Training and Education in Professional Psychology
2007, Vol. 1, No. 2, 125–134

1931-3918/07/$12.00 DOI: 10.1037/1931-3918.1.2.125

Criterion-Based Training for Rorschach Scoring

Mark J. Hilsenroth, Jocelyn W. Charnas, Jennifer Zodan
Derner Institute of Advanced Psychological Studies, Adelphi University

David L. Streiner
University of Toronto

This study addresses the effects of structured training on the development of Rorschach coding skills of graduate trainees and broadens the empirical base regarding student acquisition of these coding skills. A course outline for criterion based training in Rorschach scoring is reviewed. A training approach will be described that emphasizes a progressive "vertical" or "response segment" sequence to scoring training. The effects of this structured training protocol for graduate students Rorschach coding of Exner Comprehensive System criterion-scored protocols resulted in good to excellent levels of interrater reliability. The implications of these findings for training in Rorschach coding skills are discussed.

Keywords: Rorschach, training, graduate students, Comprehensive System

Supplemental Materials: http://dx.doi.org/10.1037/1931–3918.1.2.125.supp

The study of structured approaches in the training of clinical skills has become an area of increased interest over the last decade. Central to this discussion has been the desire for an examination of the direct relationship between specific training and the subsequent acquisition of specific clinical skills related to this training (Beutler & Kendall, 1995; Holloway & Neufeldt, 1995; Stein & Lambert, 1995). Unfortunately, graduate trainees are an understudied group regarding the effects of structured training despite the exceedingly important nature of such work to the eventual success of graduate education. In a recent summary of related clinical issues, Miller and Binder (2002) concluded that empirical evaluations of clinical training are sparse and little is known about how graduate training actually occurs or its effectiveness. For those of us involved in graduate education, these issues are of vital importance. Continued research is needed on how structured training impacts a variety of clinical skills, such as those necessary for psychological assessment, and even more specifically, Rorschach coding accuracy.

The practical implications of successful graduate training in Comprehensive System (CS) Rorschach coding are obvious. The two most recent and comprehensive surveys of predoctoral internships (Clemence & Handler, 2001; Stedman, Hatch, & Schoenfeld, 2000), each including over 300 Association of Psychology Postdoctoral and Internship Centers (APPIC) sites (most of which were programs accredited by the American Psychological Association [APA]), revealed that internship training directors greatly value the Rorschach, as well as integrated test batteries. Again, training directors reported a desire for incoming interns to have had courses on or at least a good working knowledge of the Rorschach. Finally, repeated surveys of psychological test use over the past 40 years have shown a substantial, consistent, and sustained use of the Rorschach in research and clinical settings (Butcher & Rouse, 1996; Camara, Nathan, & Puente, 2000). In fact, 90% of clinical practitioners working in the field express a belief that clinical students should be competent in Rorschach assessment (Watkins, Campbell, Nieberding, & Hallmark, 1995).

Mark J. Hilsenroth, Jocelyn W. Charnas, and Jennifer Zodan, Derner Institute of Advanced Psychological Studies, Adelphi University, Garden City, New York; David L. Streiner, University of Toronto, Ontario, Canada.

Correspondence concerning this article should be addressed to Mark J. Hilsenroth, 220 Weinberg Building, 158 Cambridge Avenue, Derner Institute of Advanced Psychological Studies, Adelphi University, Garden City, NY 11530. E-mail: hilsenro@adelphi.edu

Several studies in the peer-reviewed literature have demonstrated that the great majority (95%) of individual Rorschach variables are coded with good or excellent interrater reliability by appropriately trained raters (Meyer, 1997a, 1997b; Meyer et al., 2002; Viglione & Hilsenroth, 2001; Viglione & Taylor, 2003). This research is best appreciated in context where points of comparison are made with other instruments in psychological and cognitive assessment. (Hilsenroth & Stricker, 2004; Meyer & Archer, 2001). In addition, several authors have discussed qualitative aspects of teaching Rorschach coding to graduate trainees (Brabender, 1998; Handler, Fowler, & Hilsenroth, 1998; Hilsenroth, 1998; Weiner, 1998, 2004). However, there has been very limited empirical research that integrates these two areas by examining the effects of structured clinical training on the acquisition of Rorschach coding skills among graduate trainees.

