JOURNAL OF APPLIED BEHAVIOR ANALYSIS 2012, **45,** 827–832 NUMBER 4 (WINTER 2012)

OBSERVER TRAINING REVISITED: A COMPARISON OF IN VIVO AND VIDEO INSTRUCTION

CARRIE M. DEMPSEY

CALIFORNIA STATE UNIVERSITY STANISLAUS

BRIAN A. IWATA

UNIVERSITY OF FLORIDA

JENNIFER N. FRITZ

UNIVERSITY OF HOUSTON CLEAR LAKE

AND

NATALIE U. ROLIDER

KENNEDY KRIEGER INSTITUTE

We compared the effects of 2 observer-training procedures. In vivo training involved practice during actual treatment sessions. Video training involved practice while watching progressively more complex simulations. Fifty-nine undergraduate students entered 1 of the 2 training conditions sequentially according to an ABABAB design. Results showed that the 2 training methods produced almost identical scores on a posttraining observational test; however, the video method required fewer training sessions to complete.

Key words: observer training, measurement, reliability, video instruction

A longstanding tradition in applied behavior analysis is the use of direct observation as the primary means for collecting data. Although advances have been made with respect to recording apparatus (Thompson, Felce, & Symons, 2000), human observers conduct the actual recording of data because they can apply a wide range of coding procedures in situations that do not permit machine transduction.

Few studies have examined methods for conducting initial observer training. Research has shown that recording unpredictable (rather than predictable) events during practice produces better generalization to novel situations (Mash & McElwee, 1974), that it takes longer to acquire competence in observing a larger (compared to a smaller) number of events (Bass, 1987), and that supervised (compared to unsupervised) practice tends to generate the recording of a larger number of events (Wildman, Erickson, & Kent, 1975). These studies provided useful information for structuring the content or supervision of practice sessions; however, they did not evaluate training methods per se.

To identify current training practices, we conducted a survey of associate editors and editorial board members of the *Journal of Applied Behavior Analysis* and found that most varied techniques (lecture, written materials, video illustration, and live practice) somewhat informally. Because at least some verbal and written instruction is necessary (e.g., giving directions, having trainees study observational codes), significant variations in training are based on the use of live (in vivo) rather than video presentation of session content.

Address correspondence to Brian A. Iwata, 114 Psychology Building, University of Florida, Gainesville, Florida 32611 (e-mail: iwata@ufl.edu).

doi: 10.1901/jaba.2012.45-827

During in vivo training, a student records data during real-time sessions until a reliability criterion is met (usually 80% to 90%). During video training, the student observes recorded sessions played on a monitor. In vivo training takes advantage of existing data-collection opportunities and closely approximates field-study conditions. Although video training requires additional preparation, it permits the instructor to control the content of what is observed and to train more efficiently. Because these formats are quite different and because previous research used one format or the other but did not examine their relative merits, we compared the two formats in the present study. In vivo training was based on procedures that we have used for many years to train approximately 1,200 undergraduate observers. Video training involved exposure to simulated sessions that depict increasingly complex events.

METHOD

Subjects and Setting

Subjects were 59 undergraduate students who were enrolled in a laboratory course across consecutive semesters. All completed observer training as a course requirement, but performance during training did not enter into their final grades. Training was conducted during 3-hr blocks twice per week at subjects' lab sites (vocational or educational programs for individuals with intellectual disabilities).

Apparatus and Materials

Hardware. Subjects recorded data on Macintosh laptop computers or on Palm PDAs. During video-training sessions and all posttest sessions, subjects viewed videotapes on a TV/VCR player.

Posttest. To standardize the posttest while approximating typical observation conditions, we videotaped a variety of ongoing sessions (skill acquisition, treatment of problem behavior, parent training) that subjects would observe after training. Videos were edited into 10 3-min segments that contained 351 behavioral events. These events consisted of 17 individuals who exhibited varying frequencies of 40 responses (targets).

