scores at the response level (e.g., whether a particular code is present or absent in each response), (b) segments of scores at the response level (agreement about coding decisions *within* a particular scoring category such as Location, Determinants, or Form Quality), or (c) scores at the protocol level (aggregated response-level scores). Despite skepticism from critics (e.g., Wood et al., 1996), IRRs of most CS and R-PAS scores have been generally good to excellent when assessed from each perspective (e.g., Kivisalu et al., 2016; Meyer et al., 2002; Pignolo et al., 2017; Schneider et al., 2022; Viglione et al., 2012, 2022).

Our study of FDSHAC involves reliability of *specific scores at the response level*, an approach "designed specifically to examine the extent to which two raters understand and agree on the … scoring rules" (Meyer et al., 2002, p. 269). This is the level relevant to assuring accuracy and consistency in response-by-response configurational and sequence analyses (Bram & Peebles, 2014; Peebles-Kleiger, 2002; Schafer, 1954; Weiner, 1998). Viglione and Meyer (2008) concluded from their review that though there is strong

[1] It bears emphasis that the mere presence of a SHAC determinant is not in and of itself a pathological indicator. Rather, a respondent's "ability to recognize and verbalize the characteristics of an inkblot that contribute to [one's] response reflects an ability to and willingness to mentally represent and share [one's] internal experience" (Bram & Peebles, 2014; see also Kleiger, 1992). To what extent a given SHAC determinant can be considered part of an adaptive response is discerned through *configurational analysis* involving the simultaneous consideration of structural scores (Location; Determinants including the degree of Form Dominance; Form Quality; Cognitive Scores; among others), thematic content, and patient–examiner interactions (Bram & Peebles, 2014; Peebles-Kleiger, 2002; Schafer, 1954; Weiner, 1998). At the same time, across a protocol (rather than within a single response), it is the *sum of* SHAC determinants that points toward disturbance (Meyer et al., 2011).

[2] Tracking the sequence of configurational analyses (defined in Footnote 1) as they unfold from response-to-response facilitates inferences about conditions under which and in what ways a patient's functioning is disrupted/destabilized and recovers (Bram & Peebles, 2014; Peebles-Kleiger, 2002; Schafer, 1954; Weiner, 1998).

[3] The other five criteria are: "emotional tone"; Form Quality; "integratedness of scores, images and attitudes"; "thematic moderation and balance"; and signs of "formal thought disorder" (Schafer, 1954, p 184).

support for the response-level IRR of most individual CS variables, among those with "lower reliabilities in some studies" (p. 30) are: (a) decisions about Form Dominance in Color, Shading, and Achromatic Color and (b) distinctions among shading subtypes (Y vs. T vs. C′ vs. V). We identified two studies that reported chance-corrected reliability (Kappas) for FDSHAC determinants and provided breakdowns of reliability at each level of Form Dominance (Acklin et al., 2000; Meyer et al., 2002).

Acklin et al.'s (2000) study involved two trained raters (both graduate students) coding two sets of 20 protocols, one nonpatient and one clinical set. For the nonpatient set, the ability of the coders to agree on FDSHAC distinctions ranged from fair (T, V) to good (Y) to excellent (C′).[4] For the clinical set, there was a somewhat different pattern: reliabilities were fair (Y) and excellent (T, C′); base rates were too low to report chance-corrected statistics for V. In each set, Acklin et al. also evaluated response-level IRR according to the presence or absence of specific FDSHAC distinctions (e.g., Form Dominance over Texture [FT], Form Secondary to Texture [TF], and Formless T, each as separate dichotomous variables[5]). Kappas for the Form Dominant responses ranged from poor (FT in the nonpatient set) to fair (FY, clinical set; Form Dominance over Vista [FV], nonpatient set) to excellent (FY, nonpatient; FT, clinical; FC′, both sets). For Form Secondary responses, when they could be reported, Kappas were fair (YF, nonpatient) and good (C′F, nonpatient); but in most instances (YF, clinical; TF, both sets, Form Secondary to Vista [VF], both sets; C′F, clinical), the base rates were too low to report a chance-corrected statistic. For Formless responses, Kappa could only be reported for Formless C′ (excellent) in the nonpatient set; base rates were again too low for each Shading determinant.

Meyer et al. (2002) reported chance-corrected IRR statistics for individual CS scores at the protocol level in two separate groups of raters—one a group of students, the other, of experienced coders—and protocols. Meyer et al. did not report on the overall reliability of raters to distinguish among FDSHAC levels (e.g., FT vs. TF vs. Formless T), but they did report response-level reliability according to the presence/absence of specific FDSHAC distinctions (e.g., FT, TF, and Formless T, each as separate dichotomous variables). Among student raters, Kappas for Form Dominant scores ranged from fair (FY, FV) to good (FT, FC′); for Form Secondary, from poor (C′F) to fair (YF, TF) to good (VF); for Formless, poor (C′) and good (Y). Base rates were too low to report Kappa for Formless T and Formless V. Among experienced raters, Kappas for Form Dominant were uniformly excellent (FY, FT, FV, FC′); for Form Secondary, Kappas ranged from good (TF, VF) to excellent (YF, C′F); and for Formless, Kappas were excellent (Formless Y, Formless C′) when base rates allowed.

Together, the Acklin et al. (2000) and Meyer et al. (2002) studies reveal variability in IRR involving distinctions among FDSHAC determinants. Meyer et al.'s study in particular highlighted greater scoring difficulties for less experienced coders. Coding inconsistencies and challenges such as these inspired Viglione (2010) to augment the CS materials (Exner, 2003; Exner et al., 2001).

In our study, four experienced assessment psychologists independently scored determinants for 155 Rorschach responses, 115 (74%) of which were selected because they involved decisions about at least one SHAC determinant. Determinants were scored according to R-PAS except for SHAC responses, in which FDSHAC was coded based on Viglione (2010). We hypothesized that these guidelines would lead to good (or better) reliability for FDSHAC.

