Journal of Personality Assessment, *89*(S1), S193–S200
Copyright © 2007, Lawrence Erlbaum Associates, Inc.
ISSN: 0022-3891 print / 1532-7752 online
DOI: 10.1080/00223890701583614

The Impact of Administration and Inquiry on Rorschach Comprehensive System Protocols in a National Reference Sample

ADRIANA LIS,[1] LAURA PAROLIN,[2] VINCENZO CALVO,[1] ALESSANDRO ZENNARO,[3] AND GREGORY MEYER[4]

[1]*Dipartimento di Psicologia dello Sviluppo e della Socializzazione, Università degli Studi, Padova, Italy*
[2]*Dipartimento di Psicologia, Università degli Studi, Milano Bicocca, Italy*
[3]*Università degli Studi di Aosta, Italy*
[4]*Department of Psychology, University of Toledo*

We investigated the impact of administration and inquiry skills on Rorschach Comprehensive System (CS; Exner, 1974, 1991, 1993) protocols collected for the Italian adult nonpatient reference sample. The same research team collected CS protocols on two occasions. The initial reference sample ($N = 212$; Lis, Rossi, & Priha, 1998) was collected under the supervision of experienced psychologists who carefully studied CS administration and scoring procedures (Exner, 1986, 1990, 1993). The second sample ($N = 101$; Lis, Zennaro, Calvo, & Salcuni, 2001) was collected after the team obtained additional and sustained CS training from Rorschach workshops certified instructors. Both samples were scored, reliably but they showed large differences on many codes, with protocols from the second sample being richer and more complex than the first. The results indicate that administration skills can have a dramatic impact on CS protocols and may contribute to variations in samples collected by different investigators. Training standards should be devised to insure uniform administration procedures are followed when collecting CS protocols.

The Comprehensive System (CS; Exner, 1974, 1991, 1993; Exner & Weiner, 1995), since its appearance in 1974, has become the most widely researched (Shontz & Green, 1992) and most commonly taught (Hilsenroth, Handler, Toman, & Padawer, 1995) system for administering, scoring, and interpreting Rorschach responses. Standardization of an instrument resides in its being composed of specific stimuli, methods of administration, instructions to respondents, and coding criteria to be used uniformly. With the introduction of the CS, Exner (1986, 1993) made a serious attempt to establish psychometric proprieties for the Rorschach system, a position achieved, among other things, by presenting standard administration and scoring procedures. Exner (1993) prescribed an invariant set of instructions to be used in all Rorschach administrations, strict adherence to which helps ensure that the Rorschach data are collected in the same manner from all respondents and can be cumulated for research purposes. Exner (1993) also formulated detailed guidelines for coding individual Rorschach responses and for combining these codes into various percentages, ratios, and indexes. Reports of Rorschach research must now contain detailed interscorer reliability information. For instance, the *Journal of Personality Assessment*, which often publishes Rorschach research, requires manuscripts to include specific details of how interscorer agreement was computed and has set a minimum standard of 80% interscorer agreement for publication (Weiner, 1991). Despite these standards, Hunsley and Bailey (1999) expressed concern that because reliable Rorschach coding depends on the coding skill of individual examiners, there remains "the significant question of how reliably the Rorschach is scored in routine clinical practice (i.e. field reliability)" (p. 268).

A related issue of considerable importance that has been all but ignored by researchers is the influence of administration procedures on the protocols that are obtained from participants. Administration procedures include a number of different variables, including seating arrangements, recording devices, the instructional set given to subjects, the working relationship between the subject and examiner, and the specific steps taken when the subject is providing responses (e.g., prompts) and when those responses are being inquired by the examiner.

A survey of the Rorschach literature (Hartmann, 2001) showed that the instructions used in the Rorschach can be categorized roughly as either the very short instruction, "What might this be?" that was originally developed by Rorschach (1942, p. 16) and recommended by Exner in the CS, or as different versions of a longer, more elaborated instruction, originally developed by Hertz (1936). According to Exner (1986, 1993), the short instruction was chosen as the standard for the CS because it produces the fewest responses and directs the individual to focus on problem solving.

Exner (1969, 1986, 1993) suggested that differences in Rorschach administration procedures, such as variations in seating arrangements or instructions given to the participant, could result in protocols that might be substantially different in several ways. For instance, Exner (1986, 1993) indirectly examined the impact of different instructions by comparing Rorschach protocols collected across five major Rorschach systems: Klopfer, Beck, Rapaport, Piotrowski, and Hertz. Although the survey was retrospective and without random assignment to the different administration procedures, Exner concluded that different administration procedures produced significant variation in the number of responses given to the test. Similarly, Goetchneus (as cited in Exner, 1986) compared the instructions of Klopfer and Beck and concluded that Klopfer's instruction produced protocols that on average were six responses shorter than those administered by Beck's instructions.

Received February 23, 2004; Revised October 26, 2006

Address correspondence to Adriana Lis, Dipartimento di Psicologia dello Sviluppo e della Socializzazione, Università degli Studi, Padova, via Belzoni 80, 35141 Padova, Italia; Email: adriana.lis@unipd.it

Hartmann (2001) compared the effects of short versus long introductory instruction in a nonpatient sample. The short instruction produced significantly more questions to the examiner about the test and more brief protocols than did the long one. No between-group differences, however, were observed in the scores obtained from both instructional sets.

