Table 1
An Overview of the 2 Studies, 4 Data Sets, and Issues Addressed in Each

Source	Clinicians	Methodology	Issues Addressed
Study 1			
Data Set A	3 (the authors)	Each clinician interpreted all 55 protocols via ratings of 29 constructs on a 5-point Likert-type scale	1) Reliability for item-level versus aggregated judgments 2) Results across 3 types of reliability statistics 3) Differential use of the rating scale, ipsative scores, and statistical assumptions for reliability 4) Likert-type ratings versus Q-sorts 5) Genuine ratings versus base rate-equated random ratings
Data Set B	3 (the authors)	Each clinician interpreted all 55 protocols via Q-sorts of 29 constructs on a 7-point distribution	
Study 2			
Data Set C	17 (no overlap with Study 1)	Clinicians randomly assigned to interpret 10-11 protocols and to the 1st, 2nd, or 3rd rater position so that each protocol was rated by 3 clinicians on 29 constructs using a 5-point Likert scale	1) Reliability for item-level versus aggregated judgments 2) Impact of a problematic design on observed findings 3) Generalizability of findings from Study 1
Data Set D	8 (also gave ratings in Data Set C)	Same as Data Set A	
Comparative Analyses Across Data Sets A and D	3 - 11	Each clinician interpreted all 55 protocols via ratings of 29 constructs on a 5-point Likert-type scale	1) Individual differences in reliability 2) Agreement with psychometric true scores versus other clinicians 3) Current findings relative to meta-analyses of interrater reliability in psychology, psychiatry, and medicine

Table 2
Rorschach Rating Scale Items Used for Interpretation Across Studies

---

3. This person experiences himself as damaged, flawed, or hurt by life.
5. At least below the surface, this person is very self-critical and has painful feelings about himself.
7. This person strives to maintain an inflated belief in his personal importance or uniqueness (even though this effort may serve to counter feelings of inadequacy or inferiority).
22. This person occasionally reacts to situations with intense, poorly controlled feelings.
24. This person is bothered by distress or irritation that comes from internalizing or "holding in" feelings.
27. This person feels distant or isolated from others.
37. This person does not have a consistent coping style and frequently shifts strategies, reverses judgments, or has difficulty reaching a firm decision.
38. This person oversimplifies situations as a basic way of coping.
39. This person copes with problems by using feelings and intuitions to guide his decisions, judgments, and actions.
44. In general, this person is actively attuned to the environment and makes consistent efforts to organize and synthesize relevant information.
50. This person quickly jumps to conclusions and sizes up situations without sufficient information.
51. This person thinks about, perceives, and recalls events in a diffuse, vague, or impressionistic manner.
57. This person has difficulty shifting attention, thinking flexibly, or understanding events from more than one perspective at a time.
63. This person consistently focuses on abstract or theoretical ideas in order to minimize emotional discomfort.
72. This person relies on internal fantasies or daydreams to comfort himself or to avoid unpleasant realities in life.
86. This person sees things from an unconventional, unique, or idiosyncratic perspective.
90. This person does not perceive even relatively obvious events in a socially conventional way.
91. This person has many occasions when his perceptions of external events are clearly distorted.
92. This person has an inaccurate understanding of people or interpersonal behaviors.
95. This person has frequent and easily recognized disruptions in formal thought processes. These may be evident in a variety of ways, such as through loose associations, illogical reasoning, using words in odd ways, or having ideas that are inappropriately linked together, among other things.
112. This person enjoys social interactions and believes they can be harmonious and supportive.
122. This person has underlying oppositional tendencies and expresses anger by being contrary or resistive.

145. This person tends to perceive other people in unrealistic ways, such that his understanding is based primarily on imaginative or fantasized qualities, rather than upon a complex understanding of their actual characteristics.
155. This person has strong needs for support and nurturance.
157. This person feels lonely and has strong wishes to be emotionally connected with others.
167. This person is introspective.
1.1. This person has social and emotional limitations that make it hard for him to cope with the everyday problems of life. These limitations may be expressed in a depressive sense of helplessness and ineffectiveness, or in social difficulties where he either relies excessively on others or else disregards and avoids relationships.
2.1. This person's thinking is disorganized and his perceptions are inaccurate.
4.1. Based upon internal psychological factors, this person is at risk for suicide.