Method

Raters

The four assessment psychologists recruited as raters averaged 31.75 years of overall Rorschach experience (SD = 17.00; 7–45), 24.25 years with the CS (SD = 15.09; 5–39), and 8.50 with R-PAS (SD = 1.29; 7–10). All were active clinical Rorschach practitioners and had experience teaching/supervising various aspects of the Rorschach including coding.

Training

Raters were provided a training packet that included (a) an orientation to the study, (b) pages from the R-PAS manual (Meyer et al., 2011, pp. 103–110) on coding presence of SHAC, (c) Chapter 3 "Color, Achromatic Color, & Shading versus Form and Color for Location" from Viglione (2010), (d) a flowchart for coding FDSHAC derived from Viglione (2010; Supplemental Material S1), and (e) a reference sheet illustrating key terms (Supplemental Material S2).

After reviewing the packet, raters discussed questions with the first author. Next, they coded determinants for 20 practice Rorschach responses (extracted from Exner et al., 2001), 18 of which involved a distinction of FDSHAC. Coding of practice items was compared with benchmarks established by the first two authors, and detailed feedback and discussion followed before raters were approved to proceed with the study.

Rorschach Responses and Protocols

Raters independently coded 155 Rorschach responses, 115 (74%) of which involved a FDSHAC determination (based on benchmark consensus of the first two authors, the second of whom authored the guidelines). The 155 items were presented in the form of six complete R-PAS protocols—Response and Clarification Phases for Cards I through X—each between 25 and 27 responses. All but five of the 155 responses were extracted from child, adolescent, and adult clinical protocols obtained by the first author. All protocols were administered in English in outpatient settings in the United States (predominantly patients who did not meet criteria for psychosis), using either CS or R-PAS guidelines (roughly half of each). Those other five responses involved extremely low base-rate codes (e.g., VF and Formless Y, T, and V), so they were adapted from sample responses in Exner et al. (2001). Constructing protocols in this way circumvents the problem of low base rates that have interfered with IRR estimates in previous studies using existing patient protocols. It minimizes the standard error of measurement and increases the precision of reliability coefficients in a way that enhances confidence in potential comparisons with earlier studies not making use of Viglione's (2010) guidelines. Eliminating low base rates also facilitates the power of comparisons among IRR's of SHAC determinants within this study.

Responses sampled a mix of degrees of Form Dominance (Formless, Form Dominant, Form Secondary) for Shading (Y, T, V), and

[4] Acklin et al. (2000) treated FDSHAC distinctions (e.g., no Y vs. FY vs. YF vs. Formless Y) as "polychotomous" nominal variables. Our study involved evaluation of reliability of FDSHAC as ordinal variables. Interpretive ranges for Kappas in previous studies: fair (0.40–0.59), good (.60–0.74), excellent (.75 and above; Cicchetti, 1994; Shrout & Fleiss, 1979).

[5] To clarify, the FT variable was coded 1 if presentand 0 if absent. The same kind of binary present/absent coding was applied to the TF and T variables as well as to each of the levels of the other SHAC determinants.

Achromatic Color (C′). Based on benchmark coding, the 115 SHAC responses were comprised of the following (raw numbers with base rates): FY = 11 (1.8 per protocol), YF = 16 (2.7), Y = 6 (1.0); FT = 13 (2.2), TF = 8 (1.3), T = 6 (1.0); FV = 13 (2.2), VF = 9 (1.5), V = 7 (1.2); FC′ = 9 (1.5), C′F = 12 (2.0), C′ = 8 (1.3). These included blends among SHAC and/or other determinants. None of these FDSHAC determinants could be considered to have a low base rate (defined as <1 per protocol by Meyer et al., 2011).[6]

Plan for Interrater Reliability Analyses

Key variables and their scales, the items to be included in different versions of the analyses, and the IRR statistics to be presented are previewed in Supplemental Material S3.

Aggregate FDSHAC Variable (Ordinal)

The first row of Supplemental Material S3 shows *Aggregate FDSHAC*, an ordinal variable that captures the relative involvement of Form in *any* responses in which a Y, T, V, and/or C′ determinant was coded.[7] The second column of this row shows the 4-point ordinal Aggregate FDSHAC scale: shading or achromatic color was *not* coded = 0; Form Dominant over any SHAC determinant = 1; Form Secondary to SHAC determinant = 2; SHAC determinant with Form absent (i.e., Formless) = 3. Because of the relative ease of coding 0 for responses without a SHAC determinant, the coefficients for all 155 responses could be considered inflated. Therefore, we also conducted more conservative analyses of reliability based on only the 115 items that the benchmark raters coded a SHAC determinant.

Specific FDSHAC Variables (Ordinal)

Because FDSHAC might vary depending on the particular SHAC determinant, we also computed reliability coefficients for coding degree of Form involvement as applied *separately* to each of the Y, T, V, and C′ determinants. In the second row of Supplemental Material S3, we refer to these as ordinal *Specific FDSHAC* variables. Thus, Form Dominance for Y (Form Dominance of Diffuse Shading [FDY]), the scale was: 0 = no Y coded, 1 = FY, 2 = YF, 3 = Formless Y. Scales for Form Dominance of T, V, C′, (Form Dominance of Texture [FDT], Form Dominance of Vista [FDV], Form Dominance of Achromatic Color [FDC′]) were similar. For the same reason as above, we computed reliabilities for not only all 155 items, but we ran two more conservative analyses based on benchmark coding of the presence (a) of any SHAC determinant (again, n = 115) and (b) of the particular determinant. Thus, the latter analyses for FDY were based on 33 such items; for FDT, 27; and for FDV and FDC′ each, 29. We understood the last two versions of analyses not to be truly fair tests of reliability because (a) the 0 point of the 4-point scales was essentially eliminated, creating an artificial 1–3 scale involving only the most difficult discriminations (between Form Dominance and Form Secondary *and* between Form Secondary and Formless) and (b) they are based on a relatively small number of items. Nevertheless, we included these because we believed them informative about difficult coding discriminations.