Another important administration issue concerns the interpersonal climate between examiner and client. Clients who are mistrustful or uncooperative for other reasons are prone to give brief and simplistic protocols. Conversely, clients who feel comfortable and open are prone to give more lengthy and rich records. Relevant evidence in this regard can be found in Exner, Armbruster, and Mittman (1978), who examined protocols obtained from patients who were either tested by their own therapist in the midst of ongoing therapy or by a different therapist. The patients tested by their own therapists produced longer, more complex, and more revealing protocols than those tested by a different therapist.

The Inquiry is the phase of the test when old information is reviewed and clarified. The purpose of the Inquiry is for the examiner to see what the client saw at the time when the response was delivered, including where the percept was located in the blot and the blot features that contributed to the percept. It is a delicate phase of the test that, if misunderstood by the client or mishandled by the examiner, can lead to many problems in coding responses or interpreting the test data or both. A good inquiry is essential for accurate scoring and valid interpretation. The overall purpose of the Inquiry is to insure that the coding (scoring) of response is as accurate as possible.

Although the examiner's role in the inquiry phase of the Rorschach administration has relatively simple guidelines, the procedure is not easy to conduct adequately (Exner, 1986). According to Exner, the Inquiry is one of the most misunderstood and abused features of the Rorschach. When done correctly, it completes the richness of the test data. When done incorrectly, it often generates data that may be of clinical interest but that represent something other than CS data. Consequently, Exner referred to inquiry as "the soft underbelly of the test" (p. 75) and strongly implied that a poor inquiry nearly disqualifies the Rorschach as a useful technique for personality assessment.

Beck (1953) and Klopfer (1942) also gave specific emphasis to the critical and sensitive nature of the inquiry process. They both called for studies on the effect of inquiry, but few have taken heed.

In a well-controlled study, Blais et al. (1995) compared CS and Rapaport administration procedures. The findings showed that the CS administration, in which the Inquiry is conducted only after all responses have been given, produced significantly more color, shading, and blends than did the Rapaport administration, in which inquiry is conducted immediately after each response.

Most of the studies on Inquiry have altered some aspect of the administration procedure and observed its effect. In particular, examiners have asked directly about elements of the response (e.g., "Did the color make a difference?") rather than confining themselves to the more indirect, nonleading questions recommended in the standard inquiry (e.g., "What about the inkblot make it look that way?"). Results show that direct questions increase color, movement, and shading scores in patterns that would greatly change interpretations made from the scoring summaries (Baughaman, 1958; Kligensmith, 1956; Zax & Stricker, 1960).

One early study took a straightforward approach to the impact of the Inquiry by scoring protocols with and without the Inquiry (Gibby & Stotsky, 1953). The protocols of 240 Veterans Administration hospital psychiatric patients were scored according to the Beck (1944) system. Results showed significant increases in diffuse shading, depth from shading, color, and human movement, and a decrease in pure form when protocols included an inquiry.

Ritzler and Nalesnik (1990) extended the Gibby and Stosky procedure to the CS. The effect of the Inquiry on the CS was tested by scoring 130 protocols of patients and nonpatients with and without the Inquiry portion of the response. The presence of the Inquiry derived information significantly reduced the means for Developmental Quality–vague (DQv), Form Quality–unusual (FQu), and pure form (F) and it significantly increased the sum of shading responses (Sum Shading), the sum of color responses (SumC), diffuse shading (FY and YF), texture (FT and TF), vista (FV and VF), achromatic color (FC' and C'F), and a weighted thought disorder measure (WSum6), along with four of the six individual variables that constitute it (DV, DR, ALOG, and CONTAM). The effect on WSum6 was strongest for schizophrenic patients.

The Rorschach inkblot method is an internationally used instrument, and variations still exist in the application of this assessment method (Aronow, Reznikoff, & Moreland, 1995; Blais, Norman, Quintar, & Herzog, 1995; Hartman, Davidsen, & Molin, 1999; Killingmo, 1980; Lerner, 1998). Since the Rorschach was introduced in Italy, it has become one of the most frequently used personality assessment methods among clinical psychologists. Although many Italian clinicians and researchers acknowledge the psychometric qualities of the CS, initially many of them resisted adopting it, as they feared the CS would not allow a clinical–relational approach to the patient.

The aim of this article was to investigate the effect of administering and scoring the Rorschach test following different types of CS training for a team of psychologists. We highlight the important role that examiner training can have on the type of protocols that are collected by examining two Italian adult nonpatient reference samples obtained successively by the same team of psychologists. To illustrate this, we (1) describe the initial data collection procedures, including the examiners that were used and how they were trained and supervised; (2) describe how we came to realize that the initial data collection procedures were problematic; (3) describe the second set of data collection procedures, including the examiners who were used and how they were trained and supervised; and (4) compare both data sets across a wide range of CS variables.

Based on the existing literature, to the extent that training for the second data set improved CS administration and inquiry skills, cross-sample differences consistent with Ritzler and Nalesnik's (1990) findings could be expected for developmental quality, form quality, color, shading, pure form (F), and Special Scores. In other words, the second data set should show a decrease in DQv, FQu, and pure F, but an increase in Sum Shading, SumC, SumY, SumT, SumV, SumC', and WSum6.

