

## Sources of Unreliability in Depression Ratings

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

**Background:** Good interrater reliability is essential to minimize error variance and improve study power. Reasons why raters differ in scoring the same patient include information variance (different information obtained because of asking different questions), observation variance (the same information is obtained, but raters differ in what they notice and remember), interpretation variance (differences in the significance attached to what is observed), criterion variance (different criteria used to score items), and subject variance (true differences in the subject). We videotaped and transcribed 30 pairs of interviews to examine the most common sources of rater unreliability.

**Method:** Thirty patients who experienced depression were independently interviewed by 2 different raters on the same day. Raters provided rationales for their scoring, and independent assessors reviewed the rationales, the interview transcripts, and the videotapes to code the main reason for each discrepancy. One third of the interviews were conducted by raters who had not administered the Hamilton Depression Rating Scale before; one third, by raters who were experienced but not calibrated; and one third, by experienced and calibrated raters.

**Results:** Experienced and calibrated raters had the highest interrater reliability (intraclass correlation [ICC];  $r = 0.93$ ) followed by inexperienced raters ( $r = 0.77$ ) and experienced but uncalibrated raters ( $r = 0.55$ ). The most common reason for disagreement was interpretation variance (39%), followed by information variance (30%), criterion variance (27%), and observation variance (4%). Experienced and calibrated raters had significantly less criterion variance than the other cohorts ( $P = 0.001$ ).

**Conclusions:** Reasons for disagreement varied by level of experience and calibration. Experienced and uncalibrated raters should focus on establishing common conventions, whereas experienced and calibrated raters should focus on fine tuning judgment calls on different thresholds of symptoms. Calibration training seems to improve reliability over experience alone. Experienced raters without cohort calibration had lower reliability than inexperienced raters.

**Key Words:** reliability of results, training, depression, methods, Hamilton Depression Scale

(*J Clin Psychopharmacol* 2009;29: 82–85)

In multicenter clinical trials, good interrater reliability is essential to minimize error variance and improve study power. Greater attention is being paid to improving interrater reliability as a result of the increasing rate of failed Central Nervous System trials.<sup>1</sup> Recent efforts at improving interrater reliability include the development and use of semistructured interview guides,<sup>2,3</sup> the development of standardized scoring conventions

for use industrywide across trials,<sup>2</sup> the use of new technologies such as interactive online tutorials to improve and standardize didactic training, the use of videoconferencing to evaluate and improve raters' applied clinical skills,<sup>3,4</sup> and the use of centralized raters<sup>5</sup> (ie, a small group of tightly calibrated expert raters who remotely administer the primary outcome measure to all patients in a multisite clinical trial from a central location via videoconference or teleconference).

The achievement of good interrater reliability requires a deep understanding of the reasons why 2 raters evaluating the same patient obtain different scores. More than 25 years ago, Spitzer and Williams<sup>6</sup> proposed a classification system for the sources of unreliability in diagnostic assessment. These classifications are also relevant for examining the causes of unreliability in the administration of symptom rating scales that are typically used as outcome measures in Child Neurology Society clinical trials, such as the Hamilton Depression Rating Scale (HAMD).<sup>7</sup> These sources of unreliability include (a) information variance, when raters obtain different information as a result of asking different questions; (b) observation variance, when raters differ in what they notice and remember when presented with the same information; (c) interpretation variance, when raters differ in the significance they attach to what is observed; (d) criterion variance: when raters use different criteria to score the same information; and (e) subject variance, when true differences exist in the subject between testings, such as when the patient says different things to each rater or when the patient truly changes between the first and second interview. An empirical examination of the most common sources of unreliability in symptom rating scales has not been published to date. Such a study will help guide rater training efforts and help improve the efficacy of such efforts.

It is possible that these sources vary depending on the experience level of the rater cohort. For example, raters with limited experience conducting the scale may have more criterion variance because they are less familiar with the scale's scoring conventions and the criteria for rating different levels of an item than the 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). Furthermore, experienced raters who have also undergone rigorous training and calibration with each other may have different sources of unreliability than the other 2 groups. For example, experienced raters who have been calibrated to each other may be less likely to disagree on the criteria for scoring each item because they have a common understanding of the scale's conventions, and more likely to disagree on their interpretation of whether the patient's symptom meets the threshold for that criterion (interpretation variance; to use the previous example, they may agree that a delusion should be rated a 4 on guilt but may disagree on whether the patient's thought was a true delusion or just an overvalued ideation). Such information may be valuable in tailoring rater training efforts to better fit a rater cohort's needs. To examine this issue empirically, we videotaped and transcribed 30 pairs of interviews to examine the most common sources of rater unreliability in 3 levels of raters: inexperienced raters (ie, raters who had never administered the scale to patients

From the MedAvante Research Institute, Madison, WI.

