

## Rater Training in Multicenter Clinical Trials: Issues and Recommendations

*Kenneth A. Kobak, PhD,\* Nina Engelhardt, PhD,†  
Janet B. W. Williams, DSW,‡ and Joshua D. Lipsitz, PhD\*†*

The growing rate of failed clinical trials in neuroscience has led to increased attention being paid to methodologic factors that may contribute to this failure.<sup>1,2</sup> An issue that has been largely overlooked is that of rater training and rater competency. Given that scores on clinician-administered symptom rating scales form the foundation on which the success of a study is built, it is surprising that so little attention has been paid to this issue. There are several issues related to rater training and rater competency, which warrant examination. These include who is qualified to rate, what components should be included, how and when training should be provided, and, perhaps most important, whether rater training is effective.

### 1. Who is qualified to administer outcome measures in clinical trials?

There is great variety in the backgrounds and experience of persons administering rating scales in clinical trials conducted in the United States, ranging from psychiatrists to study coordinators with bachelor's degrees (often in fields unrelated to psychiatry) with little, if any, clinical experience. Academic credentials alone fail to insure competence in this area, because few formal academic programs include training on the use of the clinician-administered rating scales that are typically used in clinical trials. Most training that does occur takes place at the investigative site. Despite the lack of empirical data supporting acquisition of a specific set of rater skills, we believe that the following general skills are essential to conduct a competent clinical interview using clinician-administered symptom rating scales:

- *Conceptual understanding.* Raters should have didactic training in psychopathology (particularly in the disorder of interest), so they have a good conceptual understanding of the constructs being evaluated. An example of such training would be a course in psychopathology that covers current theories of depression, including diagnostic constructs and criteria.
- *Clinical experience.* Raters should have enough clinical experience with patients who have the disorder being evaluated at all levels of severity to recognize and judge the severity of each of the symptoms rated in the scale (eg, "if he does not know what retardation is, he will be unable to recognize it when it is present and unable to rate it"<sup>3</sup>). Unfortunately, some raters get little training before seeing patients in clinical trials and often learn on clinical trial patients, with clinical trial data. Often, it is their first exposure to patients with the disorder being studied.

\*Research Training Associates, Madison, WI; †Indiana University, Bloomington, IN; ‡New York State Psychiatric Institute and Columbia University, New York, NY.

Address correspondence and reprint requests to Kenneth A. Kobak, PhD, Research Training Associates, 7601 Ganser Way, Madison, WI 53719. E-mail: Kobak@charter.net.

Copyright © 2004 by Lippincott Williams & Wilkins  
ISSN: 0271-0749/04/2402-0113

- *Generic interviewing skills.* Raters should possess good generic clinical interviewing skills (eg, knowing how to pace the interview and probe appropriately to obtain the necessary information and how and when to use open-ended and close-ended questions without leading the patient and allowing the patient to describe their experience in their own words).
- *Expertise in symptom rating scale administration.* Understanding the appropriate use of symptom rating scales is also essential. Hamilton describes specific issues that should be addressed in training raters on the administration of rating scales, including the need to score items independently, avoiding the “halo effect” (ie, the tendency to rate an item high because another item was rated high) and avoiding “response set” (eg, the tendency to rate toward central tendency or, the opposite, to rate only at the extremes<sup>4</sup>).
- *Understanding of research milieu.* Understanding of the research milieu is necessary if the rater is to properly orient patients to their role as research participants. Both the rater and patient need to clearly understand the difference between research and treatment (ie, not to insinuate or promise gains with treatment) and understand that it is just as important to discover that a medication does not work as it is to find out that it does. Such an understanding should reduce “demand characteristics,” (eg, patients reporting progress to please the rater or raters noting progress to please the sponsor of the study). While it is essential to establish good rapport with the patient, the rater must be careful not to cross the boundary into psychotherapy, which could confound the experimental intervention.
- *Scale-specific expertise.* Finally, raters should be well versed in the specific rules for the scoring and administration of the scale being administered. This is especially problematic with the Hamilton Depression Rating Scale (HAMD), because Hamilton failed to provide specific probes and explicit scoring guidelines (see below).

2. What type of training should be provided to novice raters, and when should it be provided?

The training that is typically provided before the start of a clinical trial involves the showing of 1 or more videotaped clinical interviews. Raters watch the videotaped interview, rate it, and participate in a review of responses (with rationale given for correct and incorrect answers) and scale conventions. This is done either “live” at a startup meeting or remotely on a computer. Interrater reliability is assessed by observing 1 or more videotaped clinical interviews following training.