To date, only one study has examined Rorschach training with graduate clinicians and its impact on scoring accuracy. In this work, Guarnaccia, Dill, Sabatino, and Southwick (2001) were the first to study Rorschach scoring accuracy with graduate trainees. Their sample consisted of 21 second-year graduate students and 12 PhD clinicians. The graduate students in this sample underwent 25 hours of scoring instruction, as well as practiced on 50 Rorschach responses. Both the graduate students and the clinicians then scored a total of 20 Rorschach CS (Exner, 1986) criterion-scored responses (10 nonclinical responses and 10 clinical) obtained from two CS training texts (Exner, 1986; Weiner, 1998). Results demonstrated that both the students and the clinicians achieved 77% agreement on the 10 nonclinical responses and 65% and 66% agreement on the 10 clinical responses, respectively. Overall findings demonstrated few significant differences between graduate students and PhD clinicians in scoring accuracy. Based on the 25 hours of training provided to the graduate students in this study, the authors concluded that graduate students need more instruction and scoring practice than can be achieved in the usual time allotted in doctoral training programs (based on Durand, Blanchard, & Mindell, 1988; M = 22.3 hours; SD = 20.3 hours). Furthermore, they state that graduate students who receive 25 to 30 hours of instruction or practice are probably not proficient enough to use the CS in clinical practice without receiving additional training (Guarnaccia et al., 2001).

It is important, however, to note several limitations of the Guarnaccia et al., 2001 study that may have had a significant impact on their findings. One issue is that of low base rates. The reliability results reported by Guarnaccia and colleagues were based upon participant's scores from only 10 Rorschach responses for both nonclinical and clinical samples. As Meyer and colleagues (2002) have empirically demonstrated, low base rates are generally problematic and may substantially decrease interrater reliability. Second, there was no estimate of the difficulty or variations in complexity of the responses that the participants were scored. Third, the training experiences of the PhD clinicians or even whether the PhD clinicians were from the same program, and thus had received the same training was unknown. In the absence of explicit information on this issue, this set of circumstances seems plausible because it would have been easier to recruit previous doctoral-level students from the same program, who had received the same training, and thus would make the lack of differences observed between graduate students and clinicians much less surprising. Fourth, in terms of reliability, only limited data were presented in standard format. Instead, a "point assignment" approach was utilized, meaning that one point was given for each accurately scored variable, and then categorical group contrasts (analyses of variance) were performed on these total "points correct" between the two groups. That is, one point was given for each accurately scored variable. This point-assignment approach to examining reliability is psychometrically problematic, difficult to interpret, and impossible to compare with other literature on Rorschach reliability. Fifth, regarding this comparison with preexisting Rorschach reliability research, the study reported only a few percent agreement statistics even though constituent scores were available in the article to calculate percent agreement for all response segments. Related to this issue is a complete absence of Kappa (κ) or intraclass correlation coefficients (ICC) interrater reliability values that correct for chance agreement. Lastly, specifics of course content, in-class instruction, procedures, and practice were never described.

The current study is distinctive in that it attempts to broaden the empirical knowledge base regarding graduate student acquisition of Rorschach coding skills and aims to fill some of the gaps in the existing literature. The purpose of the present study was to extend the examination of scoring accuracy in graduate trainees. In addition, specific information regarding a course outline for criterion-based training in Rorschach scoring will be reviewed. A training approach will be described that emphasizes a progressive "vertical" or "response segment" sequence to scoring training. The limitations of the Guarnaccia and colleagues (2001) will be addressed, and the effects of structured training for graduate students will be examined with respect to Rorschach CS criterion-scored protocols.

Method

Participants

Participants included 29 graduate students enrolled in an APA-approved clinical PhD program. Students were 20 women and 9 men and, as in Guarnaccia et al., 2001, in the second year of training. Students completed foundation courses in psychological assessment, personality theory, and psychotherapy before beginning this training. The course instructor was male, with a Clinical PhD as well as internship training from APA-approved programs and an early career assistant professor, albeit with extensive clinical and research experience using the Rorschach.