Note. Numbers indicate RRS items. The last three entries (1.1, 2.1, 4.1) identify a global statement that had multiple subcomponents.

Table 3
Composition and Internal Consistency of Aggregated Interpretive Scales from RRS Items

Construct Description	Alpha	*M* inter-item *r*	RRS items
Factor Analytic Dimensions			
Perceptual Distortions and Thought Disorder[a]	.91	.58	86, 90, 91, 92, 95, 145, 2.1
Negative Emotionality	.84	.57	3, 5, 27, 157
Conceptually Derived Scales			
Perceptual Distortions and Thought Disorder[a]	.91	.58	86, 90, 91, 92, 95, 145, 2.1
General Distress/Dysfunction	.86	.37	3, 5, 22, 24, 27, 122, 155, 157, 1.1, 4.1
Poor Coping	.74	.29	37, 38, 44(R), 50, 51, 57, 112(R)
Defensive Idealization/Intellectualization	.52	.27	7, 63, 72

Note. $n = 213$ observer ratings. (R) indicates a reverse-coded item.
[a] The same aggregate construct scale was created by factor analysis and rational development.

Table 4

Study 1: The Reliability of Aggregated Interpretive Judgments for 3 Clinicians Using a Rating Scale (Data Set A) and Q-Sort (Data Set B) Format Across 55 Rorschach Protocols and for 3 Sets of Randomly Generated Artificial Ratings that Paralleled Those From Data Set A

Construct	Data Set A				Data Set B		
	Genuine Likert Ratings			*M r* for Artificial Ratings	Q-Sorts		
	M r	ICC (C2,1)	ICC (A2,1)		*M r*	ICC (C2,1)	ICC (A2,1)
Perceptual Distortions and Thought Disorder	.86	.83	.68	-.02	.87	.87	.84
Negative Emotionality	.93	.93	.90	-.01	.72	.72	.72
General Distress-Dysfunction	.94	.94	.76	-.05	.83	.83	.79
Poor Coping	.87	.87	.82	-.03	.88	.87	.81
Defensive Idealization-Intellectualization	.78	.77	.64	-.07	.75	.74	.71
Mean	.88	.87	.76	-.04	.81	.81	.77
Median	.87	.87	.76	-.04	.83	.83	.79

Note. *M r* = Mean correlation across 3 rater pairs; ICC = intraclass correlation; (C2,1) = consistency reliability for a single rater (i.e., one rater with another rater); (A2,1) = absolute agreement reliability for a single rater.

Table 5

Study 1: The Reliability of Individual Interpretive Judgments for 3 Clinicians Using a Rating Scale (Data Set A) and Q-Sort (Data Set B) Format Across 55 Rorschach Protocols and for 3 Sets of Randomly Generated Artificial Ratings that Paralleled Those From Data Set A

		Data Set A				Data Set B		
		Genuine Likert Ratings			*M r* for Artificial Ratings	Q-Sorts		
RRS	Item Construct	*M r*	ICC (C2,1)	ICC (A2,1)		*M r*	ICC (C2,1)	ICC (A2,1)
3	feels damaged or hurt	.88	.86	.70	-.19	.79	.78	.73
5	self-critical/pained	.86	.83	.81	.14	.69	.61	.60
7	inflated self-importance	.75	.70	.64	-.10	.71	.70	.65
22	poor affect control	.73	.67	.56	.08	.67	.64	.60
24	distressed/irritated	.73	.72	.52	.00	.67	.67	.63
27	distant or isolated	.87	.86	.86	.07	.72	.71	.64
37	inconsistent coping style	.90	.89	.89	.11	.84	.80	.80
38	oversimplifies to cope	.87	.86	.85	.00	.73	.73	.72
39	feelings guide decisions	.91	.91	.90	-.09	.64	.65	.62
44	actively organizes information	.65	.65	.60	.06	.58	.58	.53
50	jumps to conclusions	.82	.81	.80	-.05	.78	.77	.75
51	thinking is diffuse or vague	.84	.82	.73	.09	.72	.70	.70
57	inflexible thinking	.16	.21	.18	-.12	.34	.37	.31
63	focuses on abstract ideas	.85	.84	.70	-.16	.75	.73	.71
72	relies on fantasy/daydreams	.79	.78	.75	.05	.62	.62	.58
86	sees things unconventionally	.68	.67	.42	.03	.58	.58	.55
90	misses the obvious	.86	.86	.77	-.07	.71	.69	.68

