Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Psychiatric Research 40 (2006) 192–199
JOURNAL OF PSYCHIATRIC RESEARCH
www.elsevier.com/locate/jpsychires

Enriched rater training using Internet based technologies: A comparison to traditional rater training in a multi-site depression trial ☆

Kenneth A. Kobak *, Nina Engelhardt, Joshua D. Lipsitz [1]

MedAvante Inc., 7601 Ganser Way, Madison, WI 53717, United States

Received 15 March 2005; received in revised form 20 July 2005; accepted 30 July 2005

Abstract

Objective: The evaluation and training of raters who conduct efficacy evaluations in clinical trials is an important methodological variable that is often overlooked. Few rater training programs focus on teaching and assessing applied clinical skills, and even fewer have been empirically examined for efficacy. The goal of this study was to develop a comprehensive, standardized, interactive rater training program using new technologies, and to compare the relative effectiveness of this approach to "traditional" rater training in a multi-center clinical trial.

Method: 12 sites from a 22 site multi-center study were randomly selected to participate (6 = traditional, 6 = enriched). Traditional training consisted of an overview of scoring conventions, watching and scoring videotapes with discussion, and observation of interviews in small groups with feedback. Enriched training consisted of an interactive web tutorial, and live, remote observation of trainees conducting interviews with real or standardized patients, via video- or teleconference. Outcome measures included a didactic exam on conceptual knowledge and blinded ratings of trainee's audiotaped interviews.

Results: A significant difference was found between enriched and traditional training on pre-to-post training improvement on didactic knowledge, $t(27) = 4.2$, $p < 0.0001$. Enriched trainees clinical skills also improved significantly more than traditional trainees, $t(56) = 2.1$, $p = 0.035$. All trainees found the applied training helpful, and wanted similar web tutorials with other scales.

Conclusions: Results support the efficacy of enriched rater training in improving both conceptual knowledge and applied skills. Remote technologies enhance training efforts, and make training accessible and cost-effective. Future rater training efforts should be subject to empirical evaluation, and include training on applied skills.

Keywords: Rater training; Assessment; Inter-rater reliability; Depression; Hamilton depression scale; Computerized assessment; Internet; Clinical trials

1. Introduction

The evaluation and training of raters who conduct efficacy evaluations in pharmaceutical-sponsored drug trials is a feature of study design that has been largely overlooked. In spite of increasing recognition of the importance of establishing inter-rater reliability and improving interview quality in multi-center trials, little empirical research has been conducted to evaluate the effectiveness of training programs in achieving these

☆ Parts of this paper have been previously presented at the 43rd Annual Meeting, National Institute of Mental Health, New Clinical Drug Evaluation Unit (NCDEU), May 27–30th, 2003, Boca Raton, FL, USA.

* Corresponding author. Tel.: +1 608 239 3919; fax: +1 608 829 1965.
E-mail address: kkobak@medavante.net (K.A. Kobak).
[1] Columbia University and Research Training Associates.

0022-3956/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jpsychires.2005.07.012

goals. This lack of empirical data is in part a result of the fact that few such training programs exist. As the success or failure of a clinical trial rests largely in the hands of the raters administering the outcome measures, attention to the quality of the clinical assessments performed is critical.

Rater competency can be thought of along two domains: conceptual knowledge and applied skills. Conceptual knowledge refers to the academic understanding of scoring conventions, concepts evaluated by the scale, and general rules for scale administration. Applied skills refer to how well the rater can apply this knowledge in conducting a good clinical interview. A comprehensive rater training program should teach and evaluate both of these domains (Kobak et al., 2004). Unfortunately, most raters learn how to conduct these interviews at startup meetings (75% according to a recent study) (Kobak and Engelhardt, 2003), a venue which is inadequate in terms of providing the time and substance required to achieve these goals.

