A Survey of Challenges Experienced by New Learners Coding the Rorschach

Published version:

DOI:10.1080/00223891.2016.1233559

Viglione, Donald J; Meyer, Gregory J.; Resende, Ana Cristina; Pignolo, Claudia. A Survey of Challenges Experienced by New Learners Coding the Rorschach. JOURNAL OF PERSONALITY ASSESSMENT. None pp: 1-9.
DOI: 10.1080/00223891.2016.1233559

A Survey of Challenges Experienced by New Learners Coding the Rorschach

Abstract

Learning to code the imagery, communication, and behavior associated with Rorschach responding is challenging. Although there is some survey research on graduate students' impressions of their Rorschach training, research has not identified which coding decisions students find to be the most problematic and time-consuming. We surveyed students to identify what they struggled with most when learning coding and to quantify how difficult it is to learn how to code. Participants (n = 191) from the United States, Brazil, Denmark, Israel, and Italy rated 57 aspects of coding using a 4-point scale that encompassed both the time required to code and the subjective difficulty of doing so. Mean ratings for coding in general indicated that students considered the overall task challenging. Ratings also revealed that students struggled most with Cognitive Special Scores, Determinants, and extrapolating from the tables to code Form Quality for objects that were not specifically listed. The findings offer suggestions about how to improve the guidelines for some of the more difficult variables and where it is most necessary to focus teaching time. Taking these steps may help the new student in learning the Rorschach.

Keywords: Rorschach; coding difficulty; training guidelines; teaching and learning

A Survey of Challenges Experienced by New Learners Coding the Rorschach

In 1974, John Exner published the first edition of the Comprehensive System (CS), with the goal of integrating the best features of the five previous systems that had been commonly used in the United States (Beck, Klopfer, Piotrowski, Hertz, and Rapaport). Based on the research available at the time and his own investigations, Exner selected for the CS the most reliable and valid components of these systems. The CS provided a systematic approach to administration and coding, a format and procedure for calculating interpreted variables, and normative samples that grew to encompass both children and adults (Exner, 2003). Eventually, the CS became the dominant system taught in graduate training (Hilsenroth & Handler, 1995; Mihura & Weinle, 2002; Ritzler & Alter, 1986).

Although the CS is no longer evolving as a result of Exner's death in 2006, the Rorschach Performance Assessment System (R-PAS; Meyer, Viglione, Mihura, Erard, & Erdberg, 2011) was developed as a replacement for it. Four of the five R-PAS authors worked with Exner on his Rorschach Research Council, which met semiannually from 1997 through 2005 to review and complete research that would advance the CS. Although Exner planned that the Research Council would take over CS developments (Exner, 1997), no formal mechanism was in place to do so when he passed away. Nonetheless, R-PAS extends the work begun by the Research Council and aims to improve the applied use of the Rorschach by, among other things, reducing examiner variability (Meyer et al., 2011), optimizing the number of responses people give to the task (Viglione et al., 2015; Pianowski, Meyer, & Villemor-Amaral, 2016), re-anchoring normative expectations to correct over-pathologizing biases (Meyer, Erdberg, & Shaffer, 2007; Meyer, Shaffer, Erdberg, & Horn, 2015), ensuring interpretation is in line with each variable's validity evidence base (Meyer, Hsiao, Viglione, Mihura, & Abraham, 2013; Mihura, Meyer, Dumitrascu, & Bombel, 2013; Mihura, Meyer, Bombel, & Dumitrascu, 2015; Mihura, Meyer, Dumitrascu, & Bombel, 2016), and making interpretation easier (Meyer & Eblin, 2012; Meyer et al., 2011). Although these changes are important, it is also the case that most of the variables coded in R-PAS are the same as variables that were coded in the CS.

According to recent survey data collected from accredited U.S. doctoral training programs in the fall of 2015 (Mihura, Roy, & Graceffo, 2016), the Rorschach is being taught in 63% of all programs, with the CS being taught in 53% and R-PAS being taught in 37%. Of the programs teaching the Rorschach, 85% cover the CS and 60% cover R-PAS. Thus, both systems are currently in active use in the U.S. Although international data comparing CS to R-PAS instruction are not available, both systems are used internationally and have been translated into other languages.

Unlike self-report measures, the Rorschach requires extensive study and supervised practice to become proficient with its administration and scoring (Gacono, Evans, & Viglione, 2008; Meyer et al., 2011). Research has demonstrated that well-trained raters can code CS and R-PAS variables with good to excellent reliability (Kivisalu, Lewey, Shaffer, & Canfield, 2016; Meyer, 2004; Meyer et al., 2002, 2011; Viglione & Meyer, 2008; Viglione et al., 2012), and that coding reliability is very similar across different languages and cultures (Meyer, Erdberg, & Shaffer, 2007). However, Viglione and Meyer (2008) summarized some CS codes from multiple studies that revealed lower (but still acceptable) reliabilities, indicating that they are more difficult to code accurately. These codes concern vague Developmental Quality (DQv and DQv/+), the Form Dominance of color and shading variables (FC vs. CF vs. C and Form Shading vs. Shading Form vs. Shading), Form Quality (FQu and FQ+), certain Contents (Art, Ay, Sc, Bt vs. Na vs. Ls, Id), and Special Scores (DV vs. INC, ALOG, CONTAM vs. INC, PER vs. DR, Level 1 vs. Level 2). There is less data available concerning R-PAS codes, though similar instances of lower reliability have appeared for at least some of the same variables when coded in R-PAS (e.g., Vagueness, FQu%, Cognitive Codes; see Kivisalu et al., 2016; Viglione et al., 2012).

Several CS studies have investigated coding accuracy and the inter-rater reliability of coding categories[1] among students and new learners. Hilsenroth, Charnas, Zodan, and Streiner (2007) examined coding accuracy among 29 graduate students enrolled in an APA-approved clinical PhD program. The authors found an agreement of 80% or more with most of the coding categories (i.e., Location, Developmental Quality, Form Quality, Pair, Content, Popular) but lower rates of agreement for Determinants (78%) and Special Scores (65%). Similarly, estimated kappa was less than .74 for Determinants, Form Quality, Z-scores, and Special Scores. The latter coding category was the only category that showed an estimated kappa in the fair to good range of reliability (estimated $\kappa = .56$); all other estimated kappa values were higher.