Received April 29, 2009; accepted after revision September 15, 2009.

Reprints: Kenneth A. Kobak, PhD, 100 American Metro Boulevard, Suite 106, Hamilton, NJ 08619 (e-mail: kkobak@medavante.net).

Supplemental material or a copy of the complete code sheet is available on the Journal website at [www.psychopharmacology.com/](http://www.psychopharmacology.com/).

This study was supported by funds from MedAvante, Inc. The authors are employees of MedAvante, which conducts methodological research and provides rater training and centralized rater services.

Copyright © 2009 by Lippincott Williams & Wilkins

ISSN: 0271-0749

DOI: 10.1097/JCP.0b013e318192e4d7

before), raters who were experienced (had administered the scale to patients) but were not calibrated to each other, and raters who were both experienced and calibrated to each other.

## METHOD

Thirty depressed patients (major depression,  $n = 26$ ; dysthymia,  $n = 4$ ) recruited from the community were independently interviewed by 2 different raters on the same day with the structured interview guide for the HAMD (SIGHD).<sup>8</sup> Raters were blind to each other's ratings. After completion of both interviews, a project manager compared the scores obtained by both raters and asked the raters to provide written rationales for their scoring on those items for which their scores differed. A panel of independent reviewers ( $N = 6$ ) read the rationales, reviewed the transcripts of the interviews, and watched the videotapes of the interviews when necessary to code the main reason for any discrepancy on each item.

One third of the dyads ( $N = 10$ ) were conducted by inexperienced raters (ie, raters who have never conducted the HAMD before;  $n = 5$  raters); one third ( $N = 10$  dyads), by raters who had administered the HAMD but who were not calibrated to each other ( $n = 6$  raters); and one third ( $N = 10$  dyads), by raters who were both experienced on the HAMD and were formally calibrated with each other ( $n = 15$  raters). Raters in clinical trials vary in their background, experience, and training. We divided the raters into cohorts we believe represent 3 major categories of raters in clinical trials. Although many sponsors require that raters have prior experience on the scale, 1 study found that up to 25% had never administered the scale before.<sup>9</sup> Thus, an inexperienced category seemed appropriate. Experienced but not formally calibrated also seemed appropriate because most raters who have experience using the HAMD have only scored together at study startup meetings.<sup>10</sup> Reliability testing at such meetings typically does not involve rigorous calibration procedures. The most commonly used standardization method, observing and scoring videotapes, artificially inflates reliability by reducing the sources of variability described previously. It does not inform us as to whether 2 raters from different sites would obtain the same score if they each independently evaluated the same patient. Although such rigorous calibration methodology is more accurate, it is rare, if ever used, in multicenter trials because of the practical limitations in pairing large numbers of raters across sites and having them conduct independent interviews with patients. Even when using the less rigorous method of tape observation, achieving adequate levels of interrater reliability has generally not been successful.<sup>4,11</sup> Thus, a cohort of raters who have conducted the HAMD but have not been rigorously calibrated to each other using independent interviews seemed a logical cohort to include and a good proxy for most raters in traditional clinical trials. The cohort of raters who received formal calibration procedures were trained using methods previously described<sup>4</sup> and included didactic training on scoring conventions, applied training on clinical skills, and formal calibration using independent interviews. All interviews were conducted using the SIGHD.<sup>8</sup> All raters in the study had either a masters or doctoral degree in psychology, social work, or a related mental health field.