91	distorted perceptions	.79	.73	.57	-.11	.77	.77	.69
92	inaccurate view of people	.84	.82	.78	-.07	.74	.71	.68
95	disrupted thought processes	.86	.84	.82	-.11	.89	.89	.84
112	sees harmonious interactions	.58	.58	.54	.04	.63	.64	.61
122	acts contrary or resistive	.85	.84	.79	-.11	.80	.79	.76
145	fantasized qualities in others	.73	.73	.70	-.08	.67	.64	.58
155	needs support and nurturance	.81	.76	.65	.00	.59	.53	.49
157	lonely/wishes for connection	.96	.96	.96	.00	.71	.71	.69
167	introspective	.78	.75	.75	.04	.64	.63	.64
1.1	generalized coping problems	.85	.82	.81	.00	.88	.88	.88
2.1	poor thinking and perception	.76	.75	.74	.03	.80	.78	.79
4.1	psychic distress/suicide risk	.85	.84	.83	.04	.73	.70	.70
Mean		.79	.77	.71	-.02	.70	.69	.66
Median		.84	.82	.75	.00	.71	.70	.68

Note. *M r* = Mean correlation across 3 rater pairs; ICC = intraclass correlation; (C2,1) = consistency reliability for a single rater (i.e., one rater with another rater); (A2,1) = absolute agreement reliability for a single rater.

Table 6
Clinician Differences When Assigning Interpretive Ratings in Data Set A

	Clinician			Cohen's *d*
	A	B	C	A vs C
Example Item-Level Judgments				
3: feels damaged or hurt	.82	.71	.02	.89
	(0.80)	(0.92)	(0.99)	
22: poor affect control	.82	.49	-.05	.81
	(0.82)	(0.74)	(1.27)	
24: distressed/irritated	.82	.91	-.07	1.01
	(0.82)	(0.87)	(0.92)	
86: sees things unconventionally	1.25	1.00	.22	1.58
	(0.55)	(0.75)	(0.74)	
91: distorted perceptions	1.27	1.15	.40	.93
	(0.71)	(0.71)	(1.12)	
Aggregated Judgments				
Perceptual Distortions/Thought Disorder	6.55	6.93	4.02	.72
	(2.89)	(2.95)	(4.04)	
Negative Emotionality	1.22	1.15	.31	.36
	(2.49)	(2.50)	(2.56)	
General Distress/Dysfunction	5.16	4.60	1.24	.89
	(4.37)	(4.32)	(4.48)	
Poor Coping	.87	-.67	-1.24	.50
	(4.11)	(4.11)	(4.36)	

Defensive Idealization/Intellectualization	1.56 (1.57)	1.07 (1.94)	-.05 (1.81)	.95

Note. The column entries for each clinician are mean ratings (and standard deviations) across 55 patients.