Video training series. Six 10-min video segments that featured simulated acquisition, functional analysis, and treatment sessions were produced. The videos depicted one or more clients, therapists, teachers, and parents, all of whom were role-played by research assistants.[1]

Procedure

Subjects were assigned to either in vivo training or video training. Although the resulting comparison was on a between-subjects basis, subjects were not assigned randomly to one condition or another at the outset of the study. Instead, subjects were assigned to in vivo, video, or replication phases across consecutive semesters in an alternating ABABAB arrangement. Twenty-six individuals participated in Phase A, and 33 participated in Phase B. All subjects first received approximately 30 min of group instruction on data-collection apparatus, operational definitions, and reliability calculation.

In vivo training. Subjects recorded data during live sessions that typically lasted for 10 min and included a therapist and client. Prior to each session, the therapist gave the subject written and verbal descriptions of target behaviors, definitions, and procedures. The subject had access to the written description throughout the session. During the session, the subject recorded data with a highly experienced independent observer from locations approximately 2 m outside the session area. After each session, percentage agreement between the subject's and trained observer's data records was calculated. Training was completed when a subject attained a minimum of 90% agreement on three consecutive sessions with two clients (our traditional training criterion, which we

[1] A CD version of the video training program, including practice segments and scoring keys, is available from the corresponding author for the cost of media and shipping.

selected as the benchmark for comparison with video training). The criterion for the last group of in vivo subjects ($n = 10$) was modified to be identical to that used for the video subjects (see below): Training was completed when a subject attained a minimum of 90% agreement on six sessions (consecutive or nonconsecutive) across any number of clients.

Video training. Subjects recorded data from videotaped simulations of acquisition, assessment, and treatment sessions. Each video followed a script whose complexity increased across the series based on the number of individuals, target behaviors, and total events observed (see Table 1). Subjects reviewed a written description of the session prior to each session and had access to the description while recording data from a TV/VCR, which was located approximately 1 m in front of the subject (pausing or rewinding the videotape was not permitted). After each session, percentage agreement was calculated between the subject's data record and one on which 100% agreement had been obtained by trained observers. Each subject began with the first video in the sequence and continued to practice on that video until he or she attained an agreement score of at least 90%, at which time the subject proceeded to the next video in the sequence. Training was completed when a subject attained a minimum of 90% agreement across the six video segments.

Posttest. After completing either type of training, subjects recorded data from 10 3-min videotaped sessions. They had access to written descriptions of session content prior to and during the posttest. Reliability scores for the posttest were calculated by comparing the subject's data record with one on which 100% agreement had been obtained by two trained observers.

Follow-up. As an additional measure of generalization, we collected a random sample of subjects' agreement scores from actual (in vivo) sessions for which they had collected data within 1 month after training.

Table 1
Number of Clients, Targets, and Events Shown in Each Video Training Segment

Segment	Clients	Targets	Total events
1	1	1	30
2	1	2	60
3	2	3	90
4	2	4	120
5	3	5	150
6	3	6	180

Interobserver Agreement

Reliability was calculated in three ways. First, subjects compared their own practice records to the record of a trained observer (in vivo training) or a record on which two trained observers previously had obtained 100% agreement (video training). Agreement was calculated by dividing session time into 10-s intervals, dividing the smaller number of responses by the larger number of responses in each interval, and averaging these fractions across the session. These data were used as criteria for progressing through the training sequence. Mean agreement across all subjects during training segments was 92% (range, 89% to 99%). Second, a graduate student rechecked subjects' scores by recalculating the data records obtained by each subject during training. Third, two graduate students independently calculated each subject's posttest agreement score (the main dependent variable). Agreement on training records and posttest scores was calculated by dividing the smaller score by the larger score, which yielded 100% agreement for all subjects' training record and posttest scores.