### Development of Coding Sheet

We expanded the categories of information and observation variances to include subcategories to further refine the sources of unreliability (Table 1). For example, we expanded information variance to include subcategories, providing a more detailed explanation for why the raters obtained different information; for

example, one rater asked a question the other rater did not ask, one rater asked the question in a leading manner, and so on. The SIGHD includes an instruction that raters ask all bold questions verbatim and use provided follow-up probes to gather more information. Thus, we were able to code more specifically whether the reason for obtaining different information was because of the bold questions being skipped, the follow-up questions being skipped or reworded, and so on. Similarly, for observation variance, we created subcategories to further delineate the specific cause for the differences in observation; for example, "patient presented the same information to both raters, but one rater did not remember the information when scoring"; "patient presented the same information to both raters, but one rater remembered it incorrectly"; "one rater failed to notice nonverbal behavior, for example, foot tapping" that was present in both interviews; and so on. When more than 1 source of disagreement occurred, the coders used the rater's written rationale and a review of the transcript and videotape to identify the most likely source of the disagreement.

### Calibration of Reviewers

The 6 videotape reviewers had either a doctoral degree in psychology ( $N = 4$ ) or masters in social work ( $N = 2$ ) and were experienced in the scoring and administration of the HAMD (mean, 9.3 years). Calibration of the tape reviewers on the coding of the reasons for unreliability was accomplished through a 3-step process: (1) Independent coding by all reviewers followed by a group discussion and consensus coding ( $N = 10$  tapes); (2) independent coding by all reviewers followed by a discussion and consensus coding in dyads and a group discussion with all coders and consensus coding by the entire group ( $N = 4$  tapes); and (3) once the cohort of coders reached an a priori-defined level of agreement (80% of raters achieved 100% agreement for the codes on all items), coding was then done in dyads, for example, each coder independently coded the tape, followed by a discussion and consensus coding ( $N = 16$  tapes).

## RESULTS

### Total Sample

Overall, the 30 dyads evaluated 510 HAMD items (ie, 30 interviews, each with 17 items). Raters across groups agreed on 368 (72%) of the items. The reasons for disagreement on the remaining 142 items included interpretation variance (35%), information variance (27%), criterion variance (25%), subject variance (8%), and observation variance (4%; 2 items were uncodable because of recording error [1%]). Differences in how follow-up questions were paraphrased and differences in which follow-up questions were asked were the most common type of information variance (6.3% for each subcategory). The HAMD items on which raters across groups most often disagreed were psychic anxiety (12% of disagreements) and psychomotor agitation (9% of disagreements).

### Results by Cohort

Experienced and calibrated raters had higher interrater reliability ( $r = 0.93$ ) than raters who were experienced but not calibrated ( $r = 0.55$ ;  $P = 0.056$ ) and inexperienced raters ( $r = 0.77$ ;  $P = 0.246$ ), although the difference was not statistically significant. Each rater cohort administered a total of 170 items (ie, 10 interviews, each with 17 items). When omitting uncodable items and subject variance (ie, variance unrelated to rater cohort), raters who were experienced and calibrated disagreed on 34 (20%) of the 170 items they administered, raters who were trained and uncalibrated disagreed on 45 (26%) of 170 items, and inexperienced raters disagreed on 50 (29%) of

**TABLE 1.** Sources of Unreliability by Group and Sub-Categories

Group	Reason	Frequency	Percent
Experienced & calibrated	Information Variance	10	29.4
	Bolded question paraphrased differently	1	2.9
	Follow up questions paraphrased differently	1	2.9
	One rater asked a non-provided follow-up question the other rater didn't ask	4	11.8
	Insufficient follow-up probing resulted in not obtaining critical information	1	2.9
	Insufficient follow-up in order to clarify vague, ambiguous or contradictory information	1	2.9
	Provided follow-up question skipped	1	2.9
	Raters exposed to different information	1	2.9
	Observation Variance	3	8.8
	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	2.9
	When listening to the tape the patient said the same thing to both raters but one rater remembered it incorrectly	1	2.9
	One rater failed to notice non-verbal behavior, e.g., tapping foot, tears, etc.	1	2.9
	Interpretation Variance	19	55.9
	Criterion Variance	2	5.9
	Total	34	100.0
	Experienced, not calibrated	Information Variance	16
Bolded question paraphrased differently		1	2.2
Follow up questions paraphrased differently		4	8.9
One rater asked a non-provided follow-up question the other rater didn't ask		2	4.4
Insufficient follow-up probing resulted in not obtaining critical information		1	2.2
Asked leading question		3	6.7
Bolded question skipped		3	6.7
Provided follow-up question skipped		2	4.4
Observation Variance		1	2.2
When listening to the tape the patient said the same thing to both raters but one rater remembered it incorrectly		1	2.2
Interpretation Variance		11	24.4
Criterion Variance		17	37.8
Total		45	100.0
No experience	Information Variance	13	26.0
	Follow up questions paraphrased differently	4	8.0
	One rater asked a non-provided follow-up question the other rater didn't ask	3	6.0
	Bolded question skipped	1	2.0
	Provided follow-up question skipped	5	10.0
	Observation Variance	1	2.0
	Rater forget to include information when scoring	1	2.0
	Interpretation Variance	20	40.0
	Criterion Variance	16	32.0
	Total	55	100.0