Table 7
Background Characteristics of the Clinicians in Study 2

	Data Set C (17 Clinicians)				Data Set D (8 Clinicians)			
Variable	*M*	Mdn	*SD*	%	*M*	Mdn	*SD*	%
Age	48.3	48	7.8		50.0	49	8.0	
Years in Practice	14.6	15	10.9		15.4	15	9.8	
# of CS Interpretations in Career	480.6	300	510.6		356.2	275	280.9	
# Rorschachs / Month	4.8	4	2.9		5.6	4	3.0	
Ph.D./Ed.D./Psy.D.				94.1				100.0
Male				64.7				50.0
Primarily in Private Practice				88.2				75.0

Table 8
Study 2: Interpretive Reliability (M r) for One Adequate and Two Inadequate Designs

		Adequate Design	Problematic Designs	
Variables		Initial Data Set D[a]	Data Set C[b]	Data Set D After Clinicians Randomly Mixed[c]
Item-Level Judgments				
3	feels damaged or hurt	.77	.62	.60
5	self-critical/pained	.85	.46	.68
7	inflated self-importance	.76	.64	.70
22	poor affect control	.53	.67	.40
24	distressed/irritated	.61	.38	.37
27	distant or isolated	.71	.60	.51
37	inconsistent coping style	.79	.84	.78
38	oversimplifies to cope	.89	.78	.87
39	feelings guide decisions	.75	.67	.80
44	actively organizes information	.61	.34	.61
50	jumps to conclusions	.78	.62	.84
51	thinking is diffuse or vague	.63	.48	.47
57	inflexible thinking	.36	.19	.17
63	focuses on abstract ideas	.85	.64	.63
72	relies on fantasy/daydreams	.68	.51	.58
86	sees things unconventionally	.47	.17	.42

90	misses the obvious	.50	.37	.47
91	distorted perceptions	.70	.52	.55
92	inaccurate view of people	.62	.13	.55
95	disrupted thought processes	.81	.61	.68
112	sees harmonious interactions	.57	.37	.42
122	acts contrary or resistive	.81	.62	.60
145	fantasized qualities in others	.55	.38	.53
155	needs support and nurturance	.76	.61	.54
157	lonely/wishes for connection	.76	.46	.60
167	introspective	.64	.49	.41
1.1	generalized coping problems	.61	.33	.51
2.1	poor thinking and perception	.55	.50	.23
4.1	psychic distress/suicide risk	.68	.64	.38
M r Across 29 Item-Level Judgments		.68	.50	.55
Aggregated Interpretive Judgments				
	Perceptual Distortion/Thought Disorder	.75	.51	.54
	Negative Emotionality	.87	.64	.61
	General Distress/Dysfunction	.88	.69	.60
	Poor Coping	.81	.71	.78
	Defensive Idealization/ Intellectualization	.78	.66	.69
M r Across 5 Aggregated Judgments		.82	.64	.64

[a] Based on 28 sets of pairwise correlations across 55 protocols (1,540 total ratings).
[b] Based on 3 sets of pairwise correlations across 54, 54, and 52 protocols (160 total ratings).

[c] Based on 3 sets of pairwise correlations across 55 protocols (165 total ratings).

Table 9
Individual Differences in Clinician-by-Clinician Interpretive Reliability for Data Sets A and D (M r)

Data Set A Clinician	Data Set A Clinician			Data Set D Clinician	Data Set D Clinician							
	A	B	C		D	E	F	G	H	I	J	K
A	--	.80	.78	D	--	.84	.83	.80	.75	.71	.66	.66
B	**.91**	--	.78	E	**.90**	--	.81	.81	.73	.72	.64	.62
C	**.86**	**.85**	--	F	**.91**	**.89**	--	.77	.73	.71	.63	.58
				G	**.87**	**.90**	**.88**	--	.69	.67	.60	.61
				H	**.85**	**.82**	**.86**	**.83**	--	.67	.58	.52
				I	**.84**	**.85**	**.83**	**.84**	**.81**	--	.57	.55
				J	**.82**	**.78**	**.80**	**.83**	**.76**	**.74**	--	.52
				K	**.83**	**.83**	**.78**	**.81**	**.71**	**.74**	**.71**	--
Summary *M*												
Item-Level	.79	.79	.78		.75	.74	.72	.71	.67	.66	.59	.57
Aggregated	**.89**	**.88**	**.86**		**.86**	**.85**	**.84**	**.83**	**.81**	**.81**	**.76**	**.77**

Note. Raters were designated by letter after they were ordered by their average level of reliability. Coefficients above the diagonals indicate average agreement for 29 item-level judgments, while bolded coefficients below the diagonals indicate average reliability for 5 aggregated judgments. Each coefficient is based on interpretations for 55 patients.