Evaluation and training of raters' applied skills is especially critical. In a recent study (Kobak et al., 2005), interview quality was examined as a factor in study outcome. All baseline Hamilton depression scale (HAMD) (Hamilton, 1960) interviews ($N = 216$) in a multi-center depression trial were recorded and evaluated for interview quality using the rater applied performance scale (RAPS) (Lipsitz et al., 2004). Overall, the study was a failed trial (i.e., the active comparator (paroxetine) failed to separate from placebo). However, post hoc analyses found that those interviews rated "good" or "excellent" showed a large and significant placebo separation (6.8 points, $p = 0.017$) while those interviews rated "fair' or "poor" on interview quality failed to separate (−2.8 points, $p = 0.266$) (negative number reflects greater change with placebo than with drug). Thus, the quality of raters' applied clinical skills appears to be of critical importance. Unfortunately, most training at start-up meeting involve passive observation and rating of videotapes, which provide an indirect test of trainees' conceptual knowledge, but tell us nothing about the trainees' applied clinical skills. When this method is used to evaluate inter-rater reliability, it artificially inflates reliability estimates, as it reduces the "information variance" that would result if each rater evaluated the patient independently (Spitzer and Williams, 1980).

Given the importance of applied clinical skills, the question becomes "what is the quality of interviews currently being conducted in clinical trials?" The little evidence that is available is not encouraging. In one study involving 221 baseline HAMD ratings, 59% of the interviews were rated fair or poor on adherence to the interview guide instructions, 63% fair or poor on clarification skills, 55% fair or poor on follow-up questioning, and 63% fair or poor on neutrality (Kobak et al., 2005). In a second study (Feiger et al., 2003), the figures were 68% fair or poor on adherence, 61% on clarification, and 77% on follow-up. In the latter study, 45% of the interviews were under 10 min (range 3–35 min) in spite of Hamilton's suggestion that the interview should take at least a half-hour (Hamilton, 1967). The effect of poorly conducted interviews on the growing rate of failed trials (Khan et al., 2002) is unknown, but is likely to contribute to this phenomenon.

In an attempt to remedy these problems and address the shortcomings in current rater training efforts, a comprehensive, standardized, interactive rater training program was recently developed, using new technologies. The model utilized the Internet for didactic instruction, and videoconferencing and teleconferencing for applied training in clinical skills. A small, open, pilot study of this enhanced training methodology found significant improvements in didactic knowledge and attainment of good reliability between raters (Kobak et al., 2003).

The objective of the current study was to compare the relative effectiveness of a commonly used, or "traditional", approach to rater training to an enhanced rater training approach using new technologies within the context of a multi-center clinical trial. The goal was to evaluate the relative effectiveness of each approach in improving both conceptual knowledge and applied clinical skills using pre-defined performance measures before and after training was conducted. We hypothesize that enhanced training will show greater improvement pre-to-post training than traditional training on (1) didactic knowledge (as measured by pre-to post test scores) (see below) and (2) applied clinical skills (as measured by the RAPS scale (Lipsitz et al., 2004)) (see below).

2. Method

2.1. Subjects

Subjects were raters participating in a 25-center Phase III depression trial in the United States. All research sites were invited by the sponsor to participate in the study. A total of 22 sites agreed to participate. Of the three sites that did not participate, 1 site declined participation and 2 sites did not have a high-speed Internet connection required to participate in the study. Of the 22 sites willing to participate, 12 were randomly selected (using computer-generated randomization schedule) to participate in the study. Six sites (14 raters) were randomly assigned to traditional rater training and six sites (16 raters) to enriched rater training. All enriched training and post-training efficacy assessments were completed within one month prior to the start up meeting. Thus, the intervention and outcome measures were uncontaminated by exposure to the startup meeting (enriched training sites subsequently attended the startup

meeting training, per sponsors request). Sites participating in the study were compensated for their time. Six sites were non-academic research centers, two were academic sites, and two were non-academic health care centers. All raters signed informed consent documents approved by the Western Institutional Review Board (WIRB).

2.2. Rater training methods

2.2.1. Traditional rater training

"Traditional" training refers to methods commonly used by pharmaceutical companies in multi-center clinical trials to train raters on the administration and interpretation of the HAMD. In this study, traditional rater training was designed and delivered at the investigator meeting by the sponsor. It consisted of the following components:

1. Raters watched and independently scored a videotaped HAMD clinical interview.
2. The scores were tabulated and summary statistics presented, including distributions of the raw total and item scores and percent agreement and intraclass correlations of the total score and individual item scores. Discrepancies were discussed and the rationale for correct responses was provided.
3. A general overview of scoring conventions was provided.
4. A second HAMD videotape was shown and independently scored by the raters.
5. The sponsor evaluated the clinical interview skill of raters and provided feedback in groups comprised of 8–10 raters. Raters in each group were observed administering 3 items of the HAMD to another rater who posed as a depressed patient.