RESULTS AND DISCUSSION

Figure 1 (top) shows posttest scores for all subjects, which were virtually identical across training conditions. Subjects who received in vivo training scored a mean of 86% (range, 82.9% to 88.6%) with the more stringent criterion and 85% (range, 79.2% to 87.4%) with the less stringent criterion, and subjects who received video training scored a mean of 86% (range, 80.8% to 89.5%). Data from the

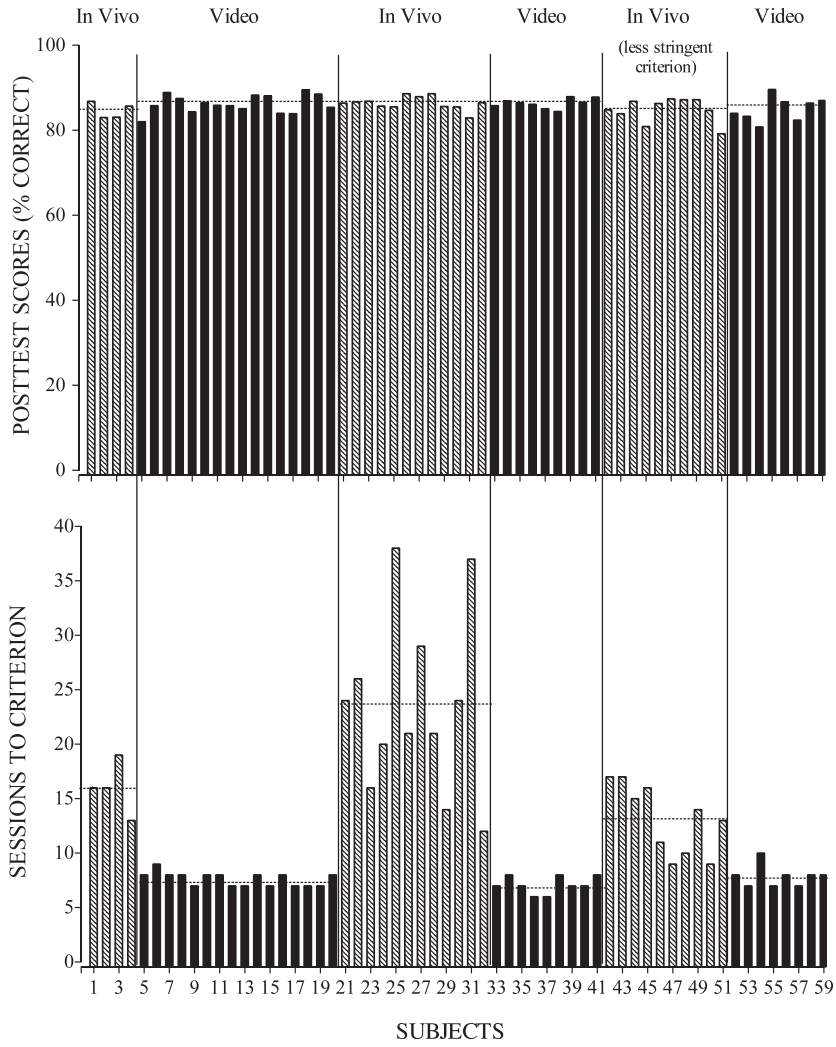

Figure 1. Posttest score (top) and number of sessions required to complete training (bottom) by subjects under in vivo and video training. Dashed horizontal lines show condition means.

follow-up sample showed that subjects who received both types of training continued to improve following training: Mean agreement scores were 98.9% (range, 96.9% to 100%) for in vivo subjects ($n = 8$) and 96.2% (range, 88.5% to 100%) for video subjects ($n = 8$).

Figure 1 (bottom) shows the number of sessions required to meet the termination criterion. With the exception of one subject (54), every subject who received video training reached criterion faster than every subject who received in vivo training. Video subjects required a mean of eight (range, 6 to 10) practice sessions to meet criterion; in vivo subjects required a mean of 22 (range, 12 to 38) sessions with the more stringent criterion and 13 (range, 9 to 17 sessions) sessions with the less stringent criterion.