Table 10
Interrater Reliability and Correlations with Psychometric True Scores for Each Clinician in Data Sets A and D

	Interrater Reliability *(M r*)		Correlation with Approximate True Scores *(M r)*	
Clinicians	Item-Level	Aggregated Judgments	Item-Level	Aggregated Judgments
Study 1 -- Data Set A[a]				
A	.79	.89	.87	.94
B	.79	.88	.86	.93
C	.78	.86	.86	.92
Study 2 -- Data Set D[b]				
D	.75	.86	.86	.92
E	.74	.85	.88	.92
F	.72	.84	.85	.91
G	.71	.83	.84	.91
H	.67	.81	.79	.89
I	.66	.81	.78	.88
J	.59	.76	.69	.85
K	.57	.77	.65	.82
Data Set A with Data Set D[c]	.74	.85	.94	.97

[a] Interrater reliability results indicate each clinician's average correlation with the two other clinicians in this sample (55 protocols; 110 total ratings). True score correlations are between each clinician's judgments and the average of the eight Data Set D clinician judgments on 55 protocols.
[b] Interrater reliability results indicate each clinician's average correlation with the seven other clinicians in this sample (55 protocols; 385 total ratings). True score correlations are between

each clinician's judgments and the average of the three Data Set A clinician judgments on 55 protocols.

[c] Interrater reliability results indicate the average of the correlations between the three Data Set A clinicians with each of the eight Data Set D clinicians (55 protocols; 1,320 total ratings). True score correlations are between the average of the three Data Set A clinician ratings and the average of the eight Data Set D clinician ratings (55 protocols; 110 averaged ratings).

Table 11
The Current Findings Relative to Meta-Analyses of Interrater Reliability in the Psychological and Medical Literature

Target reliability construct		*n*(*k*-1) = independent pairs of judgments	Reliability *r*/κ/ICC	
			scale	item
1. Measured Bladder Volume by Real-Time Ultrasound		360		.92[b]
2. Measured Size of Spinal Canal and Spinal Cord on MRI, CT, or X-Ray		200	.90[a]	
		86		.88[a]
3. Count of Decayed, Filled, or Missing Teeth (or Surfaces) in Young Children		113	.97[a]	
		237		.79[c]
4. Rorschach Oral Dependency Scale Scoring		974	.91[b]	
		6,430		.84[c]
5. Scoring the Rorschach Comprehensive System:	Summary scores	784	.91[b]	
	Response segments	11,518		.86[c]
	Scores per response	11,572		.83[c]
6. Neuropsychologists' Test-Based Judgments of Cognitive Impairment		901		.80[c]
7. Hamilton Depression Rating Scale Scoring From Joint Interviews[d]		3,847	.86[b]	

		495		.71[b]
8. Level of Drug Sedation by ICU Physicians or Nurses Check format; judges observe same material??		1,116	.86[b]	
		165		.71[c]
9. Functional Independence Measure Scoring (Joint and Separate Interviews)		1,365	.91[c]	
		1,345		.62[c]
10. TAT Personal Problem-Solving Scale Scoring		385	.85[b]	
11. Rorschach Prognostic Rating Scale Scoring		472	.84[a]	
12. Interpreting the Rorschach CS	Likert Ratings	550	.84[a]	.71[a]
	Q-Sorts	110	.81[a]	.70[a]
13. TAT Social Cognition and Object Relations Scale Scoring		934	.82[b]	
14. TAT Defense Mechanism Manual Scoring		743	.80[b]	
15. Hamilton Anxiety Rating Scale Scoring From Joint Interviews[d]		752	.80[b]	
		214		.72[c]
16. Borderline Personality Disorder (Joint and Separate Interviews)	Diagnosis	402	.82[c]	
	Specific symptoms	198		.64[c]
17. Signs and Symptoms of Temporomandibular Disorder (Separate Examinations)		192	.86[c]	
		562		.56[c]
18. Hamilton Depression Rating Scale Scoring from Separate Interviews		1,012	.82[b]	

