

# A New Approach to Rater Training and Certification in a Multicenter Clinical Trial

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

**Abstract:** Recent evidence demonstrates that the quality of raters' applied clinical skills is directly related to study outcome. As such, the training and evaluation of raters' clinical skill in administering symptom-rating scales is essential before being certified to rate patients in clinical trials. This study examined a novel approach to rater training and certification that focused on both conceptual knowledge and applied skills. Forty-six raters (MDs = 14; PhDs = 7; MA = 5; BA/LPN/RN = 20) in a large multicenter depression study went through a 2-step Hamilton Rating Scale for Depression (HAMD) certification process: didactic training, administered online via an interactive Web tutorial, and live, applied training, where raters interviewed depressed patients while being remotely observed via 3-way teleconference. Raters' applied skills were evaluated using the Rater Applied Performance Scale (RAPS), designed specifically to evaluate critical rater behaviors associated with good clinical interviews. Raters received feedback immediately following the interviews; those receiving a failing score were given 2 more opportunities to pass. Each subsequent session was accompanied by feedback, and was conducted by a different trainer, who was blind to the results of the previous session as well as to which session number it was, to avoid bias. Raters who failed on the third attempt were excluded from rating patients in the trial. All training and testing occurred prior to the startup meeting.

Results found a significant improvement pre-to-post Web training in raters knowledge of scoring conventions,  $P < 0.001$ . On the applied component, raters' RAPS scores improved significantly on the second attempt following feedback, from 9.05 to 11.58,  $P < 0.001$ , and from their second to their third session (from 9.00 to 11.00,  $P = 0.033$ ). Three raters failed all 3 attempts and were excluded from the study. Results support the efficacy of the approach in improving both conceptual knowledge and applied interviewing skill.

(*J Clin Psychopharmacol* 2005;25:407–412)

**R**atings conducted in clinical trials provide the foundation upon which the results of a study rest. Poorly conducted

interviews can reduce signal detection,<sup>1</sup> increasing the risk for failed trials and Type II errors (ie, false-negative results). As a consequence, the selection, training, and evaluation of raters conducting assessments in clinical trials are of critical importance, particularly in clinical trials employing ratings obtained from clinical interviews, such as the Hamilton Rating Scale for Depression<sup>2</sup> (HAMD).

Until recently, the relationship between interview quality and study outcome had not been empirically examined. In one recent study,<sup>1</sup> all baseline HAMD interviews (N = 216) in a pharmaceutical industry-sponsored multicenter depression trial were audio-recorded and evaluated for interview quality. Although overall, the trial failed (ie, the active comparator [paroxetine] failed to separate from placebo), post hoc analyses found that those interviews rated “good” or “excellent” on interview quality showed a large and significant placebo separation (6.8 points,  $P = 0.017$ ), whereas those interviews rated “fair” or “unsatisfactory” failed to separate (–2.8 points,  $P = 0.266$ ) (negative number reflects greater change with placebo than with paroxetine). This finding was confirmed in a subsequent study, in which interviews rated fair or unsatisfactory on one dimension of interview quality, clarification of patient response, failed to separate from placebo, whereas those with good or excellent clarification had a significant separation.<sup>3</sup> Thus, the quality of raters' applied clinical skills appears to be of critical importance in determining study outcome in some trials.

Given the importance of interview quality, the practical problem of assessing this dimension in multicenter clinical trials becomes apparent. Defining behaviors that constitute a good clinical research interview and having a validated tool to assess those behaviors is an essential first step. Recently, a model was presented outlining the skills necessary to administer symptom rating scales in clinical trials.<sup>4</sup> The model includes 2 elements: a thorough conceptual understanding of the scale's scoring conventions and proficiency in a particular set of applied clinical interviewing skills, or rater behaviors. These behaviors include adherence to the interview guidelines, use of appropriate follow-up questions, use of questions to clarify ambiguous information, and neutrality (ie, avoiding leading questions that direct the patient toward specific responses).