The in vivo and video procedures produced uniformly high performance on the posttest and during a randomly selected posttraining follow-up session. Thus, both methods were effective in training observers to a high standard of proficiency. However, subjects exposed to video

training reached the terminal criterion in fewer training sessions. Thus, decisions to use in vivo or video observer training might be based on practical considerations. In vivo training requires little advanced preparation; trainees simply practice with a competent observer. However, because content may vary in an uncontrolled way across sessions, initial skill acquisition may take longer, especially if observers record data for different types of sessions. In vivo training also may be difficult to implement with larger groups if the number of live sessions available at a given time is limited or if a limited number of trainees can practice simultaneously. Thus, in vivo training may be preferred when there are few time constraints posed by duration of training or number of trainees. Alternatively, video training allows precise control over session content, which can be manipulated to introduce complexity gradually and depict all relevant variables during practice. Although an initial time investment is required to prepare training media, its continued availability allows the training of a large number of observers and repeated use with subsequent cohorts. Consequently, video training may be preferred when efficiency is an important consideration.

Several potential limitations to the study should be noted. First, the absence of pretest scores did not allow us to determine subjects' performance prior to training, although there was no reason to suspect that untrained observers would have performed differently across the training conditions to which they were assigned.

Second, we used a video format for the posttest because it would have been impossible to conduct a uniform live posttest across trainees, raising the question of whether subjects exposed to video training might have had an unfair advantage during the posttest. However, content of the posttest (taken from actual sessions) more closely resembled what was observed during in vivo training. The similarity in scores obtained by in vivo and video subjects under both posttest (video) and follow-up (in vivo) conditions suggests that neither the format nor the content of the posttest influenced performance noticeably.

Third, the criterion for completing training by the first two in vivo training cohorts (90% accuracy on three successive sessions for two clients) may have been more stringent than the criterion imposed during video training. Therefore, we used a less stringent criterion during the final in vivo training condition. Despite this more lenient criterion, in vivo training still required approximately a third more training time than video training did.

Finally, because the content of in vivo training was not controlled, it was possible that in vivo subjects were exposed to fewer training exemplars than were video subjects. A review of the data records indicated that both groups of subjects were exposed to roughly similar content in the current study. In vivo subjects recorded data on one to three individuals who exhibited two to seven target events (eight to 123 events overall), whereas video subjects recorded data on one to three individuals who exhibited one to six targets (30 to 180 events overall). Despite the overall correspondence in content, in vivo subjects were exposed to content that varied in a relatively unsystematic fashion. For example, it was not uncommon for a session with five target responses to be followed by a session with two target responses, or for a session with 50 events to be followed by a session with three events. This observation lends support to the idea that systematic variation in complexity associated with video training contributed to its efficiency.

In summary, video-based observer training may represent an attractive method for developing observational skills in a relatively short amount of time. Differences between video and in vivo training in this study were numerous and could be examined in future studies: (a) format (video vs. live), (b) potential number of uncontrolled visual distractions during in vivo observation, (c) graduated (video) versus un-

controlled variation in complexity, and (d) simulated (video) versus actual (in vivo) session content, although some of these may be inherent differences between the two training methods. Future research also should investigate other factors that influence training efficiency, accuracy, and skill maintenance over time (e.g., see Lerman et al., 2010). Further, computerized methods of training might be examined because they allow more precise control over observational content and scoring.

REFERENCES

Bass, R. F. (1987). Computer-assisted observer training. *Journal of Applied Behavior Analysis*, *20*, 83–88. doi:10.1901/jaba.1987.20-83

Lerman, D. C., Tetreault, A., Hovanetz, A., Bellaci, E., Miller, J., Karp, H., et al. (2010). Applying signal detection theory to the study of observer accuracy and bias in behavioral assessment. *Journal of Applied Behavior Analysis*, *43*, 195–214. doi:10.1901/jaba.2010.43-195

Mash, E. J., & McElwee, J. D. (1974). Situational effects on observer accuracy: Behavioral predictability, prior experience, and complexity of coding categories. *Child Development*, *45*, 367–377. doi:10.2307/1127957

Thompson, T., Felce, D., & Symons, F. J. (Eds.). (2000). *Behavioral observation*. Baltimore, MD: Brookes.

Wildman, B. G., Erickson, M. T., & Kent, R. N. (1975). The effect of two training procedures on observer agreement and variability of behavior ratings. *Child Development*, *46*, 520–524. doi:10.2307/1128151

Received September 8, 2006
Final acceptance August 9, 2012
Action Editor, Mark Dixon