Callahan (2015) evaluated coding accuracy of CS protocols through a three-stage training experience, followed by an eight-week follow-up. The accuracy of coding all the Rorschach response segments improved over time, though the proportions of agreement

[1] In reporting our findings, we used "coding category" to refer to response segments and "coding decision" to refer to distinction between one code and another or the presence or absence of an individual code. Thus, any coding decision (e.g., Bt vs. Na vs. Ls, presence or absence of Art) occurs within a coding category (in this example, within the Content category).

with the expert scoring were generally lower for FQ (68.9%) and Special Scores (65.3%) at the 8-week follow-up protocol. Guarnaccia, Dill, Sabatino, and Southwick (2001) investigated the association of training and experience with coding accuracy using a small sample of responses. Twenty-one second level graduate students and 12 licensed psychologists coded 10 responses from clinical protocols and 10 responses from nonclinical protocols. The results showed significant but somewhat inconsistent differences in scoring accuracy. Students were more accurate for Contents (non-clinical responses) and DQ (clinical responses), whereas professionals were more accurate for FQ, Special Scores (non-clinical responses), and Contents (clinical responses). Although Popular and Pairs achieved scoring accuracy above 80% for both students and professionals, FQ and Special Scores were more difficult to score correctly for all participants.

With respect to R-PAS coding, the effects of training have not been studied extensively. However, Meyer et al. (2011) examined interrater reliability for six codes that were new to R-PAS relative to the CS (Space Reversal, Space Integration, Aggressive Content, Oral Dependency Language, Mutuality of Autonomy Health, and Mutuality of Autonomy Pathology). Six coders each independently coded a set of 50 protocols from the R-PAS normative sample. The coders varied in their previous experience coding Rorschach protocols, ranging from being highly experienced to having coded just one protocol before the study began. However, all coders were applying the draft R-PAS coding guidelines for the first time (and the final guidelines were improved and clarified by the coding challenges they encountered). Across the six codes, the average of the pairwise reliability coefficients was ICC = .81. However, for the three most experienced coders the average ICC was notably higher at .87. The other studies systematically examining R-PAS coding reliability have relied on doctoral students as the coders (Kivisalu et al., 2016; Viglione et al., 2012) but have not compared students to more experienced coders.

Taken together, these studies suggest that students and new learners show lower reliability in general than more senior coders, as well as lower accuracy with FQ, Determinants, and Special Scores, but they do not provide definitive information regarding the minimum amount of training or experience required to code reliably. Moreover, these data suggest that there are complexities in the coding process that may require further investigation, that there are coding guidelines that would benefit from further specification, and that certain coding decisions might warrant more training time.

Difficulties in coding the Rorschach accurately may be due to unclear definitions of codes that are not fully specified in the standard CS training materials (Exner, 2001; Exner, 2003). The brevity of these materials prompted Viglione (2002, 2010) to write *Coding Solutions*, which is a detailed text for coding according to CS guidelines, and similar levels of detailed guidance were incorporated into the R-PAS manual (Meyer et al., 2011). Personal experience from the first author's more than 30 years in teaching CS coding reveals that many new students struggle with confusing abstract rules and exceptions to those rules when learning to code. Possible perceptions that the guidelines are arbitrary, insufficient or less than helpful, and too difficult or too time-consuming all can undermine learning and prevent students from sustaining sufficient effort to learn how to code accurately. Thus, new learners who approach the Rorschach for the first time and graduate students who are still in training may offer valuable input on the specific scoring difficulties they encounter that might undermine their confidence of becoming a reliable coder.

The present study is the first attempt to investigate student perspectives about difficulties they encounter in learning Rorschach coding. Data collection began in 2007 (Ptucha, Viglione, & Meyer, 2008), at a time long before R-PAS was introduced (Meyer et al., 2011) when a subgroup of the R-PAS authors was considering ways to make changes to the CS. Given this, CS variables were the sole focus of the investigation then, and they are reported in the current study. Only later, well after this research was initiated, did it become clear that revisions to the CS would be impossible, which ultimately led to R-PAS being created. Some of the findings from the present study ultimately contributed to decisions that were made when creating R-PAS, most notably by dropping some coding categories and distinctions and by providing more elaborated coding instructions akin to those found in *Coding Solutions* (Viglione, 2010). However, a systematic examination of the survey results has never been published. The findings to be reported have clear relevance to training programs that continue to teach the CS, of which there still are many. In addition, to the extent that survey results identify coding challenges that are intrinsically present when coding Rorschach-based perceptions, communications, and behavior despite the elaborated guidelines that are available for the CS (Viglione, 2010) and R-PAS (Meyer et al., 2011), the results will apply to the training of both CS and R-PAS students.

Nonetheless, the purpose of our survey was to discover what students struggle with the most when learning to code the Rorschach according to the CS. Answers to these concrete questions may inform more abstract concerns pertaining to the accessibility of the test for new learners and the practical barriers for coding reliably. Moreover, we also investigated whether more experience was associated with less coding difficulty. Findings might help to identify codes that require more instructional time, more detailed guidelines, and more practice calibrating to standards to achieve mastery. Ultimately, such information could conceivably increase the number of students who become proficient and use the task, as well as increase research on the Rorschach.