There are several shortcomings to this method. First, rating videotaped clinical interviews does not test a rater’s ability to elicit and interpret clinical information. Used alone, videotaped interviews at best provide an estimate of a rater’s

knowledge of scale conventions but reveal nothing about a rater’s skill at actually administering the scale. Second, this is a flawed method for determining interrater reliability. Two people passively watching a video done by a third person (usually an expert) can achieve artificially inflated interrater reliability because of the reduction in “information variance” that would occur if the 2 raters interviewed the patient independently.<sup>5</sup>

### RECOMMENDATIONS AND FUTURE DIRECTIONS: A MODEL FOR RATER TRAINING AND MONITORING IN CLINICAL TRIALS

The following is a recommended model for rater training and monitoring of the administration of symptom rating scales in clinical trials. The model is characterized by the following attributes. It is applied before the trial startup meeting to ensure that all raters are qualified to rate well in advance of the first patient visit; it includes didactic and applied training; it incorporates testing to evaluate and document competence based on predefined standards; and it includes ongoing monitoring for quality control, calibration, and rater drift. The model consists of the following components.

- *A standard set of scoring conventions.* Several symptom rating scales, such as the HAMD, fail to provide a standardized set of probes and scoring conventions.<sup>6</sup> As a result, several sets of scoring conventions and interview guides have been developed, making interrater reliability across sites difficult to achieve. Raters are faced with the daunting challenge of having to use different guidelines in different simultaneous studies. This dilemma is most apparent with respect to pharmaceutical industry–sponsored clinical trials. Until the industry adopts a standard set of conventions to be used across studies, the problem of competing scoring conventions will remain unresolved, and little progress can be made in achieving interrater reliability among raters at diverse sites in multicenter trials. Recent efforts in this direction have been undertaken. The Depression Rating Scale Standardization Team was formed in 1999 by individuals in academia, clinical practice and research, pharmaceutical industry, and government (National Institute of Mental Health) to develop a standard approach to administering and scoring the HAMD.<sup>7</sup> Reliability of the GRID-HAMD is good for both item and total score,<sup>8</sup> and further validation studies are currently under way.
- *Use of a structured or semistructured interview guide.* The use of a semistructured interview guide is recommended for both reliability and validity. Studies have shown that use of a semistructured interview guide increases item reliability<sup>9,10</sup> and also likely improves validity by helping insure that all questions are asked of all patients. Semistructured guides standardize the initial probes, thus

reducing the amount of “information variance” while allowing raters to augment the standardized probes with improvised follow-up questions as necessary. Interviewing time is not increased, and training is greatly facilitated by an interview guide.

- *Didactic training.* Didactic training includes both a thorough review of general conventions for administering symptom rating scales (eg, avoiding the “halo effect” and “response set”) and a review of the specific scoring conventions for the scale being employed.
- *Testing of didactic knowledge.* Effective training must involve some feedback loop, where the degree of the learner’s comprehension is evaluated. This may be accomplished with multiple choice or true-false tests based on clinical vignettes of real-life patient scenarios. These vignettes would test the trainee’s conceptual understanding of both scoring conventions of the scale and general conventions for assessing patients.
- *Applied training.* Once the trainee has a thorough conceptual understanding of the scale, the next step is to learn how to apply this knowledge in conducting actual clinical interviews with patients. Training in applied skills is an essential but widely overlooked training component. For example, in a recent survey on how raters learned to conduct the HAMD, only 38% of those polled reported having been observed conducting a HAMD as part of their training.<sup>11</sup> Applied training is both time and labor intensive but is nonetheless essential. Applied training may start with observation and modeling but must include the trainee administering the rating scale with observation and feedback. Discussion should include not only rationale for scoring but also feedback on interviewing technique and skills (eg, the use of open-ended vs. leading questions, clarification, summarizing, etc). Training should focus not only on the assessment of acutely symptomatic patients entering clinical trials but also on skill in assessing patients as they improve (or worsen) over time. Reliance on the training conducted at startup meetings (typically lasting 1 to 4 hours) cannot possibly achieve the goal of preparing raters to conduct independent interviews. Clearly, significant rater training needs to take place before the startup meetings. Use of the Internet or videoconferencing technology can facilitate such training when it is logistically impossible for the trainer to meet with the trainee face to face. The number of training interviews necessary to obtain proficiency before testing will vary depending on the background and clinical skill of the trainee. An experienced clinician with a thorough didactic understanding of the scale and the disorder being evaluated may still need several training sessions when conducting the scale for the first time.
- *Testing of applied skills.* As with didactic training, there needs to be a process where the trainee’s applied skills