It is important that raters participating in clinical trials be evaluated on both their conceptual knowledge of the scale and on their applied clinical skills in administering the scale before being certified to rate patients in clinical trials. The format of rater-training programs should encompass the training

\*MedAvante, Inc., Madison, WI; †MedAvante, Inc., Princeton, NJ; ‡New York State Psychiatric Institute, New York, NY and §Neurosciences Medicine Development Center, GlaxoSmithKline, King of Prussia, PA. Received December 17, 2004; accepted after revision May 17, 2005. Address correspondence and reprint requests to Kenneth A. Kobak, PhD, MedAvante Inc., 7601 Ganser Way, Madison, WI 53719. E-mail: kkobak@Medavante.net.

Copyright © 2005 by Lippincott Williams & Wilkins

ISSN: 0271-0749/05/2505-0407

DOI: 10.1097/01.jcp.0000177666.35016.a0

in and testing of both of these skills. Unfortunately, most rater-training programs conducted today are delivered in whole or in part at the study startup meeting, and involve lectures and passive observation and rating of videotapes. These provide an indirect test of trainees' conceptual knowledge, but tell us nothing about the trainees' applied clinical skills. Recently, a comprehensive rater-training program was developed incorporating both didactic and applied elements.<sup>5</sup> Funded by the NIMH, the program uses an interactive Web tutorial to provide didactic training, and live observation of clinical interviews via videoconferencing or teleconferencing for evaluation of applied clinical interviewing skills. A pilot study found the program effective in improving didactic knowledge and interrater reliability,<sup>5</sup> and a subsequent study found it superior to traditional rater training in improving both conceptual knowledge and applied clinical skills.<sup>6</sup> These studies are the first to our knowledge that evaluated whether learning, in the form of performance on pretest and posttests of didactic knowledge and applied skills, actually occurred as a result of rater training.

The current study describes an application of this rater-training program to qualify raters for participation in a large multicenter clinical depression trial. The program includes both training and testing components, and encompasses both conceptual knowledge and applied clinical skills.

## METHODS

Forty-six raters (21 males, 25 females) from 20 sites were chosen by the sponsor as potential raters for the study. Fourteen (30%) were MDs, 7 (15%) PhDs, 5 (11%) were MAs, and 20 (44%) had a BA or AA (including RNs and LPNs). Raters were required to have prior clinical experience with depressed patients and documented experience administering the HAMD. The certification process consisted of 2 components: didactic testing and applied testing. To be certified to rate patients in the trial, raters had to pass both components.

### Didactic Training and Testing

The didactic component consisted of a Web-based tutorial on the HAMD, the primary outcome measure. The tutorial reviewed the concepts and scoring conventions for each of the items. In addition, there was a module on general interviewing skills. A 20-item multiple-choice test, consisting of clinical vignettes, was administered pretraining and posttraining (Appendix A). A score of 80% correct or greater on the posttest was required to pass. Following the posttest, the tutorial automatically provides feedback to the trainee on the items missed, and a rationale for the correct answers. Those who failed the posttest were instructed to re-review the parts of the tutorial covering the items they missed, and were given one additional chance to take the posttest using an alternative version of the test. Improvement pretraining to posttraining on the didactic examination was measured using paired *t* tests.

### Applied Training and Testing

Raters who passed the didactic examination were allowed to proceed to the applied component. This involved

live observation of the trainee conducting a HAMD interview with a depressed patient. The central training site provided the patient for the trainee to interview. An expert trainer listened to the interview via a three-way teleconference. The trainer scored the HAMD items based on the trainee's questioning of the patient, and rated the trainee's clinical interviewing skills using the Rater Applied Performance Scale<sup>7</sup> (RAPS). The RAPS was developed to assess those clinical skills integral to a good clinical assessment. Such skills include adherence to the interview guidelines, use of sufficient follow-up questions, use of questions to clarify ambiguous information, and neutrality (ie, avoiding leading questions that direct the patient toward specific responses). Each of these skills (as well as rapport and accuracy) is rated along a 4-point scale (1 = unsatisfactory, 2 = fair, 3 = good, or 4 = excellent). Validation studies found the RAPS to possess good psychometric properties.<sup>7,8</sup>

Following the interview, the trainer gave the trainee feedback on their interviewing technique, discussed the ratings for the individual HAMD items, and gave suggestions for improvement. Trainees who received a failing score on the interview were required to conduct another interview, incorporating the feedback given. Failure was defined a priori as either (a) an unsatisfactory rating on any of the 6 RAPS dimensions; (b) a score of 2.5 or less on the average of the first 4 RAPS items; (c) 3 or more individual HAMD items diverge from the trainers score by 2 points or more; (d) 6 or more individual items are off by 1 point from the trainers score; or (e) a sum of 6 total points discrepancy.