Procedure

Course materials included two Rorschach CS texts (Exner, 1993, 1995). Scoring examples of positive (i.e., accurate), ambiguous (i.e., requiring additional inquiry to clarify), and negative (i.e., inaccurate) instances of a variety of CS scores were also provided to students for in class review. Additionally, three CS criterion-scored *Rorschach Workshops* protocols were provided to students to ensure scoring accuracy for practice scoring homework. Finally, two CS criterion-scored protocols were provided to students as the basis for "midterm" and "final" evaluation for the course (these five protocols are available in the supplemental materials for this article).

At the start of the first class meeting, students reviewed the objectives of the course, and central among these was attaining a scoring proficiency of at least 80% total agreement across CS variables on each of the two CS criterion-scored protocols (midterm and final) to pass the course. There is an applied basis for this criterion as the Weiner (1991) editorial in *Journal of Personality Assessment* stated the requirement of a minimum 80% interrater agreement for publication. A student's failure to meet the criterion of 80% agreement required further scoring of Rorschach protocols (provided by the instructor) until this criterion was met. Should the 80% agreement criterion not be reached on these two protocols during the course of the semester, the student would receive an incomplete until this criterion was met. Additional course objectives were to provide students with instruction to insure the competent and appropriate use of assessment methods, familiarity with test construction issues for assessment methods such as standardization, reliability, validity, diagnostic efficiency, interpretation, bias, special populations, and recommendations for use commensurate with APA ethical standards (Section 2: Evaluation, Assessment, or Intervention). Finally, students reviewed selected readings on assessment methods which cover seminal theoretical, research, and clinical contributions.

The total course time spent on CS scoring training was 27 hours, two-1.5 hour classes each week (i.e., 3 hours a week) for 9 weeks. This is comparable to the time spent by Guarnaccia and colleagues (2001) who utilized 25 hours of scoring training. In-class activities included lectures on scoring principles, review of scoring examples (positive, ambiguous, and negative) from CS criterion sources, group discussion, and active participation stimulated by specific questions by the instructor regarding CS variable scoring criteria and definitions. The in class-activities and discussion of assigned readings were designed to amplify critical concepts in the lecture and scoring material. In addition, each week students completed practice scoring of homework assignments for in-class review and discussion that were based on three *Rorschach Workshops* CS criterion-scored clinical protocols (56 responses total, comparable in number to the 50 practice responses used by Guarnaccia et al., 2001).

The practice scoring of these three CS clinical protocols was completed progressively in "vertical/response segment" sequence from left to right as found on the Rorschach sequence of scores sheet. That is, students first scored location (Loc&S) and developmental quality (DvQ) for each of the three practice protocols for one class meeting. That scoring was then reviewed in the next class. Then, for the subsequent class meeting, students scored determinants (Det; movement, color and shading given specific focus across three individual classes) for each of the three practice protocols, to be reviewed in the next class. Then, form quality (FQ), pairs (2) and reflections (included with Det agreement), contents (Con), populars (P), *z* scores (*z*), content—special scores (spec. score) and, finally, thought disorder (SUM6)—spec. score for subsequent classes and review sessions. This progressive scoring of the same three CS practice protocols (56 total responses) provided more focused attention, repetition and repeated review designed to facilitate increased familiarity with the narratives of a few protocols, for both scoring and interpretation, as opposed to scoring greater number of protocols.

In addition to this practice scoring, students also viewed a videotaped administration of the Rorschach by the course instructor. Students were also required to videotape the administration of two Rorschach protocols with either a nonclinical volunteer(s) or a clinical patient(s). Because all psychological assessment and psychotherapy conducted at the outpatient clinic connected to this APA-approved clinical PhD program was videotaped, this did not represent a departure from standard clinic operating procedures. These videotaped Rorschach administrations were reviewed by the course instructor who provided feedback to the students. Although feedback on Rorschach administration skills was provided to the students immediately during the semester, these videotaped protocols were not reviewed by the course instructor for scoring accuracy nor was feedback regarding student coding provided until after completion of the course.

Students were evaluated based on scoring of two CS criterion-scored protocols (midterm and final) and were expected to have at least 80% total interrater reliability with the criterion CS scores on both of the protocols. To generalize to real world practice, students were allowed to consult CS workbook and training texts in scoring these two protocols. The first "protocol" or "midterm" was actually comprised of 19 responses selected from a large sample of nonclinical protocols designed to represent a wide variation of CS scores. However, because of the timing of the midterm exam, the class had yet to cover Thought Disorder Special Scores (i.e., SUM6). Therefore, no SUM6 scores were on this protocol, only content Special Scores (i.e., AG, COP, MOR, PER, etc.). The second protocol or "final" was a *Rorschach Workshops* inpatient protocol with 20 responses. Again, this protocol was selected because it included a wide variation of CS scores, including SUM6 scores.