Different Levels of FDSHAC Variables (Dichotomous)

The third row of Supplemental Material S3 presents plans to assess IRR's separately at the Form Dominant, Form Secondary, and Formless *levels* of the aggregate and specific FDSHAC scales: Each variable is dichotomous based on its presence or absence. For example, FY was coded 1 if present and 0 if absent. The same scheme was applied to YF, Formless Y, and then to each of the levels of FDT, FDV, and FDC′.[8] For each variable, we conducted three versions of the analyses with different item sets included, as described earlier.

Reliability Coefficients

We report reliability in terms of percent agreement, Fleiss' Kappa, Kendall's Tau-b, and Gwet's AC1 or AC2. The fourth column of Supplemental Material S3 displays the particular coefficients to be reported for each reliability analysis. Kappas adjust for chance agreement not reflected by percent agreement. But Kappas only credit *exact* agreement among raters, which is not ideal for ordinal variables where it is meaningful how close or apart inexactness is. Kappas have also been criticized for being too stringent in correcting for chance agreement, especially when ratings are more likely to be assigned to one category than others (which is the case for our 0–3 scales, where the 0 category is most prevalent; Zec et al., 2017) and in which raters are well trained and unlikely to guess (Gwet, 2014; McHugh, 2012). Because Kappa has been the conventional statistic for response-level reliability, we report it to compare with past studies. For another perspective, we also report Kendall Tau-b, an associational statistic crediting close-but-not-exact agreement but does not adjust for chance agreement. Finally, for ordinal variables, we report Gwet's (2014) AC2 statistic developed as an alternative to Kappa in that it corrects for chance but lowers the maximum theoretical value for chance agreement, takes into account the closeness of inexact agreements, and weights unanimity of agreement according to number of raters (see Lewey et al., 2019). Our premise is that Gwet's AC2 is the reliability statistic with the best fit for our ordinal FDSHAC variables coded by trained assessors. But we include the other statistics as they offer other vantage points for understanding patterns of IRR. For dichotomous variables, we report Gwet's AC1 in addition to Kappa because, like AC2, AC1 more appropriately adjusts for chance agreement (Gwet, 2014). So, in a way similar to interpretation of a Rorschach protocol, our approach involves interpreting each reliability statistic not in isolation but, rather, relative to one another and contextually.

We interpreted Kappa and Kendall's Tau-b according to the framework conventional for Rorschach IRR research (e.g., Meyer et al., 2002). This involves the following interpretive ranges: below.40 *poor*; .40–.59 *fair*; .60–.74 *good*; .75 or above *excellent* (Cicchetti, 1994; Shrout & Fleiss, 1979). But because Gwet's AC1 and AC2 coefficients run higher than Kappas, we applied more stringent interpretive criteria for these: below .50, *poor*; .50–.74, *moderate*; .75–.89, *good*; .90 or above, *excellent* (Koo & Li, 2016).

[6] An alternative to this "per protocol" definition of low base rate was provided by Meyer et al. (2002): <0.01 (less than one per 100 response). None of the FDSHAC determinants in this study had a base rate that low; the lowest was Formless Y at 0.04 (6 out of 155 total responses).

[7] To be consistent with the clinical literature (e.g., Bram & Peebles, 2014; Kleiger, 1997S; Peebles-Kleiger, 2002; Schafer, 1954), we define this and similar ordinal scales in terms of Form Dominance. Technically, though, such ordinal scales capture the relative involvement of SHAC determinants: 0 = absence of a SHAC determinant; 1 = SHAC secondary to Form; 2 = SHAC dominant over Form; 3 = Formless SHAC determinant.

[8] To further illustrate the dichotomous nature of these variables, coding FY as absent (FY = 0) means that the rater coded either YF, Formless Y, or no Y at all.

Results

Interrater Reliabilities of Aggregate FDSHAC

The first row of Table 1 displays the IRR's for the ordinal Aggregate FDSHAC variable. When analyzing all 155 items, the constellation of percent agreement, Kendall Tau-b, Kappa, and Gwet's AC2 indicates that raters were consistent in differentiating among the four points on the scale. Kappa is good at .67 as is Gwet's AC2 at .83. The correlational Kendall Tau-b of .80 highlights that even if raters did not achieve exact agreement, they were fairly close. Conducting the same analyses but excluding items coded 0 by the benchmark ($n = 115$) reduces percent agreement (three-quarters to two-thirds), Kendall Tau-b (excellent to good), and Kappa (good to fair), but Gwet's AC2 remains good.

Table 2's first row breaks down IRR for each of the *levels* (each coded as present vs. absent) in which a SHAC determinant was coded. Across all 155 items (first row, first and fourth columns of Table 2), raters were more consistent with each other in coding Form Dominant and Formless codes (good Kappas and AC1′s) compared to coding Form Secondary (fair Kappa, moderate AC1). Analyzing only the 115 items coded SHAC by the benchmark, we see a similar pattern (see Table 2, first row, second and fifth columns).

Interrater Reliabilities for Specific FDSHAC Determinants (FDY, FDT, FDV, FDC′)

Interrater Reliabilities for Form Dominance of Diffuse Shading

In the second row of Table 1, reliabilities reflect raters' abilities to consistently differentiate FDY. In the analyses of 155 and 115 items, Gwet's AC2′s, .89 and .82 respectively are good, and Kappas are fair. Correlation coefficients—Kendall Tau-b′s of .63 and .57—suggest that when raters disagree, they are not as close as expected. Predictably, all coefficients drop in the stringent analyses of the 33 items, but Gwet's AC2 is moderate.

The second row of Table 2 shows within-level reliabilities—FY, YF, and Formless Y (each coded dichotomously). In the 155- and 115-item sets, Gwet's AC1′s indicate reliabilities to be excellent (or just short for FY, $n = 115$) at Form Dominant and Formless levels and good for Form Secondary. In the 33-item set, Gwet's AC1′s are moderate for FY, good for Formless Y, and moderate for YF. In all three sets, Kappas are fair for FY and Formless Y and poor for YF. Again, the Form Secondary code was the most difficult to code reliably.