Method

New Learner Survey

The New Learner Survey is a 57-item measure that was developed for the present study. The coding challenges selected for the survey were largely derived from Viglione's CS coding text (2002, 2010). The topics addressed in that text were selected by tracking coding inconsistencies among multiple coders examining the same responses and by identifying common coding errors made by students in training. Most of the survey items are oriented towards common coding distinctions one must make (e.g., FT vs. TF vs. T), as opposed to rating the presence or absence of individual codes, (e.g., T vs. No T). The surveys were self-administered, and items were listed in the same order as they are encountered when coding a response using the CS, starting with Location and Developmental Quality, then moving on to Determinants, Form Quality, Pairs, Contents, Popular, and Special Scores. However, the survey began with a single item asking about difficulty learning coding for the Rorschach as a whole. Given that our aim was to investigate what codes new learners struggle with the most, we asked raters to evaluate their experience subjectively through introspection. For example, one item asked about experienced difficulty coding Location in general, while other items asked about decisions between W, D, and Dd. Students were asked to rate each item on a 4-point scale of difficulty, where 1 = “Simple, straightforward, and very easy to score. Takes very little time to score.”, 2 = “Understandable and relatively easy to score but does take some work at times. Takes a little time and effort to score.”, 3 = “Sometimes a challenge and sometimes confusing. Often consumes a significant amount of scoring time.”, 4 = “Complicated, confusing, and difficult to score. Consumes a great deal of scoring time.” The New Learner Survey rating scale deliberately targeted the two intertwined components that make coding challenging: (1) how difficult and confusing the coding experience is and (2) how much time it takes to execute a decision during the coding process. As a result, a score of 1 indicates that the coding category or decision is simple and takes very little time to code, while a score of 4 indicates that the coding category or decision is very difficult and requires a great deal of time to code. We created scale anchors that emphasized both relative difficulty and relative time because a simple decision typically can be made quickly and easily, whereas a difficult decision frequently requires additional time to sort through multiple facets of more complex coding criteria.

Participants

The New Learner Survey was administered to psychology graduate students who were in training under the supervision of or in classes with psychologists. Our aim was to investigate the opinion of beginning learners regarding the difficulties they encounter when coding Rorschach protocols. We were interested in the opinion of both new learners (e.g., graduate students who were attending their first Rorschach class) and students who already completed their first Rorschach semester but who were still in training. Participant surveys were gathered internationally from multiple sites across the

United States, as well as sites in Brazil, Denmark, Israel, and Italy. Students in Denmark, Israel, and Italy completed the surveys in English, while those in Brazil completed a Portuguese version of the survey. Overall, 207 psychology graduate students and trainees completed the survey. Approximately half of the contributors were from the US (56%) and half from international locations. Because absolute beginners may have limited knowledge about Rorschach codes, we excluded participants who had coded fewer than two protocols (5 students) or who did not indicate how many protocols they had coded (5 students). Thus, no absolute beginners were included in the analyses. At the other end of the continuum of scoring experience, students with considerable experience were excluded. Operationally, this was defined as omitting the 6 students who had coded 45 or more protocols, which placed them above the 97th percentile of coding experience.

As a result, the final sample consisted of 191 participants. The majority of the student participants (69.1%) had already completed at least one semester of Rorschach instruction, and the other participants were attending their first Rorschach class. Most of the participants had coded more protocols than they had administered themselves. The median number of coded protocols was 8 with a mean value of 12.1 (*SD* = 9.3; Range 2 - 40), whereas the median number of administered protocols was 6 with a mean value of 10.4 (*SD* = 9.0, Range 2 - 33). Absolute skew and kurtosis were lower than 1.1 (see below), so that these variables were normally distributed. The large standard deviations revealed that there was considerable variability in the number of records administered and coded. Because our aim was to investigate new learners' judgments, we also address the potential effect of experience on their judgments. In addition, because students who had not yet completed their first course of Rorschach instruction may not be able to provide informed responses about coding difficulty, we conducted control analyses that were limited to the 132 students who had already completed at least one semester of Rorschach training.

Statistical Procedures

To address the normality of the distributions for the 57 survey variables with this relatively large sample, we considered the cut-off suggested by West, Finch and Curran (1995) of 2.0 for skew and 7.0 for kurtosis to identify a moderate departure from normality. Fifty-two variables had reasonably normal distributions with an average absolute value for skew of .536 (absolute value range: .006 – 1.817) and an average absolute value for kurtosis of .457 (absolute value range: .097 – 2.314). Five variables showed a non-normal distribution [i.e., W vs. D, Pairs, H vs. (H), (H & A) vs. (Hd & Ad), and Popular].

To establish whether each mean rating for a “target” variable was higher or lower than the “overall mean” across all the ratings, we computed an overall mean using all the variables except the target variable being investigated. This is analogous to computing part-whole correlations after omitting the “part” from the “whole.” For example, we compared the mean rating of the target W vs. D coding decision to the overall mean of the ratings for the other 56 variables excluding the rating for W vs. D. We repeated this procedure for all the other 56 variables. Next, the target and overall means were compared by computing a paired-samples t-test for all the items that were normally distributed and a sign test for those that violated the assumption of normality (Table 1). The “overall mean” that was the comparison point for each of the 57 rated items had an average across all 57 items of 1.92 and a range across each specific item from 1.90 to

1.93. Each of the SDs for these overall means was in the range between 0.36 and 0.37. Because these overall Ms and SDs are so similar, they are not separately reported in Table 1. However, they form the comparison point with the target Ms and SDs, and they are used to generate the *t*-tests, *p* values, and Cohen's *d* values that also are reported in Table 1. Given the multiple comparisons, we decided to apply a correction for alpha using the conservative Bonferroni correction: alpha of .05 was divided by 57, so that a difference is considered significant if *p* is less than .0009.

Cohen's *d* was computed to evaluate the magnitude of the difference between each target coding category and its comparison overall mean score. Even though these were paired samples data, we followed statistical recommendations by Dunlap, Cortina, Vaslow, and Burke (1996) and used the standard *d* formula of computing mean differences and dividing by the pooled SD in order to document how far apart the two sets of mean values were in SD units. Consistent with characterizations in Cohen (1988), values around |.20| would indicate a small effect size, |.50| a medium effect size, and |.80| a large effect size. However, it is important to keep in mind that the *d* values reported here are different than typical *d* values because we are comparing each target variable mean to an overall mean, which is a grand mean computed across all the other rated variables, rather than to other individual target means. For instance, in the data to be presented comparing the item that was rated most difficult to code (Cognitive Special Scores) to the item rated least difficult to code (Pairs) would produce a *d* value of about |3.0|, though the *d* values we report in Table 1 comparing each of these items to their counterpart overall means is about |2.0|.

Finally, Spearman correlations were used to evaluate the relationship between the coding difficulty judgments expressed by the participants and their level of experience. For indicators of experience, we used the number of protocols scored, the number of protocols administered, and whether they had completed a semester of training.