To provide some external evaluation of the nature of these two protocols, 20 Rorschach experts were then contacted, all of whom had conducted graduate or postgraduate training involving the Rorschach, each has several Rorschach publications (peer-reviewed journal articles, chapter, books), and almost all are past or current consulting editors for the *Journal of Personality Assessment*. These experts were asked to review the midterm and final protocols (protocol narratives with free association and inquiry, the sequence of scores, and the structural summary) and to rate the scoring difficulty of the two protocols based on their experience. These experts were asked to place an "X" along a dotted line that had the descriptor "Not at all Difficult" at one end of this dotted line and "Extremely Difficult" at the other. A third descriptor of "Average" was placed at the center point (50th percentile) of this dotted line. Each expert's ranking of scoring difficulty was determined by converting where on the dotted line they had placed the "X" with a corresponding percentile rank (i.e., 0–100). All 20 (100%) of the Rorschach experts contacted returned their difficulty ratings with the midterm protocol rated at a mean scoring difficulty in the 32nd percentile and the final protocol rated at a mean scoring difficulty in the 72nd percentile. Therefore, the midterm protocol was considered to represent a fair to moderate level of scoring difficulty, whereas the final protocol was rated as a very difficult to score by these experts.

Results

Percent agreement was calculated in relation to CS criterion scores at the individual response level (exact agreement) and is reported for each response segment and total agreement across all scores on the protocol. Following procedures detailed in Meyer, 1999, (see also Meyer, 1997a, b; Meyer et al., 2002) chance agreement rates were generated based upon the actual response segment scores from the midterm and final protocols. estimated κ was then calculated from the observed and chance agreement rates for each of the response segment scores. Fleiss and colleagues (Fleiss, 1981; Fleiss & Cohen, 1973; Shrout & Fleiss, 1979) provide referents to the magnitude of standard estimates of reliability, κ or ICC, in the following ranges: <0.40 = *poor*, 0.40–0.59 = *fair*, 0.60–0.74 = *good*, >0.74 = *excellent*. Further recommendations for interpreting κ and ICC (Cicchetti, 1981, 1994) are as follows: < 0.40 = *poor*, 0.40–0.59 = *fair*, 0.60 to 0.74/0.79 = *good*, $>0.75/0.80$ = *excellent*, and >0.80 = *nearly prefect*.

Table 1 presents percent agreement and estimated κ for the midterm protocol (19 responses and scoring difficulty of 32%) for each CS response segment. All 29 graduate students obtained 80% or greater for total or overall percent agreement on this protocol, as well as greater than 80% agreement for all response segments. Estimated κ was in the excellent range (>0.74) for all response segments, except *z* and spec. scores that were both in the good to excellent range. Also presented in Table 1, as a point of comparison, is the percent agreement for the 21 graduate students on 10 nonclinical responses reported by Guarnaccia and colleagues (2001). In every case, except pairs, percent agreement for response segments was lower, in some cases substantially lower, than findings from the current study. It is important to note that the ratings of Guarnaccia and colleagues (2001) were across only 10 nonclinical responses, whereas the current study used almost twice as many responses to calculate reliability.

Table 2 presents percent agreement for the final protocol (20 responses and scoring difficulty of 70%) across each CS variable. Twenty-three of 29 graduate students (79%) obtained 80% or greater total or overall agreement on the final protocol. Of the six graduate students obtaining less than 80% agreement, four obtained 80% agreement after one additional protocol, and two obtained 80% agreement after two additional protocols. In addition, interrater reliability for each response segment demonstrated 80% agreement or more with the exception of Det at 78% and spec. scores at 65%. Likewise, estimated κ was in the excellent range (>0.74) for five of the nine response segments (Loc&S, DvQ, 2, Con, P), Det, FQ, and *z* were in the good to excellent range and spec. score were in the fair to good range of reliability. Also presented in Table 2 is the percent agreement for the 21 graduate students on 10 clinical responses reported by Guarnaccia and colleagues (2001). In every case, except pairs, percent agreement for response segments was substantially lower than findings from the current study. Again, it is important to note that the ratings of Guarnaccia and colleagues (2001) were calculated across 10 clinical responses, whereas the current study utilized twice as many responses to calculate reliability.