Interrater Reliabilities for Form Dominance of Texture

The third row of Table 1 reveals that reliability coefficients for ordinal coding of FDT are fairly strong in the 155- and 115-item analyses: Gwet's AC2′s are excellent, and Kappas are good. Kendall Tau-b′s are excellent, underscoring that when raters were not in exact agreement, they were close. The analysis restricted to the 27 items for which benchmark coded Texture showed Gwet's AC2 to be moderate and Kappa fair. Turning again to percent agreements and breakdowns among different levels of codes in Table 2 (third row), TF proved most difficult.

Table 2 underscores that, overall, raters were fairly consistent at each level of FDT, especially at the dichotomous FT and Formless T levels: In the 155- and 115-item sets, AC1′s are excellent; and Kappas, good. In the 27-item set, Kappas were fair for FT and good for Formless T, and AC1′s were moderate for FT and good for Formless T. As for the relatively weaker reliability for TF, Gwet's AC1′s were actually excellent in the 155- and 115-item sets and moderate in the 27-item set; and Kappas were fair in the first two sets and poor in the third.

Table 1
Interrater Reliabilities for Ordinal FDSHAC Scales: Means, Standard Deviations, and 95% Confidence Intervals

Row/variable	Benchmark criteria	Percent agreement	Kendall's Tau-b	Fleiss' Kappa	Gwet's AC2
1. Aggregate FDSHAC					
155 items	—	75% (4.53)	.80 (.04)[c]	.67 (.06)[b] [.61–.73]	.83 (.04)[e] [.80–.88]
115 items	SHAC > 0	67% (6.92)	.62 (.07)[b]	.52 (.08)[a] [.43–.60]	.82 (.05)[e] [.77–.86]
2. FDY					
155 items	—	82% (2.07)	.63 (.07)[b]	.54 (.05)[a] [.45–.64]	.89 (.03)[e] [.85–.93]
115 items	SHAC > 0	77% (3.11)	.57 (.04)[a]	.51 (.04)[a] [.42–.62]	.83 (.04)[e] [.76–.89]
33 items	Any Y > 0	59% (5.64)	.45 (.14)[a]	0.44 (.07)[a] [.28–.59]	.60 (0.11)[d] [.40–.80]
3. FDT					
155 items	—	91% (3.43)	.84 (.06)[b]	.74 (.09)[b] [.61–.82]	.95 (.02)[f] [.92–.98]
115 items	SHAC > 0	89% (3.82)	.81 (.06)[b]	.73 (.09)[b] [0.64–.81]	.92 (.03)[f] [.88–.97]
27 items	Any T > 0	69% (7.89)	.46 (.13)[a]	.53 (.10)[a] [.34–.70]	.74 (.08)[d] [.54–.93]
4. FDV					
155 items	—	89% (2.23)	.68 (.07)[b]	.54 (.08)[a] [.43–.65]	.94 (0.01)[f] [.92–.97]
115 items	SHAC > 0	85% (2.40)	.65 (.06)[b]	.52 (.07)[a] [.41–.64]	.92 (.03)[f] [.89–.96]
29 items	Any V > 0	47% (9.01)	.32 (.10)	.29 (.11) [.12–.42]	.41 (.09) [.24–.58]
5. FDC′					
155 items	—	91% (2.26)	.81 (.05)[c]	.73 (.06)[b] [.64–.81]	.95 (.02)[f] [.92–.98]
115 items	SHAC > 0	87% (2.70)	.78 (.05)[c]	.71 (.05)[b] [.62–.80]	.92 (.03)[f] [.88–.97]
29 items	Any C′ > 0	70% (7.15)	.59 (.11)[a]	.57 (.10)[a] [.40–.73]	.76 (.08)[e] [.61–.91]

Note. FDSHAC = form dominance in shading and achromatic color; SHAC = shading and achromatic color; FDY = form dominance in diffuse shading; FDT = form dominance in texture; FDV = form dominance in vista; FDC′ = form dominance in achromatic color.
Interpretive ranges for Kappa and Kendall Tau-b: [a] fair (0.40–0.59); [b] good (0.60–0.74); [c] excellent (0.75 and above; Cichetti, 1994; Shrout & Fleiss, 1979). Interpretive ranges for Gwet's AC2: [d] moderate (.50–.74); [e] good (.75–.89); [f] excellent (.90 and above; Koo & Li, 2016).

Table 2
Interrater Reliabilities for Different Levels of FDSHAC and FDC Scales: Means, Standard Deviations, and 95% Confidence Intervals