Results and Discussion

Coding a Rorschach record appears to be moderately difficult for new learners: About 59% of the students rated the overall difficulty in coding the Rorschach as "Sometimes a challenge and sometimes confusing. Often consumes a significant amount of scoring time" (a score of 3 in the rating scale). The mean rating for coding the CS as a whole of 2.70 (SD = .73) is slightly lower than this benchmark score. However, averaging ratings across all 57 scores produced a mean difficulty rating of 1.92 (SD = .48). The large and significant difference between the mean rating of the Rorschach as whole and the overall mean across all the individual target ratings suggests that new learners experience the challenge of scoring all the codes for a full protocol as considerably more difficult than the average, single coding decision. From a different perspective, coders may experience an additive effect of difficulty across codes and responses.

Reviewing Table 1, it can be seen that 21 coding decisions and categories have a mean rating over 2, and Cognitive Special Scores has a mean rating greater than 3. When separated into each coding category (i.e., Location, Determinants, etc.) and compared with the overall mean, the category with the highest, that is most difficult, rating is Cognitive Special Scores (M = 3.02, d = 2.05), followed by Determinants (M = 2.52, d = 1.06), and then Other Special Scores (M = 2.32, d = 0.68). Because these three coding categories were rated as the most difficult, we discuss their individual coding decisions more fully.

As the most difficult and time-consuming coding category 79% of participants rated Cognitive Special Scores as challenging (3) or complicated (4). All the individual coding decisions within this category have a mean rating significantly higher than the overall mean, with large to very large effect sizes relative to the overall mean of all remaining items (*d*s from 1.01 to 1.79). Thus, not only is the Cognitive Special Scores as a group the most complicated coding category that consumes a great deal of scoring time, but also within this group the primary coding decisions we asked about were all essentially equally difficult.

Determinants are the second most difficult category overall ($M = 2.52$, $d = 1.06$) with a large effect size. Comparing each sub-category with its overall mean (Table 1), four areas have significantly higher mean scores with medium to large effect sizes: (1) deciding which shading subtypes to code ($d = 1.28$); (2) deciding between dimensionality based on form (FD) and dimensionality based on shading (V; $d = 1.03$); (3) deciding on the degree of Form Dominance (*d*s from 0.74 to 1.22); and (4) differentiating between Diffuse Shading (Y) and Achromatic Color (C'; $d = 0.70$). Among the remaining Determinants, Reflections, the presence or absence of Color, and the distinction between Human (M), Animal (FM), and Inanimate (m) Movements were reported as 1 (*simple*) or 2 (*relatively easy*) to code by more than 90% of the students. Moreover, their means are significant lower compared to the overall mean. Thus, as a whole, determinants are challenging to code and consume significant time with decisions among shading subtypes, FD vs. V, and form dominance for both color and shading responses taking the most time and effort. In contrast, coding reflection, color vs. not, and distinctions between M, FM, and m are considered easy to code.

The category of Other Special Scores, which could be considered to be thematic codes rather than cognitive codes, was rated more than a half standard deviation above the overall mean of all the variables (M = 2.32, d = 0.68). In addition, the mean rating of the special score Abstract (AB; M = 2.15, d = 0.38) is slightly higher than the overall mean score, but Aggressive Movement (AG; M = 1.63, d = -0.55), Cooperative Movement (COP; M = 1.59, d = -0.63), and Morbid (MOR; M = 1.65, d = -0.50) have a mean rating significant lower than the overall mean score. Furthermore, most of the students rated each individual coding decision as relatively easy to code. Given these results, it remains unclear which coding decisions contribute to the perceived relative high difficulty rating of this general coding category. To speculate, it may be due to the overall number and diversity of themes of the variables involved.

Form Quality (FQ; M = 1.98, d = 0.11), is the next most difficult to code but not significantly more than the survey mean. FQ as a category was rated with a 3 (*challenging*) or 4 (*complicated*) by only the 19% of the students. It is worth noting that the mean rating of extrapolation for objects not listed in the FQ tables (M = 2.51, d = 1.01) is significantly higher than the overall mean and among the highest of the individual coding decisions. Indeed, 50% of participants found this procedure challenging or difficult and time consuming. This suggests that within the FQ coding category extrapolation may account for a large part of the challenges for new learners. Simplifying and providing guidance for this process might improve utility and help new learners. One might also wonder whether extrapolation difficulties partially account for the finding that the inter-rater reliability of FQ coding is often low, particularly for FQu.

New learners, however, perceived many scores as simple to code. The mean rating of the coding decisions related to Location (*M* = 1.63, *d* = -0.60), Pairs (*M* = 1.15, *d* = -1.99), Content (*M* = 1.67, *d* = -0.651), and Popular (*M* = 1.18, *d* = -1.82) are significantly lower than the overall mean score. For Location, coding D vs. Dd for "near Dd responses" (*M* = 2.12) is the only coding decision that has a mean rating greater than the overall mean, but the Cohen's *d* of -0.37 indicates a small to medium effect size. Interestingly, most of the Contents display medium to large differences in mean scores (Table 1), and all are lower than the overall mean score except for Idiographic contents (Id; *M* = 2.03, *d* = .18). Developmental Quality (DQ; *M* = 1.82, *d* = -0.19) and the Z-scores (*M* = 1.84, *d* = -0.12), did not show any significant difference compared to the overall mean score (Table 1), suggesting that they are in the average range of difficulty.

To address experience and training, we calculated Spearman, rank order correlations to analyze the relationship of coding difficulty ratings with three measures of experience and training: (1) the number of protocols scored, (2) the number of protocols administered, and (3) whether or not one had completed a first Rorschach course (see the final columns in Table 1). As expected, the correlations were predominantly negative with only one significantly positive correlation (completing a semester with Pairs, rho = .16). Thus, in general more experience is associated with lower ratings of CS coding difficulty.

The number of scored protocols produced 19 significant correlations with the coding categories and decisions ($p < .05$; rho > .14). Also, for 39 out of 57 items, correlation values for the number of scored protocols were greater than those obtained by the number of administered protocols (4 out of 57) and having completed the first Rorschach class (8 out of 57). It makes sense that practicing coding is the key component in the reduction of confusion encountered and time spent coding. The key findings for the number of scored protocols are the mean of the 57 correlations (rho = -.21) and the correlation with the Rorschach as a whole (rho = -.26) with a small to medium effect size. Thus, the biggest effect on ease of CS coding derives from the number of Rorschach protocols scored. Nevertheless, 38 of 57 correlations with the number of protocols scored were not significant, indicating that the effect was not detected for two thirds of the coding difficulty ratings. In part, this may be due to our decision to omit the most and least experienced student coders from the sample.