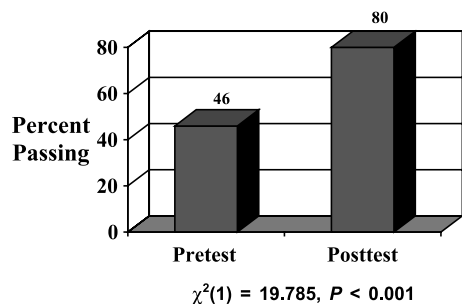
Trainees were given a total of 3 chances to pass the applied component. If the trainee failed on the third interview, they were excluded from rating patients in the trial. Each of the 3 applied tests was conducted by a different trainer, who was blind to whether it was the trainee's first, second, or third attempt. This was done to minimize bias, that is, the tendency to pass a trainee on the third attempt if their performance was marginal.

Trainees who passed both the didactic and applied testing components were certified to rate patients in the trial. Thus, all evaluation, training, and certification were accomplished prior to the startup meeting. During the study startup meeting, however, an applied "refresher" session was conducted to reinforce behaviors and help trainees maintain their skills. This consisted of small group breakout sessions in which raters took turns interviewing a standardized depressed patient (ie, a trained actor). Following each item, raters discussed the rationale for their ratings, and the trainer provided additional feedback on interview technique.

## RESULTS

### Didactic Training

A significant improvement was found pretraining to posttraining in the trainees' knowledge of scoring conventions, with the mean number of correct answers on the didactic examination increasing from 14.34 at pretest to 17.54 at posttest [ $t(45) = 7.11, P < 0.0001$  (pretest range 6–20, SD = 3.10; posttest range 15–20, SD = 1.62)]. Prior to the



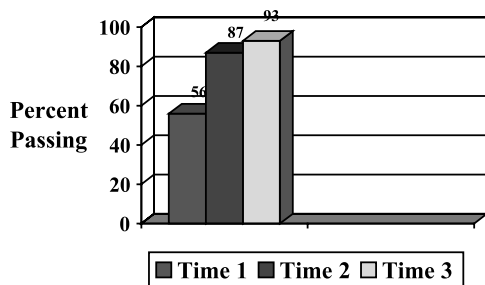
**FIGURE 1.** Pre-to-post training improvement on didactic knowledge of HAMD scoring conventions: percent passing.

didactic training, 46% passed the pretest (pass = 80% correct). After completing the Web tutorial, 89% of trainees passed the posttest on their initial attempt,  $\chi^2(1) = 19.785, P < 0.001$ , and the remainder passed on their second attempt. (Fig. 1). The most difficult items for trainees were those regarding the evaluation of nonverbal affect in depressed mood (38% wrong), evaluating weight loss in a patient who has begun to regain weight (38%), distinguishing concomitant medical conditions from somatic symptoms of anxiety (33%), and distinguishing loss of appetite from quantity of food eaten (33%).

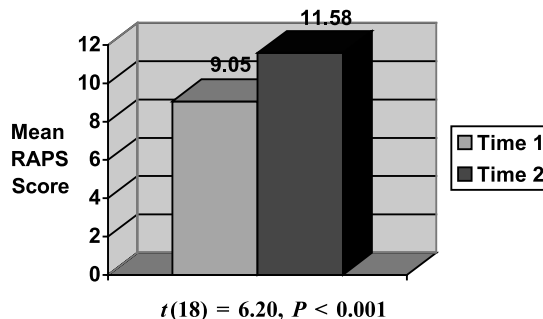
**Applied Training**

Of the 46 trainees, 26 (56%) passed the applied training examination on their initial attempt (prior to any feedback) (Fig. 2). Of the 20 trainees who required additional training and a second examination, 14 (70%) passed on their second attempt. Of the 6 who required a third attempt, 3 (50%) passed. The remaining 3 raters were excluded from participating in the study.

