

Sources of Unreliability in Depression Ratings

Kenneth A. Kobak, PhD, Brianne Brown, PsyD, Ian R. Sharp, PhD, Hollie Levy-Mack, MSW, Kurrie Wells, PhD, Felice Ockun, MSW, MS, Janet B.W. Williams, DSW

Background: In multi-center clinical trials, good inter-rater reliability is essential to minimize error variance and improve study power. However, the reasons why raters differ in scoring the same patient have not been thoroughly examined. In 1980, Spitzer and Williams identified several potential sources of unreliability: information variance (different information obtained due to asking different questions); observation variance (the same information is obtained, but raters differ in what they notice and remember); interpretation variance (raters differ in the significance attached to what's observed); criterion variance (different criteria used to score items), and subject variance (true differences in the subject). An empirical examination of the most common sources of unreliability in symptom rating scales has not been published to date. Such a study could help guide rater training efforts and improve the efficacy of such efforts. For example, the scores from raters with limited experience may have more criterion variance, as these raters are less familiar with the scale's scoring conventions than more experienced raters (an example may be one rater who thinks delusions should be rated a 3 on the guilt item and another thinks they should be rated a 4). Experienced raters who have been calibrated to each other may be less likely to disagree on the criterion for scoring each item, and more likely to disagree due to different interpretations of whether the patient's symptom meets the 'threshold' for that criterion (interpretation variance). We videotaped and transcribed 30 pairs of interviews to examine the most common sources of rater unreliability.

Method: Thirty depressed patients were independently interviewed by two different raters on the same day with the SIGH-D, blind to each other's ratings. Raters provided rationales for their scoring, and a panel of expert reviewers read the rationales, the transcripts of the interviews, and reviewed the videotapes when necessary, in order to code the main reason for the discrepancy on each HAMD item. 1/3 of the interviews (n=10 pairs) were conducted by raters who had not administered the HAMD before, (n=5 raters), 1/3 (n=10 pairs)

by raters who had administered the HAMD but were not calibrated (n=6 raters), and 1/3 (n=10 pairs) by raters with both experience and formal calibration training on the HAMD (n=15 raters). A codesheet was developed to include subcategories for information and observation variance, in order to further refine these sources of unreliability. For example, information variance was expanded to include subcategories for why the raters obtained different information, e.g., one rater asked a question the other rater didn't ask, one rater asked the question in a leading manner, etc.

Results: Experienced and calibrated raters had the highest inter-rater reliability (ICC) (r=.93) followed by inexperienced raters (r=.77) and experienced but uncalibrated raters (r=.55). When raters disagreed, the most common reason for disagreement was interpretation variance (35%), followed by information variance (28%), criterion variance (25%), subject variance (8%) and observation variance (4%). Reasons for disagreement were significantly different among the three groups. Experienced and calibrated raters had significantly less criterion variance (5%; 2/42) than experienced raters without calibration (38%;17/45), and raters with no experience (29%;16/55), 2(2)=13.70, p=.001. HAMD items most often disagreed upon across cohorts were psychic anxiety (12%) and agitation (9%).

Conclusion: Reasons for disagreement varied by level of experience and calibration. Experienced and uncalibrated raters should focus on establishing common conventions, while experienced and calibrated raters should focus on fine tuning judgment calls on different thresholds of symptomatology. Calibration training appears to improve reliability over experience alone. Experienced raters without cohort calibration had the lowest reliability, suggesting that exposure to different scoring conventions as a result of experience without rigorous calibration of the cohort may make reliability worse and speaks to the need for better training and calibration methods in clinical trials.