2.2.2. Enriched rater training intervention

"Enriched" training refers to a web-based, remote training program that consists of a didactic and an applied component.

Didactic component. The didactic component preceded the applied component in order to insure that trainees had an adequate conceptual understanding of the principles for administering and scoring the HAMD before attempting to apply these principles in conducting a clinical interview. The didactic component consisted of an interactive, web-based tutorial containing the following features:

(1) A review of the general guidelines for administering the HAMD. This included an interactive "self test", for immediate reinforcement of learning.
(2) A review of the concepts and scoring conventions for each of the 17 HAMD items. This included interactive video vignettes illustrating the concepts and scoring conventions, followed by interactive self-testing of the trainees understanding, in order to reinforce learning. Trainees were given immediate feedback on their answers and a rationale for the correct score.

Trainees could e-mail questions to the instructors at any time during the tutorial. Trainees could also print the text of the teaching modules for future reference. The entire didactic component took about 2 h to complete.

Applied component. The applied training component involved having trainees remotely conduct two HAMD clinical interviews: one with an actual patient and one with a standardized patient (i.e., an actor). The patients were provided by the central training site and were remotely interviewed by the trainee by either videoconference or teleconference (see below). The trainer was in the room with the patient, and observed the interview, providing live feedback to the trainee both during and after the interview. In addition, each trainee participated in one "group" training session via teleconference, during which trainees took turns interviewing a patient, followed by group discussion of ratings after each item. This group process was designed to reinforce learning by having trainees learn from observing each other, and provide cross calibration by including raters in each group from different sites.

2.3. Pre- and post-training outcome measures

The following measures were administered pre- and post-training:

(1) A 20-item multiple-choice test on didactic knowledge of scoring conventions.
(2) Evaluation of applied interviewing skills. Each rater completed two HAMD interviews before training and two interviews after training to measure change in clinical skill following the training intervention. Interviews were recorded and reviewed by one of two outside expert raters, who were blind to both which training intervention the rater received and whether the interviews were pre- or post-training. Interviews were rated for clinical quality using the RAPS (Lipsitz et al., 2004) scale. The RAPS scale was developed specifically for evaluating expertise in conducting clinician-administered symptom rating scales, and evaluates interviewer's skills on a four-point scale along six dimensions: *adherence* (to interview protocol), *clarification* (skill in clarifying ambiguous information), *follow-up* (use of additional probes to elicit further information), *rapport* (appropriate connection with the patient without becoming therapeutic), *neutrality* (avoiding leading questions

and minimizing expectancy effects), and *accuracy* (to gold standard clinician/trainer). In order to standardize patient difficulty between groups, each patient was interviewed at least twice (once by an enriched trainee and once for a traditional trainee, in counterbalanced order). All patients for pre- and post-testing purposes were provided by the central training site, and interviewed remotely by video- or teleconference as previously described.

(3) Each trainee completed a feedback form evaluating the training methodology. In addition, patients used for training purpose completed a feedback form on how they felt being interviewed remotely.

The study utilized the SIGH-D structured interview guide (Williams, 1988) , augmented with additional probes to more precisely determine frequency and intensity of a symptom. The anchor points (originally developed for the HAMD by Guy (Guy, 1976)) were also augmented to increase clarity and improve reliability. The use of a structured interview guide increases reliability by providing standardized probes, which reduce information variance, and helps insure all required domains are assessed (Moberg et al., 2001).

All training and testing was done using half actual patients and half "standardized" patients, i.e., medically trained actors from the University of Wisconsin Medical School. Actual patients have the advantage of authenticity, while standardized patients allow the trainer to carefully design scripts to teach specific points or test specific skills. Studies on standardized patients have found they achieve high level of stability for inter-rater reliability purposes in the assessment of depression (Badger et al., 1995). In addition, several studies have demonstrated that trainee competence as evaluated with standardized patients is a good measure of trainee competence with actual patients (Colliver and Swartz, 1997; De Champlain et al., 1997; Peabody et al., 2000; Pieters et al., 1994). Studies have found experienced physicians were unable to differentiate standardized patients from real patients when sent unannounced into a physicians office, even when the physician was told in advance that this would be occurring (Colliver and Swartz, 1997). Order of actor vs patient was counterbalanced for both training and testing. All real patients used for training purposes were recruited via newspaper advertisements and signed consent forms approved by the Western IRB. Depression diagnoses of real patients were confirmed using a structured clinical interview (Sheehan et al., 1998).