Given the discrepancies between our work and that of Guarnaccia and colleagues (2001),

Table 1
Interrater Reliability of Rorschach Response Segments for Non-Clinical Responses

Sample	Loc&S	DvQ	Det	FQ	2	Con	P	Z	Spec.Score	Total
Midterm protocol (N = 29)[a]										
Agreement (%)	96%	96%	85%	93%	91%	95%	92%	86%	89%[b]	91%
Estimated Kappa	.93	.93	.82	.81	.81	.94	.82	.71	.73[b]	
Guarnaccia et al., 2001 (N = 21)[c]										
Agreement (%)	82%	77%	75%	61%	93%	90%	87%	65%	56%	77%

Note. Spec.Score = Special Scores; Loc&S = Location and Space; DVQ = Developmental Quality; Det = Determinants; FQ = Form Quality; Con = Contents; 2 = Pairs; P = Populars; Z = Z-score.
[a] 19 non-clinical responses, expert rated scoring difficulty as 32nd percentile. [b] No thought disorder special scores (i.e., SUM6), only content special scores. [c] Ten non-clinical responses scored by 21 graduate students.

Table 2
Interrater Reliability of Rorschach Response Segments for Clinical Responses

Sample	Loc&S	DvQ	Det	FQ	2	Con	P	Z	Spec.Score	Total
Final Protocol (N = 29)[a]										
% Agreement	99%	91%	78%	80%	92%	90%	97%	83%	65%	83%
Estimated Kappa	.98	.86	.73	.71	.82	.89	.88	.65	.56	
Guarnaccia et al., 2001 (N = 21)[b]										
Agreement (%)	82%	76%	51%	61%	93%	67%	93%	72%	34%	65%
Meyer et al., 2002 (N = 4)[c]										
Agreement (%)	90%	92%	73%	82%	93%	76%	95%	[d]	82%	86%
Estimated Kappa	.86	.86	.65	.70	.82	.70	.87	[d]	.68	

Note. Loc&S = Location and Space; DvQ = Developmental Quality; Det = Determinants; FQ = Form Quality; Con = Contents; Spec.Score = Special Scores; 2 = Pairs; P = Populars; Z = Z-score.
[a] Twenty clinical responses, expert rated scoring difficulty as 72nd percentile. [b] Ten clinical responses scores by 21 graduate students. [c] 1,407 clinical responses from 66 protocols. [d] Meyer et al., 2002, do not report reliability data for Z Scores.

we examined the literature for other relevant data on student Rorschach coding reliability. We were aware of the student group in the Meyer and colleagues (2002) study examining the interrater reliability of various sets of coding groups. Although not a training study per se, Meyer and colleagues (2002) report on the scoring reliability data of four students (three graduate students and one bachelor level student with Rorschach training). A comparison of interrater reliability for the students on clinical responses from Meyer and colleagues (2002) points to a substantially different picture. At the bottom of Table 2 are the percent agreement and estimated κ for student-scored response segments from Meyer et al., 2002. It is important to note that none of these students were part of the current study or underwent the training course currently being described. Percent agreement and estimated κ reported for response segments in Meyer and colleagues (2002) are highly similar to the findings reported in the current study. It is also important to note that the student ratings reported in this study (Meyer et al., 2002) were across 1,407 responses from 66 protocols. Therefore, they have a more than adequate base rate with which to assess reliability.

In addition to the classroom setting described above, we have also developed a workshop training manual based on these same structured training initiatives for use across a variety of settings (Hilsenroth & Charnas, 2007).[1] Included with the workshop outline and training initiatives are 30 Rorschachs that include both clinical and nonclinical protocols (15 clinical and 15 nonclinical) from a range of different patient diagnoses, individual sociodemographics, and clinical administrators. Each of these 30 protocols was scored by at least three independent raters (sometimes as many as four or five), including at least one Rorschach expert in the consensus tabulation of criterion scores. Additionally, as described earlier with the midterm and final protocols, Rorschach experts were asked to rate the scoring difficulty level of each protocol. Table 3 presents the results of an interrater reliability trial of two graduate student research assistants with the expert criterion scores of 20 Rorschach protocols on the central interpretive CS variables using the current training model in a workshop format. All variables, with the exception of two content variables (A:Ad & Fd), were found to be in the good (>0.59) or excellent (>0.74) level or interrater reliability, and 89% of all ICC calculations were found to be in the excellent range of reliability.