METHOD

The Team

Initial training for the 1998 data set. Twelve second-level graduate students enrolled in the clinical training program of the Scuola di Specializzazione del Ciclo di Vita and of the Scuola di Specializzazione in Psicologia Clinica at the University of Padova, Italy, participated in the 1998 data collection. During their second-level graduate course at the Psychology Faculty at Padova University the students had completed a one-semester course in personality evaluation. In this course the students were introduced to the Rorschach CS. Administration was practiced in the classroom, and students were required to score approximately 100 responses from various protocols provided by the instructors. With in-class practice and homework assignment, students had more than 20 hours of scoring instruction and practice. After graduation, all the students enrolled in an additional training program that provided specific intensive training in CS scoring and administration. During this course the students practiced administration under supervision with 20 volunteers and received feedback on the adequacy of their administration. When combined with additional homework assignments, students had about 30 hours of administration and of scoring instruction and practice. The number of hours of training was above the average that has been reported for other programs (Durand, Blanchard, & Mindell, 1988).

The instructors were psychologists with advanced training in the Rorschach according to the Swiss School, and had taught the Rorschach for more than 15 years with this method. Their CS training was carried out through a careful reading of Exner's books (Exner, 1986, 1990, 1993). They had carried out CS administration and scoring based on this learning for about 5 years.

Subsequent training for the 2001 data set. In 1998 the two instructors were accepted as associated members at the European Rorschach Association (ERA). In that context they interacted with Anne Andronikof Sanglade, who kindly agreed to give the 1998 data collection team (students and instructors) regular CS training. Prof. Andronikof was the founder of the ERA, held an advanced training certificate from the Rorschach workshops, and had taught the Rorschach for more than 20 years.

The same team had three second-level courses with Prof. Andronikof, starting in September 1998. With in-class practice and homework assignments, the team had more than 90 hours of instruction in the general rationale of CS administration, scoring, and interpretation. Subsequently, the team received from Prof. Andronikof advanced CS courses covering childhood, forensics, psychotherapy research, and other topics. In total, during these courses, besides learning Rorschach interpretation, the participants had in-class experience in Rorschach administration, and were required to score 300 responses, for a total of approximately 20 protocols provided and supervised by Prof. Andronikof.

In addition, by 2001 the two initial instructors had experience teaching the CS for about 3 years and had obtained advanced training certificates from Rorschach workshops. After the courses, the team continued training in CS administration and scoring by meeting for about 5 hours a week for about one year. The team had about 30 hours of experience in Rorschach administration outside of the classroom. Prof. Andronikof supervised these administrations and provided team members with ongoing feedback about administration adequacy. There was also regularly scheduled contact by e-mail to discuss doubts and difficulties. As with the initial 1998 training, the number of hours of training for the second round of data collection was above the average that has been reported for other training programs (Durand et al., 1988).

Participants. In the 1998 study, 212 nonclinical adults (104 males and 108 females) aged 25–50 volunteered to participate. In the 2001 study, 101 non–clinical adults (36 males and 65 females) volunteered to participate. Although Lis, Parolin, Salcuni, and Zennaro (2007, this issue) present data for 249 adults, our research for this article was completed in 2003 before the full sample had been assembled. As such, we focus on just the first 101 adults included in the new sample of 249.

All protocols were collected and scored by the same team of examiners. For both the 1998 and 2001 studies, the experimenters recruited subjects personally in public places (libraries, university, etc.) over a period of one year. The research project was explained to subjects, and they were asked to participate without receiving compensation. The participants were treated in accordance with the Ethical Principles of Psychologists and Code of Conduct (American Psychological Association, 1992).

Subjects were excluded if they met either of the following criteria: (a) a lifetime history of psychotherapy and psychiatric treatment (including prior hospitalizations and medications), psychological disorders, criminal arrests, and substance abuse; or (b) current psychotherapy and psychiatric treatment (including medication use), psychological disorders, and substance use. Subjects were included if they were Italian adults, aged 25–65 years, and spoke Italian as their primary language. All participants lived in Northern Italy, and their ethnic background was Caucasian. Socioeconomic status was medium and similar for the 1998 and 2001 participants. None of the participants had completed the Rorschach before.

Scoring reliability. To obtain interrater reliability data for CS scoring, 20 protocols from the 1998 data set and 30 protocols from the 2001 data set were selected at random and coscored by two judges of the team. Interscorer reliability was calculated by percentage of agreement, and all codes and variables were higher than 80% (Weiner, 1991). Because percentage of agreement does not always reflect the variance in agreements and disagreements (Acklin, MacDowell, Verschell, & Chan, 2000; Wood, Nezworski, & Stejskal, 1996), and even 80% agreement can obscure major disagreements (Grønnerød, 1999), we computed iota values for response segments. Because our study focused on summary scores (e.g., total Dd scores across all protocols), it would have been optimal to compute intraclass correlations (ICCs) for the summary scores in both the 1998 and 2001 data sets. Because 20 or 30 cases constitute small reliability samples that produce unstable ICC results, however, we decided to report iota values for response segments (Janson & Olsson, 2001). Iota is an extension of Cohen's (1960) kappa that allows one to compute chance-corrected agreement for a multivariable test scored by two or more raters (Janson, 2003). The measure is applicable to nominal or interval variables, and it can be used to examine agreement on individual CS responses or CS summary scores. Response-level agreement is informative about the precision with which coders were able to apply the same categorical coding scheme to responses, and it is useful for monitoring training and practice (Janson, 2003). The *JPA* recommendations for publishing intercoder agreement (Weiner, 1991) recommended only response-level agreement for "response segments," and these are the most commonly reported measures (Meyer, 1999; Viglione

TABLE 1.—Rorschach interrater agreement on coding segments.