Half the sites utilized videoconferencing for training and testing applied skills, and the other half utilized teleconferencing. This was done to allow for an empirical comparison between the two training methodologies. We felt this was important because the use of videoconferencing requires high-speed Internet access, which many sites (especially outside the US) do not currently have. Thus, it is important to know if teleconferencing can obtain equivalent results.

With videoconferencing, trainees were provided a Logitech webcam (30 frames per second, 640×480 pixels), which is plugged into their USB port. After loading the camera and videoconferencing software ("Click to Meet Express"), they were able to connect to the videoconference session by going to the Click to Meet website and entering a user ID and password. A dedicated videoconferencing server was used, that could host up to 10 people. All passwords were stored encrypted using the DES encryption algorithm. All servers were protected in physically secure locations with monitored access and at least one form of biometric access control.

3. Results

Rater and site demographics. Rater demographics are presented in Table 1. As a whole, the raters were older (mean age = 45 years), experienced (mean = 9 years in psychiatric clinical trials research) and had participated in an average of 24 depression trials. Seventy-three percent of the raters had conducted over 100 HAMD interviews. Forty-one percent were principal investigators. Most had learned to conduct the HAMD at start up meetings. Only 38% reported having ever been observed actually conducting a HAMD as part of their HAMD training (Table 2). Seventeen percent of the sites were academic sites, 14% non-academic health care sites, and 69% non-academic research centers.

Results: didactic knowledge. A significant difference was found between enriched and traditional training interventions on pre-to-post training improvement on didactic knowledge (mean change = 4.4 points for enriched, 0.5 points for traditional, $t(27) = 4.2$, $p < 0.0001$). The mean number of correct answers on the didactic exam increased from 14.07 to 18.47 in the enriched group, $t(14) = 6.60$, $p < 0.0001$, and from 12.07 to 12.57 in the traditional group ($t(13) = 0.81$, $p = 0.433$) (Fig. 1).

Results: applied skills. The mean RAPS score improved 2.47 points with enriched training, compared to 0.14 points with traditional training, $t(56) = 2.1$, $p = 0.035$ (Fig. 2). In terms of individual RAPS dimensions, the percentage of trainees rated as "good" or "excellent" at post test on the dimensions of adherence and follow-up were significantly greater in the enriched group than in the traditional group (72.2% vs 27.8% for adherence and 69% vs 31% for follow-up, $\chi^2(1) = 15.98$, $p < 0.0001$ and $\chi^2 = 6.90$, $p < 0.009$, respectively) (Table 3). The mean improvement was also significantly greater in the enriched group on these two subscales (0.73 vs −0.03 and 0.80 vs −0.36, respectively, $t(56) = 2.1$, $p = 0.037$ and $t(56) = 3.1$, $p = 0.003$, respectively) (Fig. 3).

Table 1
Rater demographics

	Mean age (SD) (range)	Gender	Education	Mean years research experience (SD) (range)	Mean # depression trials (SD) (range)	# HAMDs	# Licensed clinicians	Role in study
Enriched ($N = 16$)	45.00 (10.26) (30–62)	6 M 10 F	MD – 37% PhD – 19% MS – 19% BS – 25% <BS – 0	9.77 (9.12) (1–30)	25.43 (33.40) (3–100)	0–100: 3 101–250: 4 >250: 9	11 (68.8%)	PI = 6 (37.5%) Other = 10 (62.5%)
Traditional ($N = 14$)	45.36 (8.16) (35–60)	9 M 5 F	MD – 57% PhD – 14% MS – 07% BS – 14% <BS – 7%	8.23 (5.70) (2–23)	23.50 (26.86) (4–100)	0–100: 5 101–250: 2 >250: 7	12 (85.71%)	PI = 6 (42.9%) Other = 8 (57.1%)
Combined ($N = 30$)	45.17 (9.14) (30–62)	15 M (50%) 15 F (50%)	MD – 45% PhD – 17% MS – 15% BS – 17% <BS – 3% UNK – 3%	9.0 (7.63) (1–30)	24.46 (29.76) (3–100)	0–100: 28% 101–250: 17% >250: 55%	23 (76.67%)	PI = 12 (40%) Other = 18 (60%)

All comparisons ns.