Row/variable	Fleiss' Kappa (155 items)	Fleiss' Kappa for items coded BM SHAC > 0 (115 items)	Fleiss' Kappa for items coded for key determinant by BM (*n* = #items)	Gwet's AC1 (155 items)	Gwet's AC1 BM SHAC > 0 (115 items)	Gwet's AC1 for items coded for key determinant by BM (*n* = #items)
1. Aggregate FDSHAC						
Form Dom. or not	.65^{b} [.57–.74]	.60^{b} [.50–.70]	—	.75^{d} [.68–.83]	.63^{c} [.53–.73]	—
Form secondary or not	.42^{a} [.32–.52]	.34 [.23–.45]		.64^{c} [.55–.74]	.46 [.34–.57]	
Formless or Not	.66^{b} [.55–.77]	.63^{b} [.52–.75]		.86^{d} [.81–.92]	.80^{d} [.72–.88]	
2. Diffuse shading (Y)			(*n* = 33)			(*n* = 33)
FY or not	.55^{a} [.39–.72]	.55^{a} [.39–.72]	.53^{a} [.33–.74]	.92^{e} [.88–.95]	.89^{d} [.83–.94]	.61^{c} [.41–.81]
YF or not	.31 [.16–.45]	.31 [.16–.45]	.29 [.10–.47]	.87^{d} [.82–.92]	.82^{d} [.75–.88]	.50^{c} [.28–.73]
Formless Y or not	.50^{a} [.28–.72]	.50^{a} [.28–.72]	.55^{a} [.29–.81]	.94^{e} [.91–.97]	.92^{e} [.88–.96]	.78^{d} [.62–.95]
3. Texture (T)			(*n* = 27)			(*n* = 27)
FT or not	.74^{b} [.61–.87]	.73^{b} [.60–.86]	.59^{a} [.37–.81]	.95^{e} [.92–.98]	.93^{e} [.89–.97]	.59^{c} [.37–.81]
TF or not	.41^{a} [.18–.64]	.40^{a} [.16–.63]	.32 [.02–.63]	.94^{e} [.90–.97]	0.91^{e} [.87–.96]	.62^{c} [.41–.84]
Formless T or not	.63^{b} [.40–.85]	.62^{b} [.39–0.85]	.64^{b} [.39–.90]	.97^{e} [.94–.99]	0.95^{e} [.92–.99]	.81^{d} [.64–.98]
4. Vista (V)			(*n* = 29)			(*n* = 29)
FV or not	.48^{a} [.26–.70]	.47^{a} [.25–.70]	.40^{a} [.14–.67]	.95^{e} [.92–.98]	.93^{e} [.89–.97]	.67^{c} [.45–.89]
VF or not	.32 [.15–.49]	.30 [.13–.48]	.11 [.00–.32]	.91^{e} [.87–.95]	.88^{d} [.82–.93]	.35 [.10–.61]
Formless V or not	.50^{a} [.23–.76]	.50^{a} [.24–.76]	.43^{a} [.14–.72]	.97^{e} [.95–.99]	.96^{e} [.93–.99]	.80^{d} [.64–.96]
5. Achromatic color C′			(*n* = 29)			(*n* = 29)
FC′ or not	.69^{b} [.52–.85]	.68^{b} [.51–.84]	.66^{b} [.45–.87]	.95^{e} [.92–.98]	.93^{e} [.88–.97]	.71^{c} [.51–.91]
CF′ or not	0.56^{a} [.38–.74]	.54^{a} [.36–.73]	.47^{a} [.24–.71]	.93^{e} [.90–.97]	.91^{e} [.86–.96]	.57^{c} [.34–.81]
Formless C′ or not	.71^{b} [.54–.89]	.71^{b} [.53–.88]	.71^{b} [.50, 0.92]	.97^{e} [.94–.99]	0.95^{e} [.92–.99]	.82^{d} [.67–.98]

Note. Each variable level coded dichotomously as present/absent. BM = benchmark; Form Dom. = Form Dominant; FDSHAC = Form Dominance and Shading in Achromatic Color; SHAC = Shading and Achromatic Color; FDC = Form Dominance for Color; FY = Form Dominance over Diffuse Shading; YF = Form Secondary to Diffuse Shading; FT = Form Dominance over Texture; TF = Form Secondary to Texture; FV = Form Dominance over Vista; VF = Form Secondary to Vista; FC′= Form Dominance over Achromatic Color; CF′ = Form Secondary to Achromatic Color. Interpretive ranges for Kappa and Kendall Tau-b: ^{a} fair (0.40–0.59); ^{b} good (.60–0.74); ^{c} moderate (.50–.74); ^{d} good (.75–.89); ^{e} excellent (.90 and above; Koo & Li, 2016).

Interrater Reliabilities for Form Dominance of Vista

In the fourth row of Table 1, the mean overall Gwet's AC2′s for FDV are excellent in the 155- and 115-item sets; whereas, similar to FDY, Kappas were fair. The fourth row of Table 2 shows that the VF level—even more so than Form Secondary for other FDSHAC determinants—challenged raters' consistency. This was specially illuminated in the 29-item set of only items that the benchmark coded Vista: For VF, Gwet's AC1 and Kappa were poor.

Interrater Reliabilities for Form Dominance of Achromatic Color

The fifth row of Table 1 illuminates that reliabilities for FDC′ were consistently robust. Gwet's AC2′s are excellent in the 155- and 115-item analyses and good in the 29-item analysis. Kappas are good for the 155- and 115-item sets and fair for the 29 items. Kendall Tau-b′s show that when agreement was not exact, raters were still relatively close (.81 for 155 items, .78 for 115 items, and .59 for 29 items). The fifth row of Table 2 shows the familiar pattern in which Form Secondary (C′F) was the least reliable. However, relative to the other SHAC determinants, conservative Kappas for the Form Secondary category was highest for C′F: .56 for C′F compared to .31 for YF, .41 for TF, and .32 for VF.

Interrater Reliabilities of Specific FDSHAC′s: Comparison to Previous Studies

To determine whether Viglione's (2010) guidelines improve reliability of scoring Form Dominant versus Form Secondary versus, Formless FDSHAC determinants, we compared our findings to the two previous studies that employed chance-corrected coefficients (Acklin et al., 2000; Meyer et al., 2002). Those studies used Kappas, so that is the only statistic we report in Supplemental Material S4. Because those prior studies involved a collection of full Rorschach protocols (in contrast to our "protocols" constructed to increase opportunities to code all FDSHAC′s), base rates were too low for many of the variable for the authors to report Kappas. Generally, the addition of Viglione's guidelines did not yield improvements over the earlier studies. In distinguishing among the levels, our Kappas are comparable to Acklin et al.'s for FDC′ (our .77 vs. their graduate students' .79 and .76 for nonpatient and clinical protocols, respectively) and FDV (our .54 vs. their .55 for nonpatient protocols).[9] For FDY, our Kappa was .54, whereas Acklin et al. reported .55 for clinical protocols and .70 for those from nonpatients. For FDT, our Kappa is .74, and Acklin et al.'s was .44 for nonpatient and .80 for clinical.

Meyer et al.'s pairs of experienced raters were more consistently reliable across levels (Kappas .86–.95 for Form Dominant, .67–.91 for Form Secondary, and .89–1.00 for Formless[10]) than our experienced raters who had the benefit of Viglione's guidelines (.48–.74 for Form Dominant, .31–.56 for Form Secondary, and .51–.71 for Formless).