	1998 Data ($N = 20$)		2001 Data ($N = 30$)	
Coding Segment	% Agree	Iota	% Agree	Iota
Whole Responses (All scores in a response)	.996	.972	.975	.863
Location & Space (2 variables)	.995	.988	.963	.911
DQ (+,o,v/+,v)	.981	.939	.883	.760
Determinants (11 variables)	.993	.943	.965	.824
FQ (None,+,o,u,–)	.961	.942	.830	.732
Pairs	1.00	1.00	.965	.925
Contents (27 variables)	.999	.995	.988	.865
P	1.00	1.00	.946	.864
Z Score	.991	.986	.883	.849
CS Special Scores (14 variables)	.997	.892	.988	.728

TABLE 2.—Differences in reference data for Italian adults collected in 1998 and 2001.

	1998 Initial Sample		2001 Corrected Sample			
Score	*M*	*SD*	*M*	*SD*	Cohen's *d*	*d* Label
R	26.77	10.62	23.56	8.50	−.33	S
W	10.77	5.51	9.54	4.37	−.25	S
D	12.69	7.29	9.72	6.15	−.44	M
Dd	3.32	4.48	4.30	3.44	.25	S
S	2.08	2.12	3.67	2.59	.67	M
DQ+	3.61	3.07	6.01	3.89	.68	M
DQo	20.34	9.12	15.52	6.58	−.61	M
DQv	2.43	2.76	1.44	1.52	−.44	M
DQv/+	.29	.60	.59	.93	.38	S
FQ+	.04	.28	.12	.38	.24	S
FQo	13.61	4.90	10.39	3.52	−.75	L
FQu	6.18	4.22	7.37	5.68	.24	S
FQ–	6.66	4.72	5.32	3.13	−.33	S
FQnone	.28	.60	.37	.70	.14	S
MQ+	.01	.15	.00	.17	−.06	S
MQo	1.33	1.35	1.77	1.40	.32	S
MQu	.41	.83	1.16	1.45	.63	M
M–	.26	.60	.96	1.31	.69	M
Mnone	.01	.10	.00	.01	−.14	S
S–	.79	1.10	1.57	1.45	.61	M
M	1.59	1.58	3.93	2.68	1.06	V L
FM	1.37	1.57	2.97	2.30	.81	L
m	.29	.64	1.62	1.71	1.03	V L
FM+m	2.47	2.33	4.59	3.28	.75	L
FC	1.48	1.56	2.61	1.75	.68	M
CF	1.25	1.37	1.50	1.25	.19	S
PureC	.17	.48	.44	.77	.42	M
Cn	.00	.00	.00	.01	.00	S
SumC	2.90	2.17	4.55	2.29	.74	L
WSumC	2.99	2.00	3.46	1.85	.24	S
SumC′	.50	.81	2.37	1.75	1.37	V L
SumT	.24	.55	1.11	1.07	1.02	V L
SumV	.52	.93	.81	1.32	.25	S
SumY	1.52	1.53	1.90	1.81	.23	S
Sum Shading	2.79	2.26	6.19	3.70	1.11	V L
Fr+rF	.39	.77	.62	1.24	.22	S
FD	.30	.70	1.29	1.12	1.06	V L
F	16.44	8.32	8.69	5.01	−1.13	V L
Pair	6.23	4.61	6.77	4.58	.12	S
Ego Index	.27	.15	.35	.17	.50	M
Lambda	2.09	1.73	.69	.55	−1.09	V L
Lambda Mdn	1.60	1.73	.50	.55	−.86	L
EA	5.00	3.02	8.05	4.15	.84	L
es	5.26	3.60	10.78	5.83	1.14	V L
D	−.08	.87	−1.06	1.74	−.71	L
AdjD	.19	.80	−.43	1.13	−.63	M
active	3.46	2.79	5.46	3.94	.59	M
passive	.97	1.39	3.08	2.06	1.20	V L
Ma	1.52	1.56	2.38	1.88	.50	M
Mp	.49	.85	1.56	1.42	.91	L
Intell Index	1.80	2.88	2.36	2.28	.22	S
Zf	12.23	5.33	13.12	5.00	.17	S
Blends	1.60	1.70	5.17	3.37	1.34	V L
Afr	.59	.21	.52	.19	−.35	S
P	4.68	1.84	4.99	1.74	.17	S
X+%	.53	.13	.47	.15	−.43	M
Xu%	.22	.10	.29	.15	.55	M
S–%	.13	.19	.31	.28	.75	L
Isolation Index	.22	.15	.23	.13	.07	S
H	2.29	2.07	2.19	1.94	−.05	S
(H)	1.27	1.27	1.21	1.19	−.05	S
Hd	1.15	1.39	1.60	1.76	.28	S
(Hd)	.19	.45	.96	.94	1.04	V L
Hx	.24	.77	.44	.95	.23	S
All H	4.91	3.38	5.96	3.67	.30	S
A	9.23	3.88	7.07	1.19	−.75	L
(A)	.45	.68	.46	1.76	.01	S