The mean RAPS score for all raters on their initial attempt was 11.06 (SD = 2.24) (maximum possible score = 16). The mean RAPS score for those failing the initial attempt was significantly lower than those who passed (9.05 vs. 12.54) [ $t(43) = 8.10, P < 0.0001$ ]. For those failing their initial applied examination, their RAPS score improved significantly on their second attempt following feedback (from 9.05 to 11.58) [ $t(18) = 6.20, P < 0.0001$ ; Fig. 3]. Similarly, those failing their second applied examination improved significantly on their third attempt following their



**FIGURE 2.** Percent of total sample (N = 46) passing applied clinical skills examination on RAPS prior to feedback (Time 1), and after 1 and 2 feedback sessions (Times 2 and 3).



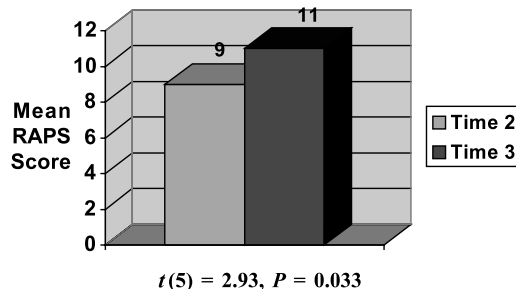
**FIGURE 3.** Improvement from Time 1 to Time 2 on RAPS (applied clinical skills) for those failing the initial clinical skills examination.

second feedback session (from 9.00 to 11.00) [ $t(5) = 2.92, P = 0.033$ ; Fig. 4]. Interestingly, the mean RAPS score did not correlate significantly with the rater's score on the Web exam,  $r = 0.108, P = 0.479$ , suggesting that these are 2 different sets of skills.

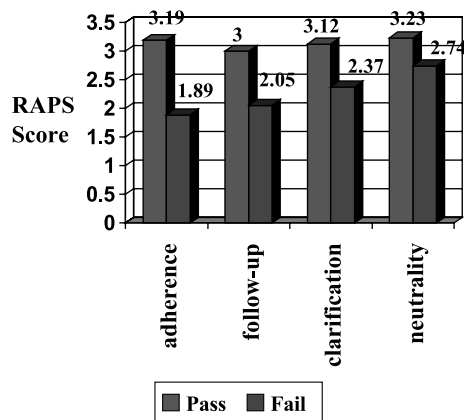
For those failing the initial applied skills examination, we were interested in learning which clinical skills were the most problematic (the best performance on each item would merit a score of 4 [excellent], the worse performance a score of 1 [unsatisfactory]). They were, in increasing order adherence (mean RAPS = 1.89), follow-up (2.05), clarification (2.37), and neutrality (2.74). Scores on all these dimensions were significantly lower for those who failed versus those who passed ( $P < 0.01$  for all comparisons) (Fig. 5).

Finally, we examined the results stratified by education to see whether the training generalized across educational levels. On the didactic training, a significantly greater percentage of MDs or PhDs passed the pretest (61.9%) compared with those with MA, BA, or AA degrees (32.0%) [ $\chi^2(1) = 4.114, P = 0.043$ ]. However, both groups significantly improved with training (mean improvement of 3.14 and 3.24 points, respectively) [ $t(20) = 4.932, P < 0.001$  and  $t(24) = 5.054, P < 0.001$ ].

On the applied training, 71.4% of MD/PhDs passed the RAPS on their initial testing, compared with 44.0% of MA/BA/AA trainees [ $\chi^2(1) = 3.494, P = 0.062$ ]. However, among those who failed the initial testing, both groups had substantial improvement on their second RAPS testing



**FIGURE 4.** Improvement from Time 2 to Time 3 on RAPS (applied clinical skills) for those failing the second clinical skills examination.