Table 2
How trainees first learned to conduct the HAMD ($N = 29$)

Training activity	%
Watch videos	72
Observe others	59
Reading	55
Supervisor observed me conducting HAMD interviews	38
Role playing	24
Group startup meeting	72

Note: Categories not mutually exclusive.

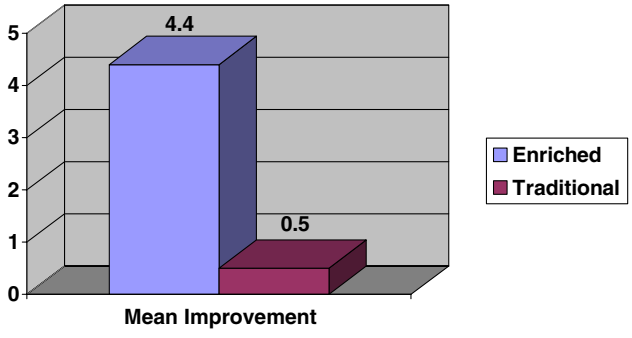

Fig. 1. Pre-to-post training improvement in scores on exam of didactic knowledge: enriched vs. traditional training ($t(27) = 4.2$, $p < 0.0001$).

Interview length. The mean length of HAMD interviews increased significantly pre-to-post training in the enriched group, from 21.04 to 27.49 min, $t(50) = 3.61$, $p = 0.007$. The mean length of HAMD interviews in the traditional group did not change significantly, decreasing from 22.0 to 21.89 min, $t(54) = 0.59$, $p = 0.558$.

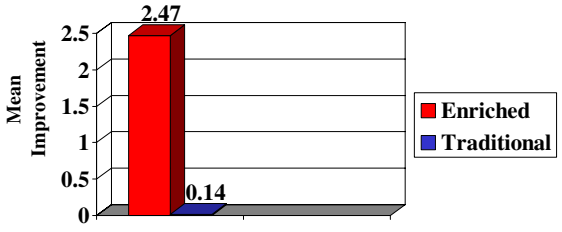

Fig. 2. Mean improvement (pre-and post training) on applied skills (RAPS scale): enriched vs. traditional training ($t(56) = 2.1$, $p < 0.035$).

Table 3
Percent rated "good" or "excellent" on RAPS dimensions post-training

	Enriched (%)	Traditional (%)	χ^2	p
Adherence	72.2	27.8	15.98	0.0001
Follow-up	69	31	6.90	0.009
Clarification	60	40	1.21	0.272
Neutrality	51.2	48.8	0.02	0.885
Rapport	52.1	47.9	0.01	0.905

Trainee satisfaction. Trainees were asked to evaluate both the didactic web tutorial and the applied training program. For the web tutorial, 67% reported they found

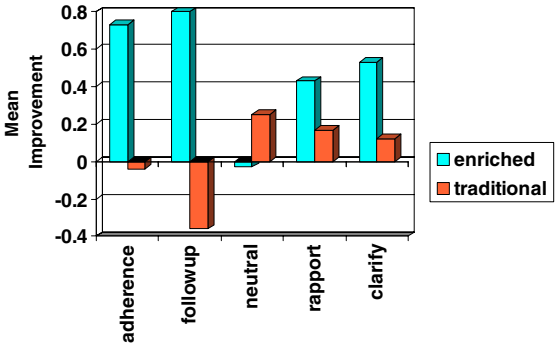

Fig. 3. Mean improvement on RAPS subscales, pre-to-post training ($p < 0.05$ for adherence and follow up).

it convenient, 88% found it helpful, and all but one (93%) found it more useful than training typically conducted at startup meetings. All trainees (100%) wanted to see similar web training on other rating scales.

For the applied training, 87% reported they found it convenient, 93% found the approach helpful, and 87% found it more useful that usual training. All trainees (100%) found the live feedback helpful, and 87% would like to receive similar applied training on other scales. Sixty-seven percent thought the length of the entire training program was just right, and 33% too long. Twenty-six percent thought the technology interfered with their ability to communicate with the patient.

A significantly greater percentage of videoconference trainees reported the technology interfered with their communication (66%) compared to teleconference trainees (0%), χ^2 (1) = 8.182, $p = 0.004$. No other significant differences were found between videoconference and teleconference. The percentage of trainees who reported liking the training methodology was not significantly different between videoconference trainees (83%) and teleconference trainees (63%).