(Continued on next page)

& Taylor, 2003). Agreement for protocol-level summary variables is most informative about the reliability of data at the level that it is used for clinical decisions or research about people. The percentages of agreement and iota values for response segments indicated satisfactory interrater reliability scores for the various CS scoring categories in both data sets (see Table 1).

Data collection. After the training for the 2001 data set was completed, two independent judges from the team carefully reviewed all of the protocols that had been collected for the 1998 data set, and no judge reviewed more than 20 protocols. The review showed that most of the 1998 protocols lacked in: (a) inquiry and/or (b) scoring accuracy. Compromised scoring accuracy in the context of good scoring reliability (interrater agreement), as described in the previous paragraph, indicated that the coders were consistent in how they assigned scores, but they did so according to inaccurate scoring rules or inaccurate scoring benchmarks.

The team was particularly frustrated with the inadequate CS inquiry and decided that, as supported by Exner (1986, 1990, 1993), the protocols could not be called CS protocols. We thus decided to throw out these protocols and to collect and to score a new group of adult protocols.

Administration. All the 1998 and 2001 participants received the instruction "What might this be" originally developed by Rorschach (1942) and recommended in the CS (Exner, 1993). The examiners were adequately prepared. Cards were organized and placed as recommended by Exner (1993), meaning that they were in the proper order when face down and out of the reach of the person being tested. Responses were written verbatim. Location sheets readily were available to use during the Inquiry. Subject and examiner were seated side by side. No abbreviations were used when recording responses either in the first or in the second administration. Any question asked by the client during the test was recorded, as was the response of the examiner. Similarly, any comment was recorded. If questions occurred after the test began, the examiners provided brief, nondirective replies (Exner, 1993). A three-column format for responses, inquiry, and scoring was used when recording the responses.

For the 1998 data set, to prepare the subjects, the subject and examiner spent time getting to know each other before beginning the testing. But according to the Italian tradition, the examiners did not specify the name of the test because of

TABLE 2.—Differences in reference data for Italian adults collected in 1998 and 2001 *(Continued)*

	1998 Initial Sample		2001 Corrected Sample			
Score	*M*	*SD*	*M*	*SD*	Cohen's *d*	*d* Label
Ad	2.26	1.85	2.66	.94	.27	S
(Ad)	.07	.29	.24	.95	.24	S
An	1.56	1.61	1.18	3.00	−.16	S
Art	.82	1.05	1.08	.71	.29	S
Ay	.36	.74	.56	2.34	.12	S
Bl	.19	.48	.31	.57	.23	S
Bt	1.66	1.69	1.76	1.40	.06	S
Cg	1.05	1.31	2.18	1.18	.91	L
Cl	.20	.55	.20	1.01	.00	S
Ex	.22	.52	.18	.56	−.07	S
Fi	.46	.73	.39	1.59	−.06	S
Food	.32	.69	.66	1.91	.24	S
Ge	.38	.71	.24	.45	−.24	S
Hh	.33	.75	.70	.50	.58	M
Ls	.48	.83	1.07	.72	.76	L
Na	1.52	1.67	1.14	.96	−.28	S
Sc	.48	.77	1.07	.57	.87	L
Sx	.31	.69	.39	1.02	.09	S
Xy	.25	.59	.18	1.33	−.07	S
Idio	3.01	2.83	1.34	1.50	−.74	L
ALOG	.01	.12	.23	.55	.55	M
CONTAM	.08	.28	.00	.30	−.28	S
WSum6	2.36	4.64	9.08	9.65	.89	L
AB	.31	1.02	.36	.69	.06	S
AG	.44	.77	.36	.61	−.12	S
COP	.37	.66	1.22	1.26	.85	L
CP	.02	.14	.00	.14	−.14	S
MOR	1.23	1.30	1.66	1.45	.31	S
PER	.13	.40	.70	1.02	.74	L
PSV	.66	1.02	.31	.61	−.42	M

S = small effect size.
M = medium effect size.
L = large effect size.
VL = very large effect size.

concern that it would induce anxiety or resistance, and they did not explain the Rorschach in great detail. Even though we instructed examiners not to accept card rejections, examiners did not give encouragement, and they accepted attempted rejections, valuing nondirective behavior on the part of the testing situation. Following Exner's guidelines, however, the final data set contained protocols with two answers on the first card and no card rejections.

For the 2001 data set, we provided a more detailed introductory interview concerning all the assessment procedures to be used, including the Rorschach, as described in detail in Exner (1993). Encouragement and attempted rejections were dealt with as recommended by Exner. If the person gave only one response to Card I, the examiners encouraged the subject by saying, "Take your time and look some more. I'm sure you'll find something else too" (Exner, 2001, p. 6). If the subjects attempted a rejection, reporting that they could not see or find anything in the blot, the examiner was reasonably firm in not accepting the rejection. The problem was circumvented by saying, "Take your time. We are in no hurry. Everyone can find something" (Exner, 2001, p. 6).