Focusing on the associations with the number of protocols coded, it is difficult to identify patterns for those most or least affected by experiences. However, it does appear that Cognitive Special Scores, Form Quality, Form Dominance, Pairs, and Other Special Scores are least influenced by experience. Alternatively, for Location, other Determinant Codes, most Content decisions, and Z-Scores, practice appears to be most helpful in the experience of new learners.

Location and Content were easier to code compared to the overall mean score, and it seems that practice makes them even easier to code. Pairs and Popular did not improve with practice, but the data suggest that they are easy to code at the beginning so that practice does not make it easier. However, FQ and Special Scores are more challenging categories, and training was not found to be particularly helpful. For these categories, better guidelines and instruction may be needed or even a reformulation of the coding processes themselves.

Finally, we recomputed mean ratings of difficulty after excluding students who were still completing their first semester of coursework. Limiting the results to the 132 students who had at least a semester of Rorschach experience produced no noticeable difference in the findings. The mean ratings showed essentially the same distribution as before across the 57 items (*M* = 1.90 vs. 1.92, *SD* = 0.46 vs. 0.47, Min = 1.15 vs. 1.15, Max = 2.98 vs. 3.02) and the correlation of the mean ratings in the refined sample with the mean ratings in the full sample was .9959. Similarly, the effect sizes obtained in the smaller but more experienced subsample were essentially unchanged relative to the effect sizes obtained in the full sample, with the correlation between these values being .9955.

Conclusions and Implications

For the New Learner Survey, the students learning the Rorschach judged each coding category and decision to identify those that are difficult, challenging, and time-consuming. Overall, our results suggest that the students consider the CS as a whole to be more difficult and time-consuming to code than most of the single coding categories or decisions in the system, thus possibly experiencing an additive effect across codes in the challenge to learn coding. The students struggle most with Cognitive Special Scores, Determinants (particularly with shading subtypes, FD vs. V, and form dominance), and extrapolation for objects not listed in the FQ tables. Location, Pair, Content, and Popular seem to be easy to code, and DQ and Z-scores had mean ratings in the average range. Finally, we found no evidence that the degree of experience as measured by the number of scored protocols, the number of administered protocols, or by having completed the first Rorschach class, influenced new learners' judgment about the perceived difficulty of Rorschach coding decisions. However, before accepting the relative independence of experience from perceived coding difficulty, further research with sufficient power would be necessary. In addition, we did not investigate the relationship between experience and accuracy (e.g., Callahan, 2015), so, despite our findings, it is highly likely that accuracy improves with experience.

The results of the present study raise interesting considerations about the interrater reliability of the CS and possibly extending to R-PAS. To address this, one can compare the New Learner Survey results with the results reported in the literature. Doing so, considerable similarity is found between the codes perceived as most difficult and time-consuming by the new learners and the codes that have lower interrater reliability. Based on findings in the literature, Viglione and Meyer (2008) indicated that form dominance, shading subtypes, Cognitive Special Scores, and the distinction between Level 1 and Level 2 cognitive codes were the coding decisions with lower reliabilities. Lower inter-rater reliabilities for Determinants, FQ, and Special Scores were reported also by Hilsenroth et al. (2007), Guarnaccia et al. (2001), and Callahan (2015) for the CS and by Kivisalu et al. (2016) for R-PAS. These coding decisions also had higher mean ratings of difficulty in the New Learner Survey. Interrater reliability statistics measure how well two raters agree in coding the same Rorschach response. The fact that new learners find some coding decisions particularly confusing and time-consuming probably increases the probability of disagreement in coding the same Rorschach variable. That is, if a coding decision is not clear and obvious, the coders could use different strategies to resolve the dilemma and those strategies may not be shared among different sites or different coders.

The importance of understanding which Rorschach variables are difficult to code by new learners can also provide guidance for improving teaching methods and identifying which codes need more detailed coding guidelines and practice materials. The findings from the present study may, therefore, suggest how to reduce scoring time in practice and how to improve the training materials and the guidelines for some of the more difficult variables and, in doing so, improve the utility of the Rorschach from a new student's perspective. Also, the correlations suggest that coding records and responses rather than administering records is the key variable in making coding easier for students.

DQ, as well as Pairs, Content, Popular, and Z-Scores were rated as easy or in the average range of difficulty, so that these categories may not need to be improved in training materials. Location as a category and its subcomponents were perceived as relatively easy to code and were generally correlated with experience. The correlation with experience suggests that training guidelines should focus on exercises in coding location with threshold examples to help student differentiate W vs. D, D vs. Dd in multi-object responses, and D vs. Dd locations in "near Dd responses".

Related to Determinants, students identified form dominance, shading subtypes, and distinctions between Y vs. C' and FD vs. Vista as the coding decisions that were more difficult to learn. Given the results of the correlations, it seems that more coding exercises may help students to become confident in distinguishing among shading subtypes, whereas more teaching time and threshold examples may lead to a better understanding of how to code the degree of form dominance and how to differentiate the two depth codes (i.e., FD vs. Vista). Although differentiating whether movement was active vs. passive was not rated as difficult, practice coding made it easier.

Students struggle with the process of extrapolation to determine FQ for objects not in the FQ tables, and experience does not help students in feeling confident about FQ codes. One specific suggestion would be to present a systematic method for FQ Extrapolation, outlining via numbered steps or a flow chart all the possibilities that students may encounter during the procedure. The same recommendations apply to Cognitive Special Scores, for which additional experience may not be helpful in reducing difficulty. Among Other Special Scores, results from the present study suggest that AB is a bit harder to learn than average and requires more teaching time. Additional guidelines and benchmark examples are likely to help with this code. In contrast, students considered AG, COP, and MOR as relatively simple to code and to learn. Additional guidance for FQ extrapolation, differentiating cognitive codes, and classifying AB has been provided by Viglione (2010) for the CS and by Meyer et al. (2011) for R-PAS.