**FIGURE 5.** Mean RAPS score on the individual RAPS items for those passing versus failing the initial applied clinical skills examination.

following the first training session [mean improvement of 3.20 and 2.29 points respectively,  $t(4) = 2.764$ ,  $P < 0.051$  and  $t(13) = 5.950$ ,  $P < 0.001$ ].

## DISCUSSION

Screening and qualifying raters who administer the primary outcome measures in clinical trials is an essential component of clinical trial methodology. Results of this study support the efficacy of both the Web tutorial in improving conceptual knowledge, and the applied training in improving clinical interview skills. The methodology appeared to work equally well with trainees both with and without advanced degrees. Administering the didactic component on the Web (versus a lecture at a startup meeting) allows raters to work at their own pace, ensures standardization of the material imparted, and improves the quality of the educational experience, by utilizing interactive learning methodologies. Our results also demonstrate that applied clinical skills can be evaluated and improved. Although somewhat labor intensive, the importance of interview quality in study outcomes makes the time and effort spent worthwhile (especially considering the time and costs of failed trials). That said, improved rater performance may not have any effect on any given trial's results. Even excellent raters will not make an ineffective compound effective, and there are many reasons besides rater quality that cause failed trials.

The entire training protocol (didactic and applied) was conducted in approximately 12 weeks using 4 trainers. Each clinical interview lasted approximately 1 hour, with feedback. The percentage of raters who failed on the first attempt (44%) is a sobering statistic. However, in this study, we were able to bring most raters up to satisfactory standards in 3 sessions. This mirrors data reported by Muller et al,<sup>9</sup> who found that 3 training sessions were optimal for training on the Positive and Negative Symptom Scale (PANSS). It should be noted, however, that a certain degree of prior clinical expertise and conceptual knowledge is necessary to achieve these goals within 3 sessions (see Kobak et al for a

review<sup>4</sup>). It is unlikely that a novice rater with no prior clinical experience with depressed patients and no understanding of clinical research interviewing will be able to achieve similar results. In fact, a retrospective analysis of the 3 raters who failed all 3 applied tests revealed such a lack of experience. Clearly, more data are needed on the quality of clinical trial assessments, but the limited data available are similar to our findings.<sup>3,6</sup>

Another interesting finding was the low correlation between scores on the applied and didactic training, suggesting that these are somewhat independent skill sets. In other words, one could have a good intellectual understanding of how to administer a scale, and a thorough knowledge of the scales conventions and not be able to apply this knowledge in conducting a competent clinical interview. Thus, conceptual knowledge may be a necessary, but not sufficient, condition for a good clinical assessment.

An important consideration is whether the improvement in interview quality is maintained throughout the trial. For this reason, we incorporated a midpoint evaluation session, where all raters are tested on their clinical skills halfway through the trial using the same methodology. Whether the good clinical skills are actually applied during the conduct of the study cannot be determined without some form of monitoring of the actual interviews conducted with study patients. The use of audiotaping has become technologically much easier with the advent of digital recorders, and audio files can be sent electronically to remote experts using secure file transfer protocols. The American Society of Clinical Psychopharmacology has recommended the use of audiotape monitoring for ongoing quality control.<sup>10</sup>

In summary, results support the efficacy of the program in improving rater's conceptual knowledge and clinical skills. As technology improves, remote training may become more widely available, using videoconferencing as well as teleconferencing. Training and screening of raters' applied clinical skills should be incorporated into clinical trial methodology, and the efficacy of such procedures should be empirically evaluated. The impact of interview quality on signal detection should be monitored in future studies.

## ACKNOWLEDGMENTS

*This study was funded by a grant from GlaxoSmithKline. This project has also been funded in whole or in part with Federal funds from the National Institute of Mental Health, National Institutes of Health, Department of Health and Human Services, under contract no. NIH-N43MH12049, and from a grant from Eli Lilly & Co.*

## REFERENCES

1. Kobak KA, Feiger AD, Lipsitz JD. Impact of interview quality on signal detection. *Am J Psychiatry*. 2005;162:628.
2. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry*. 1960;23:56-62.
3. Feiger A, Engelhardt N, DeBrota D, et al. Rating the raters: an evaluation of audiotaped Hamilton Depression Rating Scale (HAM-D) interviews. National Institute of Mental Health, New Clinical Drug Evaluation Unit, 43rd Annual Meeting. Boca Raton, FL; 2003.