Inquiry. For the 1998 data collection we followed Exner's guidelines (Exner, 1986, 1990, 1993). From Exner guidelines and recommendations all the team members were very aware that inquiry procedures were not easy to conduct adequately. The team (theoretically) fully understood that the objective of the Inquiry was for the examiner to see what the client saw, understand where in the blot it was seen, and understand what features of the blots caused it to be seen that way. The Inquiry was introduced following the standard explanation proposed by Exner to ensure that the volunteers understood why the Inquiry was being conducted and what was expected. Next, Exner's procedures were followed for the Inquiry itself. Each response was inquired by first reading verbatim the person's answer. The examiner was familiar with the coding or scoring possibilities as described in the manuals and framed subsequent inquiry questions to resolve scoring ambiguity. Inappropriate questions (Exner, 2001) were never asked because all examiners were aware how they could create unwanted sets that would ruin a potentially valid record.

Nonetheless, despite these seemingly sound procedures, after completing the training for the 2001 data set the team discovered how the Inquiry portion of the 1998 data set generally was inadequate. Only after the subsequent training did we fully understand how different the CS coding system was from the Swiss one and how it was necessary to conscientiously inquire key words or phrases to ensure a good protocol. After the 2001 training we understood how it was impossible to adequately code many responses from the 1998 data set, particularly for determinants, because the inquiry was incomplete. As a result, high Lambda protocols were very common in the 1998 data set (see Table 2).

RESULTS

The results of the two sets of data, collected in 1998 and 2001, respectively, are summarized in Table 2 (see Lis et al., 1998; Lis et al., 2001; Lis et al., 2007/this issue). The table reports the means and standard deviations for both sets of results, as well Cohen's d, which is an effect size index that indicates how many standard deviations apart the two samples are. Cohen (1988) suggested the following guidelines for interpreting the size of d values: small = .30, medium = .50, and large = .80. The differences for many variables are large or very large, and we will describe and discuss these below.

The mean number of responses (R) in the 2001 protocols was lower than in the first administration. The effect size for the difference is small. Nonetheless, this finding indicates that any higher score frequencies in the 2001 sample could not be the result of longer protocols with more responses to score.

Form Quality and M Quality

A large effect size was found for FQo, which decreased in the 2001 data collection. Several other form quality variables had less pronounced shifts, although the overall trend was consistent in indicating more unconventional perceptions in the second data set. This trend also was evident with X+%, which decreased, and Xu%, which increased.

Determinants

Large or very large effect sizes were found for M, FM, m, and FM+m. The mean of all these movement variables increased with the 2001 data collection. Although the effect size was medium for two of the color variables and small for CF, a large change was found for SumC. The mean of SumC increased in the second data collection. Very large changes were found for Sum C', Sum T, and Sum Shading, which all increased in the second round of data collection. Also a very large effect was found for FD, which again increased in the second data collection. A very large difference was found for F, which decreased dramatically in the second data collection. As would be expected, this change produced a parallel change in Lambda, which showed a large decrease in the second data set (using either the mean, listed first in the table, or the median scores, listed second). A very large effect was found for Blends: they increased greatly in the second data collection. Finally, a large and very large effect was found for EA and es, respectively, both of which increased very much in the second administration. A large effect was found also for the D score: its relative value decreased considerably in the second administration.

Ratios, Percentages, and Indexes

As mentioned above, a very large effect was found for Lambda, which decreased considerably during the 2001 data collection. Because the Lambda distribution is skewed, a secondary analysis examined changes in the median Lambda score. This value also decreased substantially in the second data set. Ultimately, the mean and median Lambda values in the 2001 data set are much more similar to the values in Exner (1993) nonpatient adult sample than those from our 1998 data set.

A large change was found for S–%, which increased in the second data collection. A large and very large effect was found for passive human movement and passive movement, both of which increased in the second data collection. The very large difference for passive movement is probably due to two reasons. First, there are more movement responses overall in 2001. Second, there was a change in the proportion of active to passive movement scores assigned in each data set; the proportion of passive scores was lower in the 1998 data set (active:passive about 11:3) than in 2001 (about 11:6).

Contents

A large or very large effect was found for several contents: (Hd), A, Cg, Ls, and Idio. With the exception of Idio, which decreased in the second administration, all the contents increased in the second data collection. In general, the findings suggest that the raters applied secondary content codes more regularly in the 2001 sample than in the 1998 sample.

Special Scores

A large effect was found for WSum6, COP, and PER, all of which increased in the second administration.

Discussion

The results clearly indicate that training and experience with the Inquiry and scoring can have a major effect on CS scores. Adults in the current Italian nonpatient reference sample look more complex, articulate, and engaged with the task than adults in the 1998 sample, although they also have more unusual perceptions and lapses in logic.