This study focused on the CS, and some of its findings ultimately contributed to refinements in an advanced coding guide for the CS (Viglione, 2010). In addition, some of the findings contributed to decisions made when creating R-PAS (Meyer et al., 2011) once it became clear that the CS source materials (Exner, 2003) would not be revised (e.g., elaborated guidelines; no longer coding form dominance for Y, T, V, C', and r; identifying ambiguous location boundaries to aid D versus Dd Location decisions; practice coding responses accompanied by commentary to illustrate benchmark standards). On the other hand, R-PAS added a number of thematic codes, differentiated ways of using background white space, and added codes related to administration behaviors. These additions might offset some of the other improvements by introducing new coding challenges.

Although refined coding guidelines, simplified coding requirements, and structured coding examples will help make it easier to learn how to code Rorschach responses, it should be recognized that some of the difficulties identified in this survey will remain difficulties for any multi-variable system of coding. In part this is because there is an almost limitless degree of unique attributions, communications, and behaviors that can be exhibited by respondents when giving responses. In part this also is because there is an irreducible degree of uncertainty associated with assigning particular kinds of codes, including the task of assessing the visual fit of images to inkblot locations based on verbal and nonverbal communications (Form Quality) and the task of classifying dimensions of disordered thought and impossible or implausible imagery in inherently confusing thought disordered communications (Cognitive Special Scores). Thus, coding the behaviors observed in the structured context of Rorschach responding will never be as simple as tallying responses to a self-report questionnaire or evaluating the correctness of verbal and nonverbal responses to a cognitive assessment measure. Alerting students to the genuinely complex demands of coding Rorschach task behaviors will not erase the challenges but it may heighten the rewards that come from mastering such a skill.

Although the present study adds important new suggestions about CS training material and describes the new learners' point of view, it has some limitations. First, the only participant information we collected was country of origin, completion of a semester-long Rorschach course, and the number of protocols coded and administered. The study did not access other demographic information or potentially relevant training information, for example age, hours of supervision, type of class attended, or whether records were administered and coded for training, clinical practice, or research. In addition, we did not determine to what extent students relied on the relatively brief standard training materials (Exner, 2001, 2003) versus supplemental material designed to make CS coding easier and more reliable (Viglione, 2002, 2010). These limitations may reduce the generalizability and validity of the findings, even though our study does provide practical implications for teaching and test utility. Secondly, when we developed the New Learner Survey, we used a subjective frame of reference in the form of asking students to report their experience of the difficulty and time involved in coding. It is not clear how the findings might be different if time and difficulty were separated in the survey. Future research might isolate the perceived difficulty of each coding decision and assess its relation to a more "objective" measure of time. Asking students about the amount of time they have actually spent in learning the Rorschach codes during their training period or the amount of time actually spent on specific coding decisions are ways of addressing time. Finally, our study did not investigate how students perceive R-PAS coding decisions – on their own or compared to the CS. It would be interesting to replicate this study with R-PAS.

References

Callahan, J. L. (2015). Evidence-based technical skills training in pre-practicum psychological assessment. *Training and Education in Professional Psychology, 9*, 21-27. doi: 10.1037/tep0000061

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*, 170-177. doi: 10.1037/1082-989X.1.2.170

Exner, J. E. (1974). *The Rorschach: A Comprehensive System*. New York: Wiley.

Exner, J. E. (1997). Rorschach Workshops and the future. *1997 Alumni Newsletter*. Asheville, NC: Rorschach Workshops, July 7.

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

Exner, J. E. (2003). *The Rorschach: A Comprehensive System* (4th ed.). New York: Wiley.

Gacono, C. B., Evans, F. B. III, & Viglione, D. J. (2008). Essential issues in the forensic use of the Rorschach. In C. B. Gacono, F. B. Evans, N. Kaser-Boyd & L. A. Gacono (Eds.), *The handbook of forensic Rorschach assessment* (pp. 3-20). New York, NY: Routledge.

Guarnaccia, V., Dill, C. A., Sabatino, S., & Southwick, S. (2001). Scoring accuracy using the Comprehensive System for the Rorschach. *Journal of Personality Assessment, 77,* 464-474. doi: 10.1207/S15327752JPA7703_07

Hilsenroth, M. J., Charnas, J. W., Zodan, J., & Streiner, D. L. (2007). Criterion-Based Training for Rorschach Scoring. *Training and Education in Professional Psychology, 1*, 125-134. doi: 10.1037/1931-3918.1.2.125

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*, 382-390. doi: 10.1080/00223891.2015.1118380

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: Vol. 2. Comprehensive handbook of psychological assessment* (pp. 315-342). Hoboken, NJ: Wiley.

Meyer, G. J. & Eblin, J. J. (2012). An overview of the Rorschach Performance Assessment System. *Psychological Injury and Law, 5*, 107-121. doi: 10.1007/s12207-012-9130-y

Meyer, G. J., Erdberg, P., & Shaffer, T. W. (2007). Towards international normative reference data for the Comprehensive System. *Journal of Personality Assessment, 89,* S201-S216. doi: 10.1080/00223890701629342

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. doi: 10.1207/S15327752JPA7802_03

Meyer, G. J., Hsiao, W.-C., Viglione, D. J., Mihura, J. L., & Abraham, L. M. (2013).

Rorschach scores in applied clinical practice: A survey of perceived validity by experienced clinicians. *Journal of Personality Assessment, 95*, 351-365. doi: 10.1080/00223891.2013.770399

Meyer, G. J., Shaffer, T. W., Erdberg, P., & Horn, S. L. (2015). Addressing issues in the development and use of the Composite International Reference Values as Rorschach norms for adults. *Journal of Personality Assessment, 97*, 330-347. doi: 10.1080/00223891.2014.961603

Meyer, G. J., Viglione, D. J., Mihura, J. L., Erard, R. E., & Erdberg, P. (2011). *Rorschach Performance Assessment System: Administration, coding, interpretation, and technical manual.* Toledo, OH: Rorschach Performance Assessment System.