		597		.52[b]
19. Therapist or Observer Ratings of Therapeutic Alliance in Treatment (Generally ratings of same session transcripts.)		(S=31)	.78[a]	
20. Job Selection Ratings by Joint Interviews		9,364	.77[a]	
21. Hamilton Anxiety Rating Scale Scoring from Separate Interviews		268	.76[b]	
		208		.58[c]
22. Axis I Psychiatric Diagnosis by SCID in Joint Interviews		216	.75[c]	
23. Type A Behavior Pattern by Structured Interview		(S=3)	.74[a]	
24. Axis II Psychiatric Diagnosis by Semistructured Joint Interviews		740	.73[c]	
25. Personality or Temperament of Mammals (variable observations)		151	.71[a]	
		637		.49[a]
26. Visual Analysis of Plotted Behavior Change in Single-Case Research		1,277		.57[b]
27. Editors' Ratings of the Quality of Manuscript Reviews or Reviewers		3,721		.54[b]
28. Presence of Clubbing in Fingers or Toes[e]		630		.52[c]
29. Stroke Classification by Neurologists		1,362		.51[c]
30. Child or Adolescent Problems:	Teacher ratings	2,100	.64[a]	
	Parent ratings	4,666	.59[a]	
	Externalizing	7,710	.60[a]	
	Internalizing	5,178	.54[a]	
	Direct observers	231	.57[a]	

	Clinicians	729	.54[a]	
31. Job Performance Ratings by Supervisors		1,603	.57[a]	
		10,119		.48[a]
32. Axis I Psychiatric Diagnosis by SCID in Separate Interviews		693	.56[c]	
33. Job Selection Ratings by Separate Interviews		3,185	.53[a]	
34. Axis II Psychiatric Diagnosis by Semistructured Separate Interviews		358	.52[c]	
35. Self and Partner Ratings of Conflict:	Men's aggression	616	.55[a]	
	Women's aggression	616	.51[a]	
36. Determination of Systolic Heart Murmur by Cardiologists		500		.45[c]
37. Abnormalities on Clinical Breast Examination by Surgeons or Nurses		1,720		.42[c]
38. Mean Quality Scores from Two Grant Panels:	Dimensional ratings	2,467		.43[b]
	Yes/No decision	398		.39[c]
39. Job Performance Ratings by Peers		1,215	.43[a]	
		6,049		.37[a]
40. Number of Factors in a Correlation Matrix by Scree Plots[f]		2,300		.35[c]
41. Medical Quality of Care as Determined by Physician Peers		9,841		.31[c]
42. Job Performance Ratings by Subordinates		533	.29[a]	
		4,500		.31[a]

43. Definitions of Invasive Fungal Infection in the Research Literature		21,653		.25[c]
44. Research Quality by Peer-Reviewers:	Dimensional ratings	31,068		.25[b]
	Yes/No decision	4,807		.21[c]

Note. Adapted from Meyer (2004), which provides a complete description of the meta-analytic data sources contributing to this table. CT = computed tomography, ICC = intraclass correlation, ICU = intensive care unit, κ = kappa, MRI = magnetic resonance imaging, *r* = correlation, S = number of studies contributing data, SCID = Structured Clinical Interview for the DSM (Diagnostic and Statistical Manual of Mental Disorders), and TAT = Thematic Apperception Test.

[a] Pearson's *r*

[b] Combination of *r* and κ or agreement ICC

[c] κ or agreement ICC

[d] Category includes videotaped interviews and instances when the patient's report fully determined both sets of ratings (e.g., identical questions in written and oral format).

[e] One study produced outlier results (κ = .90) relative to the others (κ range from .36-.45) so the results should be considered tentative.

[f] Finding should be treated cautiously because agreement varied widely across studies, with values below .10 in several samples but above .70 in several others.