170 items ( $F_2 = 1.36$ ;  $P = 0.273$ ). The amount of variability between raters (ie, the SDs of the mean differences between raters) for the calibrated group was lesser ( $SD = 2.7$ ) than for the experienced but uncalibrated group ( $SD = 5.2$ ) and the inexperienced group ( $SD = 4.3$ ). Nine percent of the dyads (9 of 10) were within 3 points of each other on total score in the calibrated group, compared with 70% (7 of 10) in the experienced and uncalibrated group and 60% (6 of 10) in the inexperienced cohort.

The reasons for disagreement differed among the 3 groups. The experienced and calibrated raters were less likely to disagree because of criterion variance (6%; 2/34) than experienced raters without calibration (38%; 17/45) and raters with no experience

(32%; 16/50). Psychic anxiety was the item most often disagreed on by both inexperienced raters (10.9% of their disagreements) and experienced and uncalibrated raters (11.1% of their disagreements [tied for first]). It was the second most common item disagreed upon by experienced and calibrated raters (14.3% of disagreements), where agitation was the most commonly discrepant item (16.7% of disagreements).

## DISCUSSION

The main source of rater unreliability varied by experience level. Experienced and calibrated raters disagreed significantly less because of criterion variance, undoubtedly because part of the calibration process includes establishing common criteria

that the cohort will use and apply. The main reason for disagreement in this cohort was interpretation variance, which suggests that once raters agree on the common criteria, when they do differ, it is more often because of differences in judgment as to whether the patient meets the threshold for that criterion rather than disagreement as to the use of different criteria per se. Thus, when training a cohort of calibrated raters, it may be best to focus the training on fine tuning judgment calls on different thresholds of symptoms. The experienced and uncalibrated raters disagreed most often because of criterion variance. Raters at different sites are trained using different sets of conventions, and different studies by different sponsors often use different sets of conventions.<sup>12</sup> Thus, training-experienced and uncalibrated raters should initially focus on establishment of common scoring conventions and criteria. Experienced and uncalibrated raters also had a high percentage of disagreement due to information variance, suggesting that standardizing the way experienced raters ask questions is critical in achieving reliability. In a study of the impact of rater quality and signal detection, adherence to a structured interview guide was one of the biggest predictors of separation of an active drug from placebo.<sup>13</sup> Interestingly, raters with no experience administering the HAMD disagreed most often because of interpretation variance rather than criterion variance. This might be explained by the fact that they had not gone through the experience of being exposed to diverse sets of scoring conventions and thus tended to focus more strictly on the anchor points contained in the SIGHD interview guide.

The ICC for the experienced and calibrated group (0.93) was higher than the ICC for the experienced and uncalibrated group (0.55), which had the lowest ICC of the 3 cohorts. This suggests that calibration training seems to improve reliability over and above experience alone. In fact, prior exposure to different scoring conventions as a result of experience without rigorous calibration of the cohort may make reliability worse, as noted by the comparison of the experienced and uncalibrated ICC (0.55) to the ICC of the inexperienced cohort (0.77). The ICC for the trained and uncalibrated cohort (assessed in this study using the more rigorous method of independent interviews) may be a rough estimate of what the true ICC is for raters in clinical trials in general and speaks to the need for better training and calibration methods in clinical trials.

It is interesting to note that although the actual number of disagreements was not dramatically different between the calibrated, uncalibrated, and inexperienced cohorts (34, 45, and 50, respectively), the ICCs do seem to be meaningfully different (0.93, 0.55, and 0.77, respectively). This suggests that although uncalibrated raters may not disagree significantly more often than calibrated raters, when they do disagree, the magnitude of disagreement may be greater. This hypothesis is supported by the larger SDs of the difference scores between raters for the uncalibrated (SD = 0.43) and the inexperienced (SD = 5.2) groups compared with the calibrated cohort (SD = 2.7) and by the wider range of discrepancies in the latter 2 cohorts.