are evaluated with actual or standardized patients. This should include not only the evaluation of the degree to which scores obtained by the trainee agree with the “gold standard” but also their demonstration of good interview techniques. Recently, a scale was developed to assess rater behaviors along several dimensions, including adherence to the interview guide, use of follow-up probes to elicit further information, clarification of ambiguous or contradictory answers, and neutrality.<sup>12</sup>

- *Testing the efficacy of the training intervention.* Pretesting and posttesting should be conducted to empirically examine whether the training was effective. Testing should include evaluation of (a) improvements in conceptual understanding of scoring conventions, (b) improvement in accuracy (ie, the degree to which scores obtained by the trainee agree with the “gold standard”), (c) improvements in quality (eg, interview skill and technique), and (d) improvement in interrater reliability (ie, the degree to which scores within the training cohort agree with one another). As previously discussed, testing on the latter should be done using independent interviews of patients by the trainees and not the passive observation and scoring of videotaped interviews.
- *Posttraining monitoring for interview quality and rater drift.* Even trainees who achieve high levels of competence and reliability with each other before the trial tend to “drift” over time in how they administer and score an interview. Ongoing monitoring and calibration during the trial ensure that raters both stay on track with each other and continue to conduct the interview correctly. This may be accomplished by ongoing monitoring of audiotapes with feedback to the rater and by group “refresher sessions” in which trainees take turns interviewing a real or “mock” patient with feedback and group discussion. The use of digital tape recorders enables recordings to be sent immediately over the Internet as e-mail attachments for review and feedback. This allows for both a quality review of actual study interviews and a quick correction of any existing problems. The American Society of Clinical Psychopharmacology recently recommended the use of audiotape monitoring for ongoing quality control.<sup>2</sup>

Ongoing monitoring may also help attenuate demand characteristics and other subtle forms of rater bias. Several studies comparing clinician version and self-report (ie, computer-administered) versions of the HAMD and the Hamilton Anxiety Scale show a large “inflation” of scores at baseline in studies that require a minimum score for entry, which subsequently drops once patients are randomized.<sup>13,14</sup>

Because ongoing monitoring of interview quality is not often implemented, there is little empirical data on the quality of interviews conducted in clinical trials. However, the little evidence that exists is not encouraging. In a study of baseline (visit 1) HAMD interviews audiotaped during a multicenter

trial, 77% of tapes were rated unsatisfactory or fair on adequate follow-up questioning, 58% were unsatisfactory or fair on adequacy of information obtained, and 68% rated unsatisfactory or fair on adherence to the structured interview guide.<sup>15</sup> In this study, these numbers had real significance: when raters who were rated fair or unsatisfactory on clarification were dropped from the analyses, the study compound in question separated from placebo, whereas before it did not. In spite of Hamilton's suggestion that the interview should take at least a half hour, the mean time to administer the HAMD in this study was 13 minutes (range, 3 to 35 minutes).

### 3. Is rater training effective?

Despite the near-universal use of some form of rater training in clinical research trials, there is a paucity of empirical research on the effectiveness of rater training. Early studies by Gibbon et al<sup>16</sup> demonstrated that comprehensive training could improve interrater reliability in diagnostic interviewing. Their model included many of the components described above. Tracey et al<sup>17</sup> used an expanded version of this model to establish and maintain good interrater reliability in a 9-site multicenter trial examining treatment of tardive dyskinesia. Training limited to startup meetings has proved less favorable. A study by Demitrack et al<sup>18(p.20)</sup> found large differences in agreement at posttesting and concluded that "there was no evidence of improved rating performance across the 6 hours of reliability training."

More recent attempts at comprehensive training conducted prior to a startup meeting have been proven more successful. In an innovative program sponsored by Boehringer Ingelheim, 21 raters from 6 Canadian sites met for 2.5 days of intensive HAMD training at a rural location where trainees would not be distracted. Correlations between HAMD scores conducted by raters during the trial and the 3 HAMD experts who reviewed the audiotapes were 0.93, 0.94, and 0.89, respectively.<sup>19</sup> The trial also had a larger than usual drug-placebo separation (6.2 points).