Mihura, J. L., Meyer, G. J., Bombel, G., & Dumitrascu, N. (2015). Standards, accuracy, and questions of bias in Rorschach meta-analyses: Reply to Wood, Garb, Nezworski, Lilienfeld, and Duke (2015). *Psychological Bulletin, 141*, 250-260. doi: 10.1037/a0038445

Mihura, J. L., Meyer, G. J., Dumitrascu, N., & Bombel, G. (2013). The validity of individual Rorschach variables: Systematic reviews and meta-analyses of the Comprehensive System. *Psychological Bulletin, 139*, 548-605. doi: 10.1037/a0029406

Mihura, J. L., Meyer, G. J., Dumitrascu, N., & Bombel, G. (2016). On conducting construct validity meta-analyses for the Rorschach: A reply to Tibon-Czopp and Zeligman (2016). *Journal of Personality Assessment, 98*, 343-350. doi: 10.1080/00223891.2016.1158182

Mihura, J. L., Roy, M., & Graceffo, R. A. (2016). *Psychological assessment training in APA-accredited clinical psychology doctoral programs*. Manuscript submitted for publication.

Mihura, J. L., & Weinle, C. A. (2002). Rorschach training: Doctoral students’ experiences and preferences. *Journal of Personality Assessment, 79,* 39-52. doi: 10.1207/S15327752JPA7901_03

Pianowski, G., Meyer, G. J., & Villemor-Amaral, A. E. (2016). The impact of R-Optimized Administration modeling procedures on Brazilian normative reference values. *Journal of Personality Assessment, 98*, 408-418. doi: 10.1080/00223891.2016.1148701

Ptucha, K., Viglione, D. J., & Meyer, G. J. (2008, March). *New Learners’ Survey: Experience and Challenges in Learning Comprehensive System Coding*. Paper presented at the Society for Personality Assessment Annual Conference, New Orleans, LA.

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

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

Viglione, D. J., Blume-Marcovici, A. C., Miller, H. L., Giromini, L., & Meyer, G. J. (2012). An initial inter-rater reliability study for the Rorschach Performance Assessment System. *Journal of Personality Assessment, 94*, 607-612. doi: 10.1080/00223891.2012.684118

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., Meyer, G., Jordan, R. J., Converse, G. L., Evans, J., MacDermott, D., & Moore, R. (2015). Developing an alternative Rorschach administration method to optimize the number of responses and enhance clinical inferences. *Clinical Psychology & Psychotherapy*. *22,* 546-558. doi: 10.1002/cpp.1913

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with non-normal variables: Problems and remedies. In R. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56-75). Thousand Oaks, CA: Sage.

Table 1. Descriptive statistics, rating percentages, and paired-sample t-test results comparing each of the 57 target coding categories and decisions to the overall mean of the other 56.

					Ratings (%)							***rho***		
	Rank	***M***	***SD***	***Mdn***	**1**	**2**	**3**	**4**	***t*(190)**	***p***	**Cohen's *d***	***# scored prot***	***# admin. prot***	***Completed semester***
Rorschach as a whole	3	2.70	.73	3	7	26	59	9	15.62	< .001[a]	1.40	-.26**	-.19*	-.18*
Location as a group	**42**	**1.63**	**.62**	**2**	**44**	**50**	**6**	**1**	**-7.38**	**< .001[a]**	**-0.60**	**-.28****	**-.27****	**-.25****
W vs. D	55	1.20	.45	1	82	16	2	0	-11.83[b]	< .001[a]	-1.77	-.29**	-.23**	-.20**
D vs. Dd for multiple objects responses	28	1.86	.72	2	32	51	15	2	-1.14	.255	-0.10	-.25**	-.21**	-.23**
D vs. Dd for "near Dd" responses"	18	2.12	.71	2	18	55	25	2	4.28	< .001[a]	0.37	-.31**	-.22**	-.24**
DS vs. DdS	34	1.82	.72	2	36	48	16	1	-2.08	.039	-0.18	-.10	-.08	-.04
DQ as a group	**36**	**1.82**	**.74**	**2**	**36**	**49**	**13**	**2**	**-2.29**	**.023**	**-0.19**	**-.16***	**-.15***	**-.06**
Evaluating synthesis: (DQ+ or v/+ vs. DQo or v)	30	1.84	.72	2	33	52	13	2	-1.61	.108	-0.14	-.19**	-.19**	-.13
Evaluating form demand: (DQo or + vs. DQv or v/+)	27	1.87	.75	2	33	48	17	2	-0.98	.328	-0.09	-.13	-.11	.04
Determinants as a group	**7**	**2.52**	**.78**	**3**	**8**	**42**	**40**	**10**	**11.62**	**< .001[a]**	**1.06**	**-.19***	**-.18***	**-.25****
M vs. FM vs. m	45	1.53	.65	1	55	38	7	1	-9.55	< .001[a]	-0.78	-.12	-.04	-.01
Active vs. passive	21	2.02	.78	2	26	49	22	3	1.98	.049	0.17	-.22**	-.18*	-.09
Color vs. No Color	50	1.36	.57	1	69	26	5	0	-13.81	< .001[a]	-1.22	-.10	-.16*	-.09
FC vs. CF vs. Pure C	15	2.32	.75	2	13	46	37	4	8.71	< .001[a]	0.74	-.10	-.06	-.04
Shading subtypes: Y vs. T vs. V	5	2.65	.79	3	7	35	46	13	14.65	< .001[a]	1.28	-.23**	-.21**	-.12
Y vs. C'	14	2.33	.82	2	16	41	37	6	8.32	< .001[a]	0.70	-.20**	-.13	-.10
C' vs. No C'	32	1.83	.76	2	37	44	17	2	-1.92	.056	-0.16	-.10	-.04	.05
FY vs. YF vs. Pure Y	11	2.48	.71	2	6	46	42	6	13.51	< .001[a]	1.07	-.05	.00	-.05
FT vs. TF vs. Pure T	12	2.39	.75	2	10	48	36	6	11.35	< .001[a]	0.88	-.06	-.01	-.01
FV vs. VF vs. Pure V	6	2.54	.68	3	4	44	46	6	15.87	< .001[a]	1.22	-.14	-.07	-.06
FC' vs. C'F vs. Pure C'	13	2.33	.76	2	13	46	36	5	9.49	< .001[a]	0.75	.02	.08	.07
Depth: FD vs. Vista	10	2.51	.79	3	10	38	43	9	11.63	< .001[a]	1.03	-.12	-.10	-.12
Reflections	49	1.41	.63	1	66	26	7	0	-11.27	< .001[a]	-1.04	.05	.03	.08