Discussion

Initial findings regarding the Rorschach scoring accuracy led Guarnaccia and colleagues (2001) to state that graduate students need more instruction and scoring practice than can be achieved in 25 to 30 hours of instruction or practice and are probably not proficient enough to use the CS in clinical practice without receiving additional training. However, the findings of

[1] This manual is available from the author and also at http://dx.doi.org/10.1037/1931-3918.1.2.125.supp.

Table 3

Interrater Reliability (ICC 1,1) for Two Graduate Student Raters With 20 Criterion Scored Rorschach Protocols on the Central Interpretive CS Variables

Ratios, percentages, and derivations				

R = .96	L = .99			COP = .82 AG = .90
				Food = .57

EB = .96:.94
eb = .88:.98
	EA = .97
	es = .94
	Adj es = .92
		D = .83
		AdjD = .77
			FC:CF + C = .81:.79
			Pure C = .83
			C′:WSumC = .74:.94
				Isolate/R = .95
				H:(H)Hd(Hd) = .97:.94
				(HHd):(AAd) = .91:.55

FM = .96
m = .76
	C′ = .74
	V = .87
		T = .88
		Y = .80
			S = .94
			Blends% = .93
				H + A:Hd + Ad = .80:.90
				GHR = .90
				PHR = .90

a:p = .91:.92
Ma:Mp = .93:.91
2AB + Art + Ay = .82
M− = .80
	Sum6 = .88
	WSum6 = .84
	P = .68
		XA% = .88
		WDA% = .85
		X+% = .87
		F+% = .97
		X−% = .86
		S−% = .84
		Xu% = .72
			Zf = .95
			Zd = .93
			W:D:Dd = .99:.91:.97
			W:M = .99:.96
			DQ+ = .86
			DQv = .60
				3r + (2)/R = .88
				Fr + rF = .79
				FD = .88
				An + Xy = .92
				MOR = .96

EII = .92	PTI = .65	DEPI = .84	CDI = .95	S-CON = .88	HVI = .91

Note. ICC(1,1) = One-Way Random Effects Model. Fleiss and colleagues (Fleiss 1981; Fleiss & Cohen, 1973; Shrout & Fleiss 1979) provide referents to the magnitude of standard estimates of reliability, Kappa or ICC, in the following ranges: < .40 = poor; .40 − .59 = fair; .60 − .74 = good; > .74 = excellent. Further recommendations for interpreting Kappa and ICC (Cicchetti, 1994; Cicchetti, 1981) are as follows: < .40 = poor, .40 to .59 = fair, .60 to .74/.79 = good, > .75/80 = excellent, and > .80 as nearly perfect. Definitions and calculations of all abbreviations are provided in Exner, 2003.

the current study stand in contrast to the conclusions offered by Guarnaccia and colleagues (2001). Specifically, the current findings suggest that significant ability in scoring accuracy is possible in the usual time allotted to students in graduate training programs (Durand et al., 1988). Furthermore, graduate students who receive 25 to 30 hours of instruction and practice may be more than capable to begin to use the CS in clinical practice. In light of these contradictory findings, it seems imperative that we understand what specific aspects of instruction can positively affect the relationship between training and skill acquisition.

Examining the impact of training on the subsequent acquisition of clinical skills, it is important to discuss which aspects of training were similar across the present investigation and the Guarnaccia and colleagues (2001) study. These similarities include the training texts that were used, as well as the time spent on CS scoring training (27 and 25 hours, respectively). Also, the student participants were in their second year of training, and these participants scored a comparable number of practice responses (56 and 50, respectively) as part of the training process. In addition, students in both studies were evaluated using both clinical and nonclinical responses.