The National Institute of Mental Health recently funded a pilot study to examine the use of new technologies for rater training.<sup>20</sup> A Web-based, interactive rater education program on the HAMD was developed specifically to train raters at diverse sites in multicenter trials. The program also included remote teaching and observation of trainees' applied skills using videoconferencing. Result found a significant increase in didactic knowledge pretesting to posttesting, and interrater reliability (intraclass correlation) based on trainees' actual interviews (as opposed to ratings of a videotaped interview) was 0.97 ( $P < 0.0001$ ). In a National Institute of Mental Health-funded follow-up study, this training methodology was tested against "traditional" rater training in the context of a multisite clinical trial.<sup>7</sup> Trainees in the enriched group improved significantly on both their didactic knowledge and interview quality (assessed by a blind rater), while the

traditionally trained group did not improve in either skill. Trainees rated the enriched program positively, and all stated they see similar training on other rating scales.

Others have also employed new technologies to enhance interrater reliability, with positive feedback from trainees. For example, Shayegan and Stahl<sup>21</sup> developed a program to enhance reliability by use of audience response keypads at startup meetings to develop consensus on scoring conventions. This was followed up with live, interactive sessions on the Internet following the startup meeting to monitor rater drift. They found this approach to have improved reliability compared with traditional videotape observation without feedback.

## THE NEED FOR RATER CERTIFICATION

Typically, the responsibility for training raters has fallen to the individual investigative sites. However, there needs to be a way to standardize training across sites and across studies to ensure that a diverse group of raters participating in multicenter trials are all administering the rating scale in the same manner as well as to ensure that raters participating in a trial are competent. Thus, there is a need for some type of centralization of the training process. Obviously, this cannot be accomplished until the field adopts a single version of the scale being used (eg, the HAMD) with accompanying conventions. That said, there needs to be some type of certification process. If a single set of objective standards and standardized modes of evaluation of rater competence are developed and accepted by the field, then diverse groups may conduct rater training, and the results of the training can be objectively evaluated. Perhaps, a centralized and independent "credentialing" body can be developed to review the training programs offered by different training groups, similar to the accreditation process of universities. Such an accreditation body may be housed in a nonprofit, neutral vehicle.

Rater certification may involve several levels (eg, initial certification and annual "checkups" to ensure that raters have maintained their skills). With such a process, raters can be certified across several trials their site may be participating in, and the need to learn various sets of rules and scoring conventions would be eliminated. Pharmaceutical companies would be freed from the responsibility of rater training while at the same time be assured that raters participating in their trials have met minimum standards of competency.

## CONCLUSION

Given the increasing number and cost of failed clinical trials, there is an urgent need to improve and standardize clinical trial methodology and improve rater education, training, and competence. While there would be increased costs associated with proper rater training, the costs of failed

trials are significantly higher. The quality of ratings in clinical trials needs to be monitored. Uniform standards need to be adopted, and a plan for rater evaluation and certification needs to be implemented. Training methodologies need to undergo rigorous empirical evaluation to determine their effectiveness. New technologies offer a way of not only standardizing but also improving rater training, thus improving rater competence and increasing interrater reliability across diverse settings.