					Ratings (%)							rho		
	Rank	*M*	*SD*	*Mdn*	**1**	**2**	**3**	**4**	***t*(190)**	***p***	**Cohen's *d***	***# scored prot***	***# admin. prot***	***Completed semester***
FQ as a group	**23**	**1.98**	**.70**	**2**	**24**	**57**	**17**	**2**	**1.33**	**.184**	**0.11**	**.00**	**-.03**	**.06**
FQo vs. (FQu & FQ-)	35	1.82	.79	2	39	41	17	2	-2.00	.047	-0.17	-.02	.01	.07
FQu vs. FQ-	19	2.07	.81	2	25	48	23	4	2.83	.005	0.26	.01	-.03	.11
Extrapolation for objects not in FQ Table	9	2.51	.82	3	10	40	39	11	11.56	< .001[a]	1.01	-.13	-.10	-.05
Pairs	**57**	**1.15**	**.42**	**1**	**87**	**11**	**2**	**0**	**-12.41[b]**	**< .001[a]**	**-1.99**	**-.03**	**-.05**	**.16***
Content as a group	**38**	**1.67**	**.62**	**2**	**41**	**51**	**8**	**0**	**-6.74**	**< .001[a]**	**-0.51**	**-.21****	**-.17***	**-.08**
H vs. (H)	53	1.23	.47	1	79	20	1	1	-11.54[b]	< .001[a]	-1.69	-.12	-.07	.10
Whole vs. Detail: (H & A) vs. (Hd & Ad)	52	1.27	.53	1	77	20	3	1	-10.87[b]	< .001[a]	-1.48	-.15*	-.10	.03
Animal vs. Human: (A & Ad) vs. (H & Hd)	54	1.21	.43	1	80	19	1	0	-22.65	< .001[a]	-1.80	-.16*	-.14	.03
Ay vs. Art and other scores	26	1.88	.76	2	33	48	17	2	-0.74	.460	-0.06	-.07	-.02	.05
Cl vs. Na and other scores	44	1.57	.70	1	55	33	12	0	-8.34	< .001[a]	-0.66	-.13	-.07	.01
Isolation Contents: Ls vs. Bt vs. Na	31	1.84	.75	2	36	47	16	2	-1.82	.070	-0.15	-.15*	-.10	.02
Xy vs. An	48	1.44	.57	1	60	37	4	0	-14.63	< .001[a]	-1.05	-.08	-.04	.01
An vs. Animal/Human Detail	37	1.68	.69	2	44	46	9	1	-5.72	< .001[a]	-0.47	-.06	.02	.06
Fd vs. Bt/A/Ad	51	1.29	.51	1	74	24	3	0	-20.00	< .001[a]	-1.47	-.02	-.00	.04
Hh	47	1.44	.63	1	62	33	4	1	-13.27	< .001[a]	-0.98	-.08	-.06	.00
Hx vs. No Hx	33	1.83	.75	2	36	47	15	2	-1.91	.058	-0.17	.00	.02	.04
Sc	46	1.51	.66	1	58	34	9	0	-10.96	< .001[a]	-0.81	-.19**	-.11	.00
Sx vs. Hd/Ad vs. not	39	1.66	.69	2	46	42	11	1	-6.29	< .001[a]	-0.49	-.12	-.07	.02
Id vs. Not	20	2.03	.85	2	28	47	19	6	2.11	.036	0.18	-.22**	-.14	-.08
Popular	**56**	**1.18**	**.46**	**1**	**85**	**13**	**2**	**1**	**-11.54[b]**	**< .001[a]**	**-1.82**	**-.11**	**-.11**	**-.06**
Cognitive Special Scores as a group	**1**	**3.02**	**.73**	**3**	**2**	**19**	**54**	**25**	**23.74**	**< .001[a]**	**2.05**	**-.10**	**-.00**	**-.08**
Present vs. Absent	8	2.52	.85	3	12	37	40	12	11.14	< .001[a]	1.01	-.15*	-.09	-.10
Deciding what Cognitive Special Scores applies: DV, INCOM, DR, FABCOM, ALOG, or CONTAM	2	2.93	.78	3	3	25	49	24	20.27	< .001[a]	1.79	.01	.08	-.08

					Ratings (%)							rho		
	Rank	*M*	*SD*	*Mdn*	1	2	3	4	*t*(190)	*p*	Cohen's *d*	# *scored prot*	# *admin. prot*	*Completed semester*
Level 1 vs. Level 2 Distinction	4	2.66	.88	3	7	39	34	20	13.09	< .001[a]	1.21	.07	.14	.02
Other Special Scores as a group	**16**	**2.32**	**.81**	**2**	**13**	**50**	**29**	**8**	**7.77**	**< .001[a]**	**0.68**	**-.10**	**.03**	**-.09**
AB	17	2.15	.85	2	23	45	25	6	4.34	< .001[a]	0.38	.04	.09	.03
AG	41	1.63	.69	2	48	42	9	1	-7.64	< .001[a]	-0.55	-.14	-.10	-.02
COP	43	1.59	.72	1	53	38	7	2	-8.36	< .001[a]	-0.63	-.09	-.03	-.06
GHR vs. PHR	22	1.98	.91	2	35	40	18	7	1.17	.245	0.10	-.03	.02	-.04
MOR	40	1.65	.71	2	47	42	10	1	-6.23	< .001[a]	-0.50	-.09	-.03	.01
PER	24	1.97	.82	2	32	44	21	4	1.05	.296	0.09	-.14	-.03	-.11
PSV	25	1.93	.79	2	32	44	21	2	0.27	.784	0.02	-.10	.02	-.06
Z-Scores	**29**	**1.84**	**.84**	**2**	**40**	**39**	**17**	**4**	**-1.30**	**.196**	**-0.12**	**-.17***	**-.13**	**-.16***

Note. * $p \leq .05$; ** $p \leq .01$, [a] $p < .0009$; [b] sign test. Bold was used for coding categories.
The rating categories are as follows: 1 = Simple, straightforward, and very easy to score. Takes very little time to score; 2 = Understandable and relatively easy to score but does take some work at times. Takes a little time and effort to score; 3 = Sometimes a challenge and sometimes confusing. Often consumes a significant amount of scoring time; 4 = Complicated, confusing, and difficult to score. Consumes a great deal of scoring time.