Patient satisfaction. Patients were also asked to evaluate what the experience was like for them to be interviewed remotely. A total of 20 real (not standardized) patients were solicited for feedback. In general, patients felt comfortable being interviewed (81%), thought the interviewer was able to evaluate them well using the technology (93%), were willing to be interviewed remotely to avoid having to travel to the office (89%), and thought is was a useful way to receive a psychological evaluation when other means were limited or unavailable (96%). Almost all (98%) said they would like to be interviewed again using the technology, but 22% said they thought the technology interfered with their ability to communicate with the interviewer. No significant difference in level of satisfaction was found between phone and with videoconference interviews.

Videoconference vs teleconference. No significant difference was found in improvement in applied clinical skills between those who received applied training via videoconference and those trained by teleconference (mean RAPS improvement of 2.61 and 2.25, respectively, $t(28) = 0.278$, $p = 0.783$).

Standardized vs real patients. The Intraclass correlation (ICC) between trainees tested using actors was 0.9613 (95% C.I. 0.9210–0.9813), compared to an ICC of 0.8695 (95% C.I. 0.7461–0.9354) for trainees tested using real patients, not a significant difference. Similarly, there was no difference between actors and real patients in RAPS scores either at pre-test (12.3 vs 11.6) or post-test (13.8 vs 12.6), $t(58) = 0.98$, $p = 0.329$ and $t(56) = 1.56$, $p = 0.125$, respectively. Overall, trainees accurately guessed whether a patient was real or standardized 66% of the time. However, a significant difference was found between interview modes in this regard. When interviewed by videoconference, trainee's guesses were accurate only 53% of the time (about chance), but when interviewed by telephone, they were accurate 80% of the time, $\chi^2(1) = 9.50$, $p = 0.002$.

4. Discussion

The results of this study provide empirical support for the efficacy of enriched rater training in improving both conceptual knowledge and applied clinical skills. Enriched rater training produced significant improvements pre-to-post training, while traditional rater training failed to produce significant changes in either conceptual knowledge or clinical skills. Part of the reason for the efficacy of the enriched rater training program may simply be a function of time: by providing the training outside of the start-up meeting, trainees were not constrained by time limitation, and thus could study the didactic component at their own pace, and have enough time for individual tutoring on applied clinical practice. Theoretically, similar enhancements could occur without the Internet technology if enough time was devoted at the startup meeting. However, we believe that the use of technology improved the quality of the training effort. Use of the Internet enabled the didactic component to be interactive and multi-modal, two components associated with increase retention of knowledge (Mott, 2000). The use of videoconferencing and teleconferencing enabled remote, personalized, instruction and testing, allowing training to be delivered to diverse sites from a central location. This also helped insure the standardization and quality of the material, i.e., that all trainees received the same information.

The trainees in this study were, on average, a rather experienced group. It was encouraging that the enhanced training intervention had an impact even with experienced clinicians, who may be 'set in their ways'. It is likely, however, that the improvement was in part due to the finding that the majority of raters never received formal instruction on how to administer the HAMD, or on the unique rules for administering

symptom-rating scales in a research context. These include such concepts as avoiding the "halo effect" (i.e., the tendency to rate an item high because another item was rated high), and avoiding "response set", e.g., the tendency to rate toward central tendency or the opposite, to rate only at the extremes (Hamilton, 1974). Few formal academic programs include training on the use of the clinician-administered symptom rating scales that are typically used in clinical trials, thus this gap is understandable. A recent study found that it was the amount of training received on a rating scale that resulted in higher competence, not years of clinical experience per se (Targum, 2005). New technologies allow for the easy dissemination of this training through the use of interactive didactic tutorials, such as used in this study.

Perhaps the most important feature of this approach is the enabling of teaching and testing of applied clinical skills, a component that has been virtually absent from current rater training efforts. This is especially critical given recent evidence of the relationship between these skills and signal detection (Kobak et al., 2005). The administration of symptom rating scales requires a unique blend of clinical skills that are not necessarily the same skills used in psychotherapy or taught in academic clinical training programs. These include developing sufficient follow-up probes, avoiding leading questions, clarification of ambiguous information, and adherence to the interview protocol. Of these various skills, the training intervention in this study had its most significant impact on follow-up and adherence. These two dimensions were also among the most predictive in differentiating raters with good and bad signal detection (Kobak et al., 2005).