Discussion

Applying Viglione's (2010) guidelines for coding FDSHAC, IRR's were stronger based on the more recently developed Gwet's AC2 and AC1 statistics than the traditional Kappas that overcorrect for chance and are less appropriate for assessing reliability of ordinal variables (Gwet, 2014; McHugh, 2012). Gwet's AC1′s and AC2′s were generally in the good-to-excellent range, whereas Kappas tended fair to good. This was the case when we looked at IRR's for both Aggregate FDSHAC and the specific FDY, FDT, FDV, and FDC′ scales. Further, more stringent analyses—involving exclusion of items coded 0 by the benchmark for the specific SHAC determinant—revealed that reliabilities varied among FDY, FDT, FDV, and FDC′: Gwet's AC2′s show that raters were most consistent with each other in making the finer coding distinctions (Form Dominant vs. Form Secondary vs. Formless) among levels of FDT (.74 for 27 items coded for any Texture by the benchmark) and FDC′ (.76 for the 29 items coded for any Achromatic Color by the benchmark), somewhat less so for FDY (.60 for the 33 items coded for any Diffuse Shading by the benchmark), and had most difficulty with FDV (.41 for the 29 items coded for any Vista by the benchmark).

When examining reliabilities at specific levels of Form involvement at each level of FDSHAC determinant, Form Secondary was most difficult to code consistently and had the greatest variability. This was not surprising as coding Form Secondary includes making distinctions from *both* the Form Dominant and Formless levels. Among the dichotomous Form Secondary variables, VF showed the poorest reliability and greatest variability.

When comparing to the two published studies (Acklin et al., 2000; Meyer et al., 2002) reporting chance-corrected reliabilities for FDSHAC, our raters' training with and access to Viglione's (2010) guidelines were not associated with improved reliabilities. Notably, Meyer et al.'s (2002) pairs of experienced raters consistently showed stronger reliabilities at each level of FDSHAC than our four experienced raters. Possible factors contributing to the superiority of Meyer et al.'s reliabilities include that (a) five of their six experienced raters were published Rorschach researchers in their own right so perhaps had more mastery of coding and established reliability going into their study and (b) that their four paired sets of raters were trained in the same setting or otherwise had experience collaborating on Rorschach coding projects so were already well calibrated with one another.

Implications for Rorschach Coding Systems and Clinical Assessment

Coding FDSHAC Determinants

Even with Viglione's (2010) guidelines for coding FDSHAC, our experienced raters had relative difficulty with consistency in distinguishing between the Form Dominant and Form Secondary levels and between Form Secondary and Formless. Though it is not certain if our raters performed better than they would have if only guided by the CS materials, there is clearly room for current CS authors to strengthen instructions for coding distinctions along the continuum of coding FDSHAC, especially around Form Secondary.

We found the strongest support for IRR's of Texture and Achromatic Color. Likely the verbal and nonverbal communication used for Texture (e.g., "feels soft"; "fluffy" while rubbing a shaded area) and Achromatic Color ("the blackness makes it looks evil") are more readily and accurately pegged to a determinant category compared

[9] Recall that Meyer et al. (2002) did not report reliabilities differentiating among Form Dominance, Form Secondary, and Formless SHAC.

[10] Because of low base rates, Meyer et al. (2002) did not report Kappas for Formless T or Formless V.

to communications for other SHAC′s. This is consistent with Kivisalu et al.'s (2016) suggestion to bolster training materials to consistently identify Diffuse Shading- and, especially, Vista-related scores. Interestingly, the recently published revised Comprehensive System (CS-R; Exner et al., 2022) has eliminated coding of Form Dominance for Vista and Texture (now each dichotomous). Our findings support the decision for the former but not for the latter.

Assessing Ego Involvement Using Configurational/Sequence Analyses

In psychodynamic assessment, response-by-response configurational and sequence of Rorschach scores (Location, Content, Determinants, Form Quality, Cognitive Codes, etc.), along with simultaneous tracking of patient–examiner interactions, offer a window into the nuances of how a patient moves through and metabolizes various kinds of emotional experiences (Bram & Peebles, 2014; Peebles-Kleiger, 2002; Schafer, 1954; Weiner, 1998). Consideration of FDSHAC (and FDC) in terms of degree of ego involvement or cognitive control is central to inference-making about under what conditions and in what way a person becomes emotionally destabilized and recovers. A shift in the sequence from Form Dominant responses to Form Secondary or Formless is a clue to possible destabilization, that is, the patient shows more difficulty cognitively containing their emotional experience. A shift back to Form Dominance is a clue to possible recovery. Other such clues include Form Quality, Cognitive Codes, and the presence or absence of inanimate movement (Bram & Peebles, 2014; Peebles-Kleiger, 2002). The greater convergence of such clues, the more confident we can be in inferences about conditions-under-which dynamics of a person's emotional functioning. Findings from configurational and sequence analyses that attend to FDSHAC have their greatest value in psychodynamically attuned treatment planning (Bram & Peebles, 2014). For instance, with a patient who loses Form Dominance, lapses in Form Quality, and merits Cognitive Codes in responses also coded for Achromatic Color and/or Morbid content, an inference would be that they are prone to being overwhelmed and destabilized when stirred by dysphoric feelings. Implications involve the therapist's careful pacing, timing, and support around the processing of the patient's depressive experiences.

Based on our findings, we encourage assessors applying this interpretive approach to be cautious and thoughtful in making FDSHAC distinctions. Although the empirical jury is still out on whether Viglione's (2010) guidelines improve reliability above CS instructions for coding FDSHAC (elaborated below), it makes sense for assessors to have more rather than less guidance about the various contingencies involved in making these fine-tuned determinations.

Limitations

There were three important limitations to our methodology that potentially impact generalizability. First, our raters were already experienced in coding the Rorschach, so it is not certain to what extent reliability would be as solid with new learners even if trained similarly. As Viglione et al. (2017) found, new learners of the CS reported some of their greatest challenges in making FDSHAC distinctions as well as differentiating among the specific SHAC determinants. Second, our raters were aware that we were interested in studying FDSHAC, so they were possibly more vigilant to "get right" these coding distinctions. Third, the Rorschach "protocols" were artificial (compared to intact clinical protocols) in being loaded for responses requiring an FDSHAC determination. This meant raters were continually faced with these decisions, amplifying their vigilance as well as familiarity with the coding parameters.