## REFERENCES

1. Robinson D, Rickels K. Concerns about clinical drug trials. *J Clin Psychopharmacol*. 2000;20:593–596.
2. Klein DF, Thase ME, Endicott J, et al. Improving clinical trials: American Society of Clinical Psychopharmacology recommendations. *Arch Gen Psychiatry*. 2002;59:272–278.
3. Hamilton M. Rating depressed patients. *J Clin Psychiatry*. 1980;41:21–24.
4. Hamilton M. General problems of psychiatric rating scales (especially for depression). In: Pichot P, ed. *Modern Problems of Pharmacopsychiatry: Psychological Measurements in Pharmacopsychiatry*. Vol. 7. Basel: Karger; 1974:125–138.
5. Spitzer RL, Williams JBW. Classification in psychiatry: classification of mental disorders and *DSM-III*. In: Kaplan HI, Freedman AM, Sadock BJ, eds. *Comprehensive Psychiatry*. Baltimore: Williams & Wilkins; 1980.
6. Williams JBW. Standardizing the Hamilton Depression Rating Scale: past, present, and future. *Eur Arch Psychiatry Clin Neurosci*. 2001;251(suppl 2):II/6–II/12.
7. Bech P, Engelhardt N, Evans K, et al. A proposal for a standardized HAMD scoring system: a collaboration among the pharmaceutical industry, academia, and government. Paper presented at: 41st Annual Meeting of the National Institute of Mental Health, New Clinical Drug Evaluation Unit; May 2001; Phoenix, AZ.
8. Kalali A, Gibertini M, Kobak KA, et al. A reliability study of the new GRID-HAMD—initial results. Paper presented at: 41st Annual Meeting of the American College of Neuropsychopharmacology; December 2002; San Juan, Puerto Rico.
9. Moberg PJ, Lazarus LW, Meshulam 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(1):35–40.
10. Williams JBW. A structured interview guide for the Hamilton Depression Rating Scale. *Arch Gen Psychiatry*. 1988;45:742–747.
11. Engelhardt N, Kobak KA. Traditional vs. enriched rater training for a multi-site depression trial: a pilot study of effectiveness and feasibility. Paper presented at: 43rd Annual Meeting of the National Institute of Mental Health, New Clinical Drug Evaluation Unit; May 2003; Boca Raton, FL.
12. Lipsitz J, Feiger A, Kobak KA, et al. The Rater Applied Performance Scale (RAPS): development and reliability. Paper presented at: 43rd Annual Meeting of the National Institute of Mental Health, New Clinical Drug Evaluation Unit; May 2003; Boca Raton, FL.
13. DeBrota DJ, Demitrack MA, Landin R, et al. A comparison between interactive voice response system-administered HAMD-D and clinician-administered HAM-D in patients with major depressive episode. Paper presented at: 39th Annual Meeting of the National Institute of Mental Health, New Clinical Drug Evaluation Unit; June 1999; Boca Raton, FL.
14. Feltner D, Kavoussi R, Crockatt J, et al. *Evaluation of an Interactive Voice Response HAMA in a GAD Relapse Prevention Trial*. Paper presented at: 41st Annual Meeting of the National Institute of Mental Health, New Clinical Drug Evaluation Unit; June 2002; Boca Raton, FL.
15. Feiger AD, Engelhardt N, DeBrota D, et al. *Rating the Raters: An Evaluation of Audiotaped Hamilton Depression Rating Scale (HAMD) Interviews*. National Institute of Mental Health, New Clinical Drug Evaluation Unit; May 2003; Boca Raton, FL.
16. Gibbon M, McDonald-Scott P, Endicott J. Mastering the art of research interviewing: a model training procedure for diagnostic evaluation. *Arch Gen Psychiatry*. 1981;38:1259–1262.
17. Tracy K, Adler LA, Rotrosen J, et al. Interrater reliability issues in multicenter trials: part I. Theoretical concepts and operational procedures used in Department of Veterans Affairs Cooperative Study 394. *Psychopharmacol Bull*. 1997;33:53–57.
18. Demitrack MA, Faries D, Herrera JM, et al. The problem of measurement error in multisite clinical trials. *Psychopharmacol Bull*. 1998;34:19–24.
19. Feiger AD, Lipsitz JD, Kobak KA, et al. Impact of a comprehensive HAMD inter-rater reliability training in a multi-site trial. Paper presented at: 41st Annual Meeting of the National Institute of Mental Health, New Clinical Drug Evaluation Unit; May 2001; Phoenix, AZ.
20. Kobak KA, Lipsitz JD, Feiger AD. Development of a standardized training program for the Hamilton Depression Scale using Internet-based technologies: results from a pilot study. *J Psychiatr Res*. 2003;37:509–515.
21. Shayegan DK, Stahl SM. Enhancing inter rater reliability utilizing multimedia and distance learning. Paper presented at: 23rd Annual Meeting of the Collegium Internationale Neuro-Psychopharmacologicum; June 2002; Montreal, CA.

## Editors' Note

### Welcome to New Members of the Editorial Board

We would like to welcome two new members to the Editorial Board of the *Journal of Clinical Psychopharmacology*. They are: Wolfgang W. Fleischhacker, MD, from the Universitätsklinik für Psychiatrie in Innsbruck, Austria and Andrew A. Nierenberg, MD, from the Depression Clinical and Research Program at Massachusetts General Hospital in Boston, Massachusetts.

We thank Drs. Fleischhacker and Nierenberg for their willingness to serve on the Editorial Board.

The Editors-in-Chief