These experience effects are not exclusive for the Rorschach test, but it is possible to expect similar improvement with more fully trained examiners using other instruments, such as WAIS or neuropsychological tools. In terms of the Rorschach, our results, however, highlight that not all scoring categories were affected by the refined data collection procedures. Location, developmental quality, and the organized activity categories did not change much between the 1998 and 2001 data sets. The same was true of populars, form quality, pairs and reflections. These findings suggest that these scoring categories can be adequately inquired and scored with less training than is necessary for other types of scores.

Some essential categories vital for accurate interpretation however, were, affected greatly by the additional training we provided examiners in administration and inquiry. Foremost are the categories important for the interpretation of emotional expressivity and control. The scoring of color and shading determinants of nearly all types was reduced in 1998 relative to 2001. As a consequence, the frequency of pure F scores was inflated, and the number of blend responses deflated in 1998 compared with 2001. This means that subjects in 1998 were likely to appear less affectively oriented, more emotionally controlled, and less vital than they should have been if their Rorschach protocols had been administered and inquired accurately.

In the 1998 data set, experience actual (EA) and experience stimulation (es) were attenuated by the respective decrements in color and shading. The D score was affected significantly, however, because of a bigger reduction in shading as compared with the color determinants. Lambda was inflated by the increment in pure F resulting from insufficient inquiry and the underscoring of color and shading.

Another area of CS scoring and interpretation affected by the problematic administration and inquiry procedures was the special score categories that compose most of the weighted index of thought disorder. These results were in line with Exner's hypotheses and with Ritzler and Nalesnik (1990).

As in Ritzler and Nalesnik (1990), the presence of an accurate inquiry significantly reduced the means for pure form (F), and it significantly increased the sum of shading responses (Sum Shading), the sum of color responses (SumC), texture (SumT), achromatic color (SumC′), and a weighted thought disorder measure (WSum6). Unlike Ritzler and Nalesnik, however, we did not find large changes for vague DQv, FQu, diffuse shading (SumY), or vista (SumV).

Many European data sets presented at the Madrid ERA meeting, where we first presented the results for the initial 1998 data set, reported high Lambda values. As a result, many presenters thought that there were cultural differences between European and U.S. samples on this variable. The findings presented here, however, suggest an alternative explanation. After pursuing additional training, we perceived that our 1998 protocols could be judged as being "without sufficient inquiry" as defined by the CS. This impression was supported during the second data collection effort, as our Lambda values decreased dramatically to the point where they were consistent with the values reported by Exner (2001).

These results underline the importance of structured CS training beyond the study and discussion of Exner's text and workbook. We noted that examiners had different levels of understanding of these resources, including a superficial use of the guidelines that did not follow the standard rules and standard use of the CS. We also believe that certain specific methodological characteristics such as the active cooperation of the client, the relational nature of prompting, and other aspects concerning the interactive nature of administration and the Inquiry are particularly difficult to apply without a formal training.

More specifically, after obtaining further CS training, we realized that an active and forthright Inquiry is the only way to get the subject's point of view and to see things in the real way the subject saw them. In both data sets our scoring was reliable, but the quality and richness of the protocols changed dramatically. This was much more important because it has to do with the fundamental nature of the Rorschach data: "Inquiry is one of the most misunderstood and abused features of the Rorschach. . . . When done incorrectly it can muddle a protocol terribly and often generates data that may be of clinical interest but which represents something other than Rorschach data" (Exner, 1995, p. 11).

We believe that our findings can demonstrate that the protocols themselves (i.e., the verbalized transcript of the subjects' response to the blots) changed in a deep and dramatic way. Specifically, we highlight the important role that examiner training can have on the data that are collected. Our 1998 sample compared with the 2001 sample clearly illustrates this issue. Because our data demonstrate substantial differences between these samples, the comparison serves as an appropriate caution for all of us around the world who are collecting this kind of information. In addition, we hope that our decision to throw out all the old data serves as a guide for others to follow if they discover similar limitations affecting their data.

Actually, in the regular course of our clinical practice in Italy, we still often encounter Rorschach protocols collected with the Swiss method as it taught in Italy, which means with little or no recorded inquiry. After our training in the CS we realize these protocols cannot be adequately rescored into the CS. This is most unfortunate when the Rorschach test has been administered to specific categories of patients that have not been studied sufficiently with the CS. We feel very frustrated because the existing protocols cannot be scored for the CS and cannot contribute to a CS evidence base. In essence, the protocols need to be discarded, as happened with the protocols of our first effort to collect an Italian nonpatient reference sample. Although it was relatively easy to collect nonpatient protocols again, this is rarely the same for patient data.

Although the findings presented here suggest that differences in examiner training can have a large impact on the types of CS scores observed in a sample, the current data set is limited because the design of this study was naturalistic. We did not experimentally manipulate training or inquiry expertise and we did not hold other potentially relevant variables constant. As such, it is possible that some of the 2001 versus 1998 differences may be due to genuine differences in the two samples or to other types of artifacts (e.g., differences in scoring conventions). Controlled prospective studies that compare the effects of variations in the administration procedure on various Rorschach variables in nonpatients and patient samples still are needed to validate Exner's (1986, 1993) conclusion that different administration procedures produce different types of test protocols.

REFERENCES

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

American Psychological Association. (1992). *Publication manual of American Psychological Association*. Washington, DC: Author.