However, given the differences in scoring accuracy observed between the two studies, we are invariably led to examine which structured training experiences might improve Rorschach scoring accuracy. The differences in training between the present investigation and Guarnaccia and colleagues (2001) include an explicit training criterion of 80% total or overall agreement that was identified at the beginning of the course as requirement for its' successful completion. This was not the case in the Guarnaccia and colleagues (2001) study. Additionally, training in the current study included a review of scoring examples (positive, ambiguous, and negative) from CS criterion sources and weekly reviews of practice scoring on three CS criterion-scored clinical protocols in a progressive vertical/response segment sequence, both of which were not included in the training provided by Guarnaccia and colleagues (2001). The present study also evaluated twice the number of CS criterion-scored, nonclinical and clinical, responses, and the scoring difficulty of these protocols was examined. Lastly, the present study included review of videotape administration with feedback by the instructor, which was not included in the Guarnaccia and colleagues (2001) study.

Implications

Based on the current study, there are several implications for Rorschach scoring training in programs where the Rorschach is part of the curriculum. We would encourage graduate training courses in Rorschach coding to include the use of an explicit training criterion (80% agreement) linked to successful completion of the course, the use of a variety of CS criterion scoring examples for in-class discussion, and weekly systematic practice as well as evaluation on as large a sample of CS criterion-scored responses as possible (40–50 responses) in order to provide sufficient base rates for an effective evaluation of scoring accuracy. Regarding the lack of sufficient sized samples on which to calculate interrater reliability, Meyer and colleagues (2002) present empirical data that very clearly demonstrates "Small samples often produce misleading reliability results. When this occurs, the small sample results underestimate the true reliability about five times more often than they overestimate it." (p. 252). As a practical point of comparison, a *Rorschach Workshops* beginning CS scoring tutorial is completed over 5 days with 35 total hours of training, and a 3-day advanced CS class meets for 21 total hours. It seems that even without the benefit of the current findings, these *Rorschach Workshops* training programs have been organized in a manner that is highly consistent with and capitalizes on the length and focus of training experiences described in this study. Finally, we would recommend that faculty use the information about these structured training initiatives to aid in more creatively designing training experiences to meet the individual goals of a given program. That is, these training initiatives may be implemented as part of a semester long course, or during an intersession, summer, or weekend workshop to improve scoring accuracy.

Although the findings of this study are quite positive regarding the CS scoring accuracy of trainees there are still areas for improvement in the training process. Even though the final inpatient CS protocol was rated by experts as being in the top third of Rorschach protocols for scoring difficulty based on their experience, our findings demonstrate that extra time and emphasis should be given during training to the scoring categories of Det, FQ, and *z*. Furthermore, it is probably essential that the coding of spec. scores be given at minimum twice the amount of allotted training time as any of the other scoring categories. For those involved in Rorschach training and research, these varying degrees of scoring difficulty have been recognized, and several teaching aids have been developed to address these issues. First, improvement in training for the coding of CS variables is the focus of a new supplemental reference book (Viglione, 2002) that provides several scoring examples of positive, ambiguous (requiring additional inquiry to clarify), and negative instances of a variety of CS scores in the identical fashion as was provided to students in this study for in-class review and discussion. Second, training tapes by experts provide a useful guide and template for students to observe and model behavior. As such, a new digital video is available that not only demonstrates CS Rorschach administration by Rorschach experts but also addresses several commonly asked questions among students regarding the Rorschach administration process (Sciara & Ritzler, 2006). These innovations capitalize on some of the key aspects in this study that may have led to increased scoring accuracy.

In conclusion, it is important to note that the findings of this study may be attributable to one, several, or some interaction of the different training experiences that were used. Regardless of whether one aspect of this training had a larger effect than another or some synergistic interaction occurred, the results of this study support the positive benefits these various structured clinical training initiatives have on CS Rorschach scoring accuracy. Nevertheless, future research will need to replicate and further examine these, as well as other, innovative training experiences to ascertain whether the same findings can be replicated across instructors and with varying levels of both protocol scoring difficulty (i.e., 20th, 40th, 60th, and 80th percentiles), as well as trainee clinical or Rorschach experience (i.e., predoctoral, internship, postdoctoral workshop, etc.).

References

Beutler, L., & Kendall, P. (1995). Introduction to the special section: The case for training in the provision of psychological therapy. *Journal of Consulting and Clinical Psychology, 63,* 179–181.

Brabender, V. (1998). Teaching that first Rorschach course. In L. Handler, & M. Hilsenroth (Eds.), *Teaching and learning personality assessment* (pp. 215–234). Hillsdale, NJ: Erlbaum.