Considerations for Future Research

Although our IRR's did not exceed those of previous studies (Acklin et al., 2000; Meyer et al., 2002), our project does not directly answer the question of whether Viglione's (2010) guidelines improve reliability over the original parameters presented in the CS. To answer this question, four or more coders experienced with CS coding could: (a) review the CS material on making FDSHAC determinations; (b) independently score determinants for a large set of Rorschach responses based on the CS; (c) train with Viglione's (2010) guidelines; and (d) apply the latter to rescore the set or score a second set.

We also believe that there would be benefit to a study similar to this one but with a specific aim of enhancing IRR at the Form Secondary level. As Form Secondary is the relative Achilles' heel in coding FDSHAC, we have in mind ways of sharpening training to more clearly operationally define and provide examples of questions in the decision trees, for example, what is meant by a response that "incorporates specific/fixed shape" or a "nonspecific shape" and thresholds for determining when a SHAC determinant is "*emphasized* before Form in the Clarification Phase" (see Supplemental Material S2). Further, to demonstrate the clinical generalizability of findings, it would be advantageous to conduct such follow-up studies with raters who are Rorschach novices or at least with less expertise than our experienced group.

An alternative, more pragmatic approach to addressing the challenges of Form Secondary is to evaluate reliabilities of 3-point versions of ordinal FDSHAC scales. Three-point versions would combine Form Secondary level with either the Form Dominant (into a "Form Integrated" category) or Formless (into "Form Deemphasized") level. Though a 3-point version sacrifices some interpretive nuance, it retains more than R-PAS's dichotomizing of SHAC′s.

Though there is still a need to tighten reliability of coding FDSHAC distinctions, our findings indicate that Viglione's (2010) guidelines yield reliabilities that are more than acceptable to consider assessing its construct validity in implicit measures of ego involvement in affect regulation. There are at least three ways the construct validity of FDSHAC could be studied. First, FDSHAC could be examined as it is combined with FDC into a single implicit measure of affect regulation. Second, coding FDSHAC could be part of revamping a more reliable version of Rapaport et al.'s (1968) "new F%," a measure of Form Dominance across a protocol.[11] Or third, FDSHAC could be further investigated as assessing different implicit aspects of emotional regulation compared to FDC. Although in clinical protocols the base rates of the individual codes (e.g., FY, YF, Y, FT, TF, T, and so forth) are low, when SHAC variables are aggregated they occur with a similar frequency ($M = 4.7$ per protocol in normatively) as Color responses ($M = 4.3$ per protocol; Meyer et al., 2011; Bruce Smith, personal communication, April 11, 2020).

[11] The only study aiming to validate "new F%" (Gardner, 1951) did not attend to interrater reliability in coding Form Dominance, and it also involved a small sample size. Such factors likely contributed to the lack of success in establishing validity (Murphy & Davidshofe, 2005).

Our findings suggest that IRR's of Aggregate FDSHAC are good when experienced assessors apply Viglione's guidelines.

Finally, though additional research is warranted to establish a sound evidence base justifying routine coding of FDSHAC in clinical assessment, we hope this study reawakens interest in this psychodynamically meaningful variable. Recently, Rorschach researchers have emphasized developing variables where there is a clear conceptual link between interpretation and the "response process," that is, the "psychological operations and behaviors that occur while [a person gives] a response" (Mihura et al., 2018, p. 235). The FDSHAC variable meets this criterion as its "psychological operations" involving the degree of Form used in articulating affect-laden determinants are readily mapped onto interpretations of the involvement of the frontally based, organizing, affect-regulating function of the ego.

摘要

尽管在评估训练中对精神动力学思维的重视已经减弱,但崛起的罗夏表现评估系统(R-PAS;Meyer et al., 2011)将精神分析概念重新整合到实证的罗夏评估中:R-PAS增加了涉及客体关系、内隐依赖、攻击性意念和自我障碍的分数。然而,R-PAS排除了评估自我参与焦虑/焦虑调节的精神动力学框架,方法是消除阴影和消色差(FDSHAC)中的优势形式编码,该编码一直是综合系统的部分(CS;Exner, 2003)。该决定部分是基于对FDSHAC的优势形式、次要形式和无形式之间有效可靠的编码差异的担忧。为了确定这种区分可以被可靠地编码,我们采用了补充指南(Viglione, 2010)来评估四名经验丰富的评估员的可靠性,他们为155个罗夏反应编码了决定因素,其中115个需要FDSHAC确定。将Gwet的AC2'应用于序数量表,评分者间的信度从良好到优秀。纹理和消色差的优势形式的信度最强,漫反射阴影中的优势形式则不大,而远景的优势形式则存在问题。在FDSHAC变量的形式参与水平中,评分者最难区分次要形式。我们讨论了临床编码的考虑因素,用于解释的精神动力学配置分析,并构建了验证研究。

关键词：罗夏，自我参与，情感调节，优势形式

References

Acklin, M. W., McDowell, C. J., II, Verschell, M. S., & Chan, D. (2000). Interobserver agreement, intraobserver reliability, and the Rorschach Comprehensive System. *Journal of Personality Assessment*, *74*(1), 15–47. https://doi.org/10.1207/S15327752JPA740103

Bram, A. D., & Peebles, M. J. (2014). *Psychological testing that matters: Creating a road map for effective treatment*. American Psychological Association. https://doi.org/10.1037/14340-000

Bram, A. D., & Yalof, J. (2018). Two contemporary Rorschach systems: Views of two experienced Rorschachers on the CS and R-PAS. *Journal of Projective Psychology and Mental Health*, *25*, 35–43.

Bram, A. D., Yalof, J., & Gottschalk, K. A. (2018). Psychological testing and psychoanalysis: Fixable gap or great divide? *Journal of Projective Psychology and Mental Health*, *25*, 166–172.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Exner, J. E. (2003). *The Rorschach: A comprehensive system: Vol. 1. Basic foundations and principles of interpretation* (4th ed.). Wiley.

Exner, J. E., Andronikof, A., & Fontan, P. (2022). *The Rorschach: A Comprehnsive System—Revised administration and coding manual*. Rorschach Workshops.