Aronow, E., Reznikoff, M., & Moreland, K. (1995). The Rorschach: Projective technique or psychometric test? *Journal of Personality Assessment, 64*, 213–228.

Baughaman, E. (1958). A new method of Rorschach inquiry. *Journal of Projective Techniques, 22*, 381–389.

Beck, S. J. (1944). *Rorschach's Test 1: Basic processes*. New York: Grune & Stratton.

Beck, S. (1953). Comments on "The two tests in the Rorschach" by Levin. *Journal of Projective Techniques, 17*, 475–476.

Blais, M. A., Norman, D. K., Quintar, D., & Herzog, D. B. (1995). The effect of administration method: A comparison of the Rapaport and Exner Rorschach systems. *Journal of Clinical Psychology, 51*, 100–107.

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

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

Exner, J. E. (1969). *The Rorschach system*. New York: Grune & Stratton.

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

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

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

Exner, J. E. (1991). *The Rorschach: A Comprehensive System: Vol. 2. Interpretations* (2nd ed.). New York: Wiley.

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

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

Exner, J. E. (2001). *A Rorschach workbook for the comprehensive system* (5th ed.). Asheville, NC: Rorschach Workshop.

Exner, J. E., Armbruster, G., & Mittman, B. (1978). The Rorschach response process. *Journal of Personality Assessment, 42*, 27–38.

Exner, J. E., & Weiner, I. B. (1995). *The Rorschach: A Comprehensive System: Vol. 3. Assessment of children and adolescents* (2nd ed.). New York: Wiley.

Gibby, R., & Stotsky, B. (1953). The relation of Rorschach free association to inquiry. *Journal of Consulting Psychology, 17*, 359–363.

Grønnerød, C. (1999). Rorschach interrater agreement estimates: An empirical evaluation. *Scandinavian Journal of Psychology, 40*, 115–120.

Hartman, E., Davidsen, P. E., & Molin, P. K. (1999). Administering av Rorschachmetoden. En teoretisk drØfting av ulike instruksjoner [The administration of Rorschach method. A theoretical discussion of different instructions]. *Tidsskrift for Norsk Psykologforening, 36*, 1048–1058.

Hartmann, E. (2001). Rorschach administration: A comparison of the effect of two instructions. *Journal of Personality Assessment, 76*, 461–471.

Hertz, M. R. (1936). The method of administration of the Rorschach ink-blot test. *Child Development*, 7, 237–254.

Hilsenroth, M. J., Handler, L., Toman, K. M., & Padawer, J. R. (1995). Rorschach and MMPI-2 indices of early psychotherapy termination. *Journal of Consulting and Clinical Psychology, 63*, 956–965.

Hunsley, J., & Bailey, J. M. (1999). The clinical utility of the Rorschach: Unfulfilled promises and an uncertain future. *Journal of Psychological Assessment, 11*, 266–277.

Janson, H. (2003, March). *Calculating and reporting Rorschach intercoder agreement*. Half day Workshop conducted at the Midwinter Meeting of the Society for Personality Assessment, San Francisco, CA.

Janson, H., & Olsson, U. (2001). A measure of agreement for interval or nominal multivariate observations. *Educational and Psychological Measurement, 61*, 277–289.

Killingmo, B. (1980). *Rorschachmetode og psykoterapi [The Rorschach method and psychotherapy]*. Oslo, Norway: Universitetsforlaget.

Kligensmith, S. (1956). *A study of the effects of different methods of structuring the Rorschach inquiry on the determinant scores*. Unpublished doctoral dissertation, University of Pittsburgh.

Klopfer, B. (1942). *The Rorschach technique*. Yonkers-on-Hudson, NY: World Book.

Lerner, P. M. (1998). *Psychoanalytic perspectives on the Rorschach*. Hillsdale, NJ: The Analytic Press.

Lis, A., Parolin, L. Salcuni, S., & Zennaro, A. (2007, this issue). Rorschach Comprehensive System data for a sample of 249 adult nonpatients from Italy. *Journal of Personality Assessment, 89*. (Suppl. 1), S80–S84.

Lis, A., Rossi, G., & Prina, S. (1998, August) *The Exner method: Data in a normal Italian adult sample*. Paper presented at the ERA (European Rorschach Association) Congress, Madrid, 28–30.

Lis A., Zennaro, A., Calvo, V., & Salcuni, S. (2001, March). *Italian normative data on adults: A contribution to Exner Comprehensive System*. Paper presented at the Society for Personality Assessment Midwinter Meeting, Philadelphia.

Meyer, G. J. (1999). Introduction to a special series on the utility of the Rorschach for clinical assessment. *Psychological Assessment, 11*, 235–239.

Ritzler, B., & Nalesnik, D. (1990). The effect of inquiry on the Exner Comprehensive System. *Journal of Personality Assessment, 55*, 647–656.

Rorschach, H. (1942). *Psychodiagnostics*. New York: Grune & Strutton.

Shontz, F. C., & Green, P. (1992). Trends in research on the Rorschach: Review and recommendations. *Applied-and-Preventive-Psychology, 1*, 149–156.

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

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

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

Zax, M., & Stricker, G. (1960). The effect of structured inquiry on Rorschach scores. *Journal of Consulting Psychology, 24*, 328–332.