Butcher, J., & Rouse, S. (1996). Personality: Individual differences and clinical assessment. *Annual Review of Psychology, 47,* 87–111.

Camara, W., Nathan, J., & Puente, A. (2000). Psychological test usage: Implications in professional use. *Professional Psychology: Research and Practice, 31,* 141–154.

Cicchetti, D. V. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency, 86,* 127–137.

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

Clemence, A., & Handler, L. (2001). Psychological assessment on internship: A survey of training directors and their expectations for students. *Journal of Personality Assessment, 76,* 18–47.

Durand, V., Blanchard, E., & Mindell, J. (1988). Training in projective testing: Survey of clinical training directors and internship directors. *Professional Psychology: Research and Practice, 19,* 236–238.

Exner, J. (1986). *The Rorschach: A comprehensive system, Vol. 1: Basic foundations* (2nd ed.). New York: Wiley.

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

Exner, J., (1995). *A Rorschach workbook for the comprehensive system* (4th ed.). Asheville, NC: Rorschach Workshops.

Exner, J. (2003). *The Rorschach: A comprehensive system, Vol. 1: Basic foundations* (4th ed.). New York: John Wiley & Sons, Inc.

Fleiss, J. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley.

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

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

Handler, L., Fowler, C., & Hilsenroth, M. (1998). The use of a group classroom experience to learn the process of personality assessment. In L. Handler, & M. Hilsenroth (Eds.), *Teaching and learning personality assessment* (pp. 431–452). Hillsdale, NJ: Erlbaum.

Hilsenroth, M., & Charnas, J. (2007). *Training manual for Rorschach Interrater Reliability* (2nd ed.). Unpublished manuscript, The Derner Institute of Advanced Psychological Studies, Adelphi University, Garden City, NY.

Hilsenroth, M., & Stricker, G. (2004). A consideration of challenges to psychological assessment instruments used in forensic settings: Rorschach as exemplar. *Journal of Personality Assessment, 83,* 141–152.

Hilsenroth, M. J. (1998). Using metaphor to understand projective test data: A training Heuristic. In L. Handler, & M. Hilsenroth (Eds.), *Teaching and learning personality assessment* (pp. 391–412). Hillsdale, NJ: Erlbaum.

Holloway, E., & Neufeldt, S. (1995). Supervision: Its contributions to treatment efficacy. *Journal of Consulting and Clinical Psychology, 63,* 203–213.

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

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

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

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

Meyer, G. J. (1999). Simple procedures to estimate chance agreement and kappa for the interrater reliability of response segments using the Rorschach Comprehensive System. *Journal of Personality Assessment, 72,* 230–255.

Miller, S., & Binder, J. (2002). The effects of manual-based training on treatment fidelity and outcome: A review of the literature on adult individual psychotherapy. *Psychotherapy, 39,* 184–198.

Sciara, A., & Ritzler, B. (2006). *The little book on administration for the Rorschach Comprehensive System.* Asheville, NC: Daniels Graphics.

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

Stedman, J., Hatch, J., & Schoenfeld, L. (2000). Preinternship preparation in psychological testing and psychotherapy: What internship directors say they expect. *Professional Psychology: Research and Practice, 31,* 321–326.

Stein, D., & Lambert, M. (1995). Graduate training in psychotherapy: Are therapy outcomes enhanced? *Journal of Consulting and Clinical Psychology, 63,* 182–196.

Viglione, D., & Hilsenroth, M. (2001). The Rorschach: Facts, fictions, and future. *Psychological Assessment, 13,* 452–471.

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

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

Watkins, C., Campbell, V., Nieberding, R., & Hallmark, R. (1995). Contemporary practice of psychological assessment by clinical psychologists. *Professional Psychology: Research and Practice, 26,* 54–60.

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

Weiner, I. (1998). Teaching the Rorschach Comprehensive System. In L. Handler, & M. Hilsenroth (Eds.), *Teaching and learning personality assessment.* (pp.-). Hillsdale, NJ: Erlbaum.

Weiner, I. (2004). Rorschach assessment: Current status. In M. Hilsenroth, & D. Segal (Eds.), *Comprehensive handbook of psychological assessment* (pp. 562–572). New York: Wiley.

Received April 20, 2006
Revision received September 26, 2006
Accepted November 6, 2006 ■