4. Kobak KA, Engelhardt N, Williams JBW, et al. Rater training in multicenter clinical trials: issues and recommendations. *J Clin Psychopharmacol*. 2004;24:113–117.
5. Kobak KA, Lipsitz JD, Feiger A. 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(6):509–515.
6. 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*. In press.
7. Lipsitz J, Kobak KA, Feiger A, et al. The Rater Applied Performance Scale (RAPS): development and reliability. *Psychiatry Res*. 2004;127:147–155.
8. Lipsitz J, Feiger A, Kobak K, et al. A scale for rating the applied performance of clinical raters in multi-site trials: development and reliability. National Institute of Mental Health, New Clinical Drug Evaluation Unit, 43rd Annual Meeting, Boca Raton, FL. Boca Raton, FL; 2003.
9. Muller MJ, Rossbach W, Dannigkeit P, et al. Evaluation of standardized rater training for the Positive and Negative Syndrome Scale (PANSS). *Schizophr Res*. 1998;32(3):151–160.
10. Klein DF, Thase ME, Endicott J, et al. Improving clinical trials American Society of Clinical Psychopharmacology Recommendations. *Arch Gen Psychiatry*. 2002;59:272–278.

**APPENDIX A: SAMPLE QUESTIONS FROM THE 20-ITEM DIDACTIC EXAM**

1. A 45-year-old woman reports little interest in sex the past week. She states that she has never had a strong interest in sex and that this is not any change from usual for her. Assuming that she has only been depressed for the past year, what would the rating be for sexual interest?  
0—absent  
1—mild  
2—severe
2. Josephine really noticed a change in her mood this past week. Her friends have remarked on the difference. She is laughing more and is more like her old self. She still has very little appetite but now feels this can't be related to her depression because her depressed mood has lifted.
  - A. Loss of appetite is no longer rated.
  - B. Loss of appetite is rated at a lower level of severity to account for doubts about whether it is related to depression.
  - C. Loss of appetite is rated only if other symptoms such as insomnia are still present.
  - D. Loss of appetite is still rated as observed, regardless of change in other symptoms.
3. Janine came for treatment of her depression after gaining 10 lbs in the 2 months since her depression began. Now after 6 weeks of treatment, she is feeling much better, has lost 5 or 6 lbs and is nearing her predepression weight level.
  - A. A 2 is rated for weight loss.
  - B. A 1 is rated for weight loss.
  - C. A 0 is rated for weight loss.
  - D. More information is needed to make a rating on this item.
4. Before her depression began, Marian would usually sleep from 11:30 PM to 7 AM. Although she has no trouble falling asleep, she has been waking up at about 2 AM and stays awake for the rest of the night, tossing and turning until she gets up to get ready to go to work at 7 AM.
  - A. This is considered both middle and late insomnia and scored on both items.
  - B. This is considered middle insomnia because she wakes in the middle of the night.
  - C. This is considered late insomnia because she stays awake until morning.
  - D. Neither middle nor late insomnia is scored because the symptoms do not quite fit.
5. Your patient appears very depressed. She has a sad face, tends to slump in her chair, and becomes tearful 3 times during the interview. However, she describes her mood on the 6 other days as only mildly depressed, and she was not been tearful or pessimistic. What rating would you give on depressed mood?  
0—absent  
1—indicated only on questioning (occasional, mild depression)  
2—spontaneously reported verbally (persistent, mild to moderate depression)  
3—communicated nonverbally (ie, facial expression, posture, voice, tendency to weep; persistent, moderate to severe depression)  
4—virtually only depressed mood; this in spontaneous verbal and nonverbal communication (persistent, very severe depression, with extreme hopelessness or tearfulness)

Copyright © Kenneth A. Kobak & Joshua D. Lipsitz. All rights reserved. May not be reproduced in whole or in part in any form or by any means without written permission of the authors.