Exner, J. E., & Colligan, S. C., Hillman, L. R., Metts, A. S., Ritzier, B. A., Rogers, K. T., Sciara, A. D., & Viglione, D. J. (2001). *A Rorschach workbook for the comprehensive system* (5th ed.). Rorschach Workshops.

Gardner, R. W. (1951). Impulsivity as indicated by Rorschach test factors. *Journal of Consulting Psychology*, *15*(6), 464–468. https://doi.org/10.1037/h0061368

Gwet, K. (2014). *Handbook of inter-rater reliability* (4th ed.). Advanced Analytics Press.

Kivisalu, T. M., Lewey, J. H., Shaffer, T. W., & Canfield, M. L. (2016). An investigation of interrater reliability for the Rorschach Performance Assessment System (R–PAS) in a nonpatient U.S. sample. *Journal of Personality Assessment*, *98*(4), 382–390. https://doi.org/10.1080/00223891.2015.1118380

Kleiger, J. H. (1992). A conceptual critique of the EA:es comparison in the Comprehensive Rorschach System. *Psychological Assessment*, *4*(3), 288–296. https://doi.org/10.1037/1040-3590.4.3.288

Kleiger, J. H. (1997). Rorschach shading responses: From a printer's error to an integrated psychoanalytic paradigm. *Journal of Personality Assessment*, *69*(2), 342–364. https://doi.org/10.1207/s15327752jpa6902_7

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting Intraclass Correlation Coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lewey, J. H., Kivisalu, T. M., & Giromini, L. (2019). Coding with R-PAS: Does prior training with the Exner Comprehensive System impact interrater reliability compared to those examiners with only R-PAS-based training? *Journal of Personality Assessment*, *101*(4), 393–401. https://doi.org/10.1080/00223891.2018.1476361

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

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*(2), 219–274. https://doi.org/10.1207/S15327752JPA7802_03

Meyer, G. J., Viglione, D. J., & Giromini, L. (2014). An introduction to Rorschach-based performance assessment. In R. P. Archer & S. R. Smith (Eds.), *A guide to personality assessment: Evaluation, application, and integration* (2nd ed., pp. 301–370). Erlbaum.

Meyer, G. J., Viglione, D. J., Mihura, J. L., Erard, R. F., & Edberg, P. (2011). *Rorschach performance assessment system: Administration, coding, interpretation, and technical manual*. Rorschach Performance Assessment System.

Mihura, J. L., Dumitrascu, N., Roy, M., & Meyer, G. J. (2018). The centrality of the response process in construct validity: An illustration via the Rorschach Space response. *Journal of Personality Assessment*, *100*(3), 233–249. https://doi.org/10.1080/00223891.2017.1306781

Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing: Principles and applications* (6th ed.). Pearson Prentice Hall.

Peebles-Kleiger, M. J. (2002). Elaboration of some sequence analysis strategies: Examples and guidelines for level of confidence. *Journal of Personality Assessment*, *79*(1), 19–38. https://doi.org/10.1207/S15327752JPA7901_02

Pignolo, C., Giromini, L., Ando', A., Ghirardello, D., Di Girolamo, M., Ales, F., & Zennaro, A. (2017). An interrater reliability study of Rorschach Performance Assessment System (R-PAS) raw and complexity-adjusted scores. *Journal of Personality Assessment*, *99*(6), 619–625. https://doi.org/10.1080/00223891.2017.1296844

Piotrowski, C. (2015). On the decline of projective techniques in professional psychology training. *North American Journal of Psychology*, *17*(2), Article 259.

Rapaport, D., Gill, M., & Schafer, R. (1968). *Diagnostic psychological testing* (rev. ed.). International Universities Press.

Schafer, R. (1954). *Psychoanalytic interpretation of Rorschach testing*. Grune & Stratton.

Schneider, A. M. A., Bandeira, D. R., & Meyer, G. J. (2022). Rorschach Performance Assessment System (R-PAS) interrater reliability in a Brazilian adolescent sample and comparisons with three other studies. *Assessment*, *29*(5), 859–871. https://doi.org/10.1177/1073191120973075

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Viglione, D. J. (2010). *Rorschach coding solutions: A reference guide for the Comprehensive System* (2nd ed.).

Viglione, D. J., Blume-Marcovici, A. C., Miller, H. L., Giromini, L., & Meyer, G. (2012). An inter-rater reliability study for the rorschach performance assessment system. *Journal of Personality Assessment*, *94*(6), 607–612. https://doi.org/10.1080/00223891.2012.684118

Viglione, D. J., de Ruiter, C., King, C. M., Meyer, G. J., Kivisto, A. J., Rubin, B. A., & Hunsley, J. (2022). Legal admissibility of the Rorschach and R-PAS: A review of research, practice, and case law. *Journal of Personality Assessment*, *104*(2), 137–161. https://doi.org/10.1080/00223891.2022.2028795

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

Viglione, D. J., Meyer, G. J., Resende, A. C., & Pignolo, C. (2017). A survey of challenges experienced by new learners coding the Rorschach. *Journal of Personality Assessment*, *99*(3), 315–323. https://doi.org/10.1080/00223891.2016.1233559

Villanueva van den Hurk, A. W., Jowers, C. E., Ales, F., Keene, A., McCord, D. M., & Mihura, J. L. (2021). *When test revisions disrupt our lives, what's a psychologist to do? Updates to Mihura et al.'s (2017) survey of doctoral training in adult personality tests* [Manuscript in Preparation]. University of Dayton, Department of Psychology.

Weiner, I. B. (1998). *Principles of Rorschach interpretation*. Lawrence Erlbaum.

Wood, J. M., Nezworski, M. T., & Stejskal, W. J. (1996). The Comprehensive System for the Rorschach: A critical examination. *Psychological Science*, *7*(1), 3–10. https://doi.org/10.1111/j.1467-9280.1996.tb00658.x

Zec, S., Soriani, N., Comoretto, R., & Baldi, I. (2017). High agreement and high prevalence: The paradox of Cohen's Kappa. *The Open Nursing Journal*, *11*(1), 211–218. https://doi.org/10.2174/1874434601711010211