The use of structured interview guides in general enhances reliability,<sup>14</sup> most likely by reducing information variance through the use of standardized questions and reduced criterion variance due to more explicit scoring anchors. In the current study, all 3 groups used the same structured interview guide (SIGHD). It is likely that conducting the HAMD via an unstructured interview would have decreased reliability and increased information variance. The GRID-HAMD<sup>2</sup> was recently developed through a broad-based, international consensus process that contains even more structured questions to be asked, finely delineated anchor points, and specific scoring algorithms

involving frequency and severity. It would be interesting to see if its use would have a differential impact on reliability, for example, by reducing information and criterion variance in the experienced but not calibrated group, who are more likely to have developed idiosyncratic ways of probing and scoring.

Results of this study point to the need for improved methodology in establishing interrater reliability when training raters. Reliability should be evaluated using independent interviews, and raters should debrief to identify the causes for their disagreements. Such an analysis will help identify, for example, whether there is systematic misunderstanding of conventions within the rater cohort or whether differences are due to errors in clinical administration, such as variability in how questions are asked, resulting in raters obtaining different information. The former can be remedied by better clarification of the scoring conventions, whereas the latter can be remedied by stricter adherence to the interview guide and standardization of probe or follow-up questions. Audiotaping or videotaping of interviews in combination with rater debrief is a useful way to identify sources of unreliability in both initial calibration efforts and remediation efforts to control rater drift.

## REFERENCES

1. Khan A, Brodhead AE, Kolts RL, et al. Severity of depressive symptoms and response to antidepressants and placebo in antidepressant trials. *J Psychiatr Res.* 2005;39:145–150.
2. William JBW, Kobak KA, Bech P, et al. The GRID-HAMD: standardization of the Hamilton Depression Rating Scale. *Int Clin Psychopharmacol.* 2008;23(3):120–129.
3. Kobak KA, Opler MG, Engelhardt N. PANSS rater training using Internet and videoconference: results from a pilot study. *Schizophr Res.* 2007;92:63–67.
4. Kobak KA, Engelhardt N, Lipsitz JD. Enriched rater training using Internet-based technologies: a comparison to traditional rater training in a multi-site depression trial. *J Psychiatr Res.* 2006;40:192–199.
5. Kobak KA, Kane JM, Thase ME, et al. Why do clinical trials fail? The problem of measurement error in clinical trials: time to test new paradigms? *J Clin Psychopharmacol.* 2007;27:1–5.
6. Spitzer RL, Williams JBW. Classification in Psychiatry. In: Kaplan HI, Freedman AM, Sadock BJ, eds. *Comprehensive Textbook of Psychiatry/III.* Baltimore: Williams & Wilkins; 1980:1035–1072.
7. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry.* 1960;23:56–62.
8. Williams JBW. A structured interview guide for the Hamilton Depression Rating Scale. *Arch Gen Psychiatry.* 1988;45:742–747.
9. Bullinger A, Targum SD. Rater experience in CNS clinical trials. In: National Institute of Mental Health, New Clinical Drug Evaluation Unit, 44th Annual Meeting; 2004; Phoenix, AZ.
10. Kobak KA, Engelhardt N. Standardized training on the Hamilton Depression Scale using Internet-based technologies. In: *Drug Information Association 39th Annual Meeting*; 2003; San Antonio, TX.
11. Demitrack MA, Faries D, Herrera JM, et al. The problem of measurement error in multisite clinical trials. *Psychopharmacol Bull.* 1998;34:19–24.
12. Williams JBW. Standardizing the Hamilton Depression Rating Scale: past, present, and future. *Eur Arch Psychiatry Clin Neurosci.* 2001;251:II/16–II/12.
13. Kobak KA, Feiger AD, Lipsitz JD. Interview quality and signal detection in clinical trials. *Am J Psychiatry.* 2005;162:628.
14. Moberg PJ, Lazarus LW, Mesholam RI, et al. Comparison of the standard and structured interview guide for the Hamilton Depression Rating Scale in depressed geriatric inpatients. *Am J Geriatr Psychiatry.* 2001;9:35–40.