Length of interview was also significantly longer post-training in the enriched training group. Length of interview was found in one study to be associated with greater signal detection (Feiger et al., 2003). However, whether interview length in and of itself is predictive of improved signal detection remains to be determined (e.g., it is conceivable to have a poorly conducted interview of long duration). In a recent study, length was a necessary, but not sufficient condition for an adequate interview (Kobak et al., 2005).

No significant difference in improvement on applied skills was found between those trained via videoconference and those trained via teleconference. While this may seem counter-intuitive, several studies have found that the video signal is important in adding a "social presence" to the interview, but may not add much in terms of signal detection (Cukor et al., 1998) (it should be noted however that these studies used lower bandwidth speeds than are currently available today).

All the sites in the current study used a typical high-speed Internet connection (DSL or cable). While theoretically this should provide for speeds up to 30 frames per second, in reality, this rarely occurred, due to the fluctuation in bandwidth resulting from the number of users. As a result, the picture in this study was often choppy, and there was some audio delay. On two occasions the signal was lost during the interview. This no doubt contributed to the high percentage of trainees reporting that the mode of administration interfered with their ability to communicate with the patient when using videoconference. Interestingly, in spite of this, a higher percentage of trainees liked the video compared to telephone. Improvements in technology in time should obviate this problem (e.g., ISDN lines running at 384 kps have an almost flawless picture, and the price for this service is decreasing). In the meantime, results support the use of teleconferencing for the training of applied skills. Both trainees and patients found either method acceptable. All trainees found the feedback in the live training helpful, and all also wanted to see web trainings developed for other rating scales.

Results also confirm previous findings on the equivalence of standardized and real patients for training and testing purposes. The correlations and RAPS scores were slightly higher using standardized patients than with real patients, though these differences were not statistically significant. Interestingly, trainees were significantly more accurate in their guesses as to who were real vs standardized patients using the telephone (80% accuracy) as opposed to video (53% accuracy). While counter-intuitive, there is some data suggesting that visual cues may actually reduce accuracy in assessing human emotion, e.g., Cruz found that lower camera resolution resulted in more accurate assessment of facial affect, which he attributes to reduction of cognitive overload (Cruz, unpublished manuscript). Similarly, Strauss found that interviewers judgments of intelligence, conscientiousness and extraversion of job applicants were more accurate in job interviews conducted by telephone than by videoconference (Strauss, unpublished manuscript).

The cost of delivering enriched training is comparable to the cost of delivering rater training at a startup up meting, given the time and expense involved in travel to startup meetings. The largest expense is for clinician trainer time. If done by telephone, there could be significant cost savings. More importantly, the costs of a failed trail by using ineffective raters make the investment in proper training critical. Although labor intensive, sufficient training resources can be made available to train the raters needed for clinical trials using this approach, if the providers of rater training services focus their efforts in this direction. There is some indication that this is already occurring (Kobak et al., in press).

One caveat to the study findings is the question of stability of the results. Whether raters retain the clinical skills they have learned from the training during the course of the clinical trial has yet to be determined. Newly acquired skills may erode over time. Relatedly,

even raters who are capable of conducting a good clinical interview may conduct a poor interview due to time or enrollment pressures, or an unconscious desire to include patients into studies who have no other access to health care. Ongoing monitoring for quality control is critical, and should help obviate these problems. The American Society of Clinical Psychopharmacology recently recommended the use of audiotape monitoring for ongoing quality control (Klein et al., 2002). Rater drift may also be addressed by individual or group "refresher sessions" in which trainees take turns interviewing a real or "mock" patient, with feedback and group discussion.

In summary, the current study demonstrates that rater training can be done effectively and both didactic and applied skills are amenable to change. New technologies hold promise for enhancing training efforts, and make the knowledge accessible and cost-effective. The quality of clinical trial ratings is an important methodological variable worthy of attention and study. Future rater training efforts should be subject to empirical evaluation, and should focus on both applied skills and didactic knowledge.

Acknowledgments

This project has been funded in whole or in part with Federal funds from the National Institute of Health, National Institutes of Health, Department of Health and Human Services, under Contract No. NIH-N43MH12049 to Research Training Associates, and from a grant from Eli Lilly & Co. The authors acknowledge Dr. William Z. Potter for his generous support of this methodologic research, and Dawn Sikich, for her expert review of the clinical assessments.