Table 1. Inter-rater Reliability (ICC) by Experience Level

Group	ICC	95% Confidence Interval	F	P
Experienced & Calibrated	.93	.74, .98	26.2	<.0001
Experienced Uncalibrated	.55	-.05, .86	3.4	.0349
No Experience	.77	.33, .93	7.6	.0020

Table 2. Reasons for Disagreement by Rater Cohort

Group	ICC	95% Confidence Interval	F	P
Experienced & Calibrated	.93	.74, .98	26.2	<.0001
Experienced Uncalibrated	.55	-.05, .86	3.4	.0349
No Experience	.77	.33, .93	7.6	.0020

Figure 1. Frequency Distribution of Total Score Differences between Raters by Rater Group

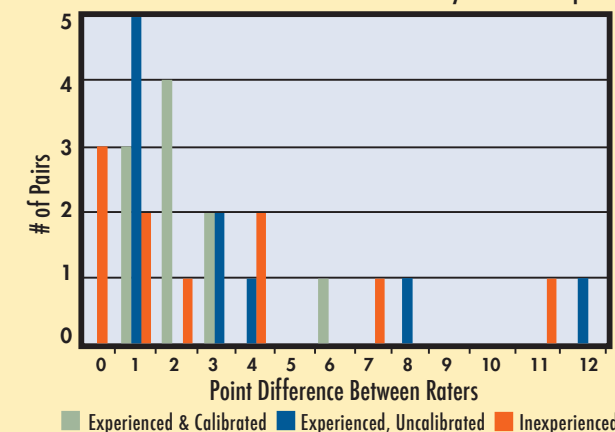


Table 3. Reasons for Disagreement by Sub-Category and Rater Cohort

Group	Reason Code	Reason	Frequency	Percent
Experienced & calibrated	INFORMATION VARIANCE		10	23.8
	1A	Bolded question paraphrased differently	1	2.4
	1B	Follow up questions paraphrased differently	1	2.4
	1D	One rater asked a non-provided follow-up question the other rater didn't ask	4	9.5
	1E	Insufficient follow-up probing resulted in not obtaining critical information	1	2.4
	1F	Insufficient follow-up in order to clarify vague, ambiguous or contradictory information	1	2.4
	1K	Provided follow-up question skipped	1	2.4
	1L	Raters exposed to different information	1	2.4
	OBSERVATION VARIANCE		3	7.1
	2A	When listening to the tape the patient said the same thing to both raters, but one of the raters did not catch what the person said	1	2.4
	2B	When listening to the tape the patient said the same thing to both raters but one rater remembered it incorrectly	1	2.4
2C	One rater failed to notice non-verbal behavior, e.g., tapping foot, tears, etc.	1	2.4	
3	INTERPRETATION VARIANCE	19	45.2	
4	CRITERION VARIANCE	2	4.8	
5A	SUBJECT VARIANCE	8	19.0	
TOTAL			42	100.0
Experienced, not calibrated	INFORMATION VARIANCE		16	35.6
	1A	Bolded question paraphrased differently	1	2.2
	1B	Follow up questions paraphrased differently	4	8.9
	1D	One rater asked a non-provided follow-up question the other rater didn't ask	2	4.4
	1E	Insufficient follow-up probing resulted in not obtaining critical information	1	2.2
	1I	Asked leading question	3	6.7
	1J	Bolded question skipped	3	6.7
	1K	Provided follow-up question skipped	2	4.4
	OBSERVATION VARIANCE		1	2.2
	2B	When listening to the tape the patient said the same thing to both raters but one rater remembered it incorrectly	1	2.2
	3	INTERPRETATION VARIANCE	11	24.4
4	CRITERION VARIANCE	17	37.8	
TOTAL			45	100.0
No experience	INFORMATION VARIANCE		13	23.6
	1B	Follow up questions paraphrased differently	4	7.3
	1D	One rater asked a non-provided follow-up question the other rater didn't ask	3	5.5
	1J	Bolded question skipped	1	1.8
	1K	Provided follow-up question skipped	5	9.1
	OBSERVATION VARIANCE		1	1.8
	2D	Rater forget to include information when scoring	1	1.8
	3	INTERPRETATION VARIANCE	20	36.4
	4	CRITERION VARIANCE	16	29.1
	5	SUBJECT VARIANCE	3	5.5
	Uncodable		2	3.6
TOTAL			55	100.0