References

Badger LW, deGruy F, Hartman J, et al.. Stability of standardized patients' performance in a study of clinical decision making. Family Medicine 1995;27:126–31.

Colliver JA, Swartz MH. Assessing clinical performance with standardized patients. JAMA 1997;278:790–1.

Cukor P, Baer L, Willis BS, et al.. Use of videophones and low-cost standard telephone lines to provide a social presence in telepsychiatry. Telemedicine Journal 1998;4:313–21.

De Champlain AF, Margolis MJ, King A, Klass DJ. Standardized patients' accuracy in recording examinees' behaviors using checklists. Academic Medicine 1997;72:S85–7.

Feiger A, Engelhardt N, DeBrota D, et al. Rating the raters: an evaluation of audiotaped Hamilton Depression Rating Scale (HAMD) interviews. In: 43rd annual meeting, National Institute of Mental Health, New Clinical Drug Evaluation Unit, Boca Raton, FL; 2003.

Guy W. ECDEU assessment manual for psychopharmacology, revised. National Institute of Mental Health, US Dept. of Health, Education, and Welfare, Rockville, MD. publication ADM 76–338; 1976.

Hamilton M. A rating scale for depression. Journal of Neurology, Neurosurgery and Psychiatry 1960;23:56–62.

Hamilton M. Development of a rating scale for primary depressive illness. British Journal of Social and Clinical Psychiatry 1967;6:278–96.

Hamilton M. General problems of psychiatric rating scales (especially for depression). Modern Problems of Pharmacopsychiatry Psychological Measurements in Psychopharmacology 1974;7:125–38.

Khan A, Leventhal RM, Khan SR, Brown WA. Severity of depression and response to antidepressants and placebo: an analysis of the food and drug administration database. Journal of Clinical Psychopharmacology 2002;22:40–5.

Klein DF, Thase ME, Endicott J, et al.. Improving clinical trials American Society of Clinical Psychopharmacology Recommendations. Archives of General Psychiatry 2002;59:272–8.

Kobak, KA, Engelhardt N. Standardized training on the Hamilton Depression Scale using Internet-based technologies. In: Drug information association 39th annual meeting, San Antonio, TX; 2003.

Kobak KA, Engelhardt N, Williams JBW, Lipsitz JD. Rater training in multicenter clinical trials: issues and recommendations. Journal of Clinical Psychopharmacology 2004;24:113–7.

Kobak KA, Feiger AD, Lipsitz JD. Interview quality and signal detection in clinical trials. American Journal of Psychiatry 2005;162:628.

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. Journal of Psychiatric Research 2003;37:509–15.

Kobak K, Lipsitz J, Williams J, Engelhardt N, Bellew K. A new approach to rater training and certification in a multi-center clinical trial. Journal of Clinical Psychopharmacology [in press].

Lipsitz J, Kobak KA, Feiger A, Sikich D, Moroz G, Engelhardt N. The Rater Applied Performance Scale (RAPS): development and reliability. Psychiatry Research 2004;127:147–55.

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. American Journal of Geriatric Psychiatry 2001;9:35–40.

Mott VW. The development of professional expertise in the workplace. New Directions for Adult and Continuing Education 2000;86:23–31.

Peabody JW, Luck J, Glassman P, Dresselhaus TR, Lee M. Comparison of vignettes, standardized patients, and chart abstraction: a prospective validation study of 3 methods for measuring quality. JAMA 2000;283:1715–22.

Pieters HM, Touw-Otten FW, De Melker RA. Simulated patients in assessing consultation skills of trainees in general practice vocational training: a validity study. Medical Education 1994;28:226–33.

Sheehan DV, Lecrubier Y, Sheehan K, et al.. The mini international neuropsychiatric interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM IV and ICD-10. Journal of Clinical Psychiatry 1998;59:22–3.

Spitzer RL, Williams JBW. Classification in Psychiatry. Baltimore: Williams & Wilkins; 1980.

Targum SD. Rater competency for mood disorders scales. In: 45th annual meeting, National Institute of Mental Health, New Clinical Drug Evaluation Unit, Boca Raton, FL; 2005.

Williams JBW. A structured interview guide for the Hamilton Depression Rating Scale. Archives of General Psychiatry 1988;45:742–7.