JOURNAL OF APPLIED BEHAVIOR ANALYSIS 2010, **43,** 195–213 NUMBER 2 (SUMMER 2010)

APPLYING SIGNAL-DETECTION THEORY TO THE STUDY OF OBSERVER ACCURACY AND BIAS IN BEHAVIORAL ASSESSMENT

DOROTHEA C. LERMAN, ALLISON TETREAULT, ALYSON HOVANETZ, EMILY BELLACI, JONATHAN MILLER, HILARY KARP, ANGELA MAHMOOD, MAGGIE STROBEL, SHELLEY MULLEN, ALICE KEYL, AND ALEXIS TOUPARD

UNIVERSITY OF HOUSTON, CLEAR LAKE

We evaluated the feasibility and utility of a laboratory model for examining observer accuracy within the framework of signal-detection theory (SDT). Sixty-one individuals collected data on aggression while viewing videotaped segments of simulated teacher–child interactions. The purpose of Experiment 1 was to determine if brief feedback and contingencies for scoring accurately would bias responding reliably. Experiment 2 focused on one variable (specificity of the operational definition) that we hypothesized might decrease the likelihood of bias. The effects of social consequences and information about expected behavior change were examined in Experiment 3. Results indicated that feedback and contingencies reliably biased responding and that the clarity of the definition only moderately affected this outcome.

Key words: behavioral assessment, college students, data collection, observer bias, signal-detection theory

Direct observation and measurement of behavior are the cornerstones of effective research and practice in applied behavior analysis. Trained observers use various methods to record occurrences of precisely defined target behaviors and other events during designated observational periods. In application, behavioral consultants often rely on parents, teachers, and direct-care staff to collect data on target behaviors. The behavioral consultant examines these data to obtain information that is key to program effectiveness, such as the baseline level of responding, the conditions under which a behavior occurs, and changes in responding with the introduction of treatment or modifications to existing procedures. Little research has been conducted on the accuracy of data collected by direct-care staff or the best way to train people to collect these data.

Interobserver agreement, which is determined by having two observers record the same events at the same time, is routinely reported in published research to provide some degree of confidence in the accuracy of the reported data. However, practitioners do not routinely collect data on interobserver agreement (i.e., reliability). Furthermore, agreement is not synonymous with accuracy (i.e., two observers could agree but incorrectly score the behavior; Kazdin, 1977; Mudford, Martin, Hui, & Taylor, 2009).

A number of studies have identified variables that might lead observers to record data inaccurately. Research findings indicate that factors related to the measurement system (e.g., number of different behaviors scored), characteristics of the observers (e.g., duration of training), characteristics of the setting (e.g., presence of other observers), and consequences for scoring (e.g., social approval for recording changes in the level of the target behavior) can influence the accuracy and reliability of behavioral measurement (see Kazdin, 1977; Repp, Nieminen, Olinger, & Brusca, 1988, for reviews). A large portion of these studies, however, focused on interobserver agreement rather than accuracy. Furthermore, the nature of inconsistencies or inaccuracies in data collection was not systematically examined. For example, sources of random error versus nonrandom error (i.e., observer bias) have not been differentiated in previous research.

We thank Jennifer Fritz for her assistance in conducting this study.

Address correspondence to Dorothea C. Lerman, 2700 Bay Area Blvd., Box 245, Houston, Texas 77058 (e-mail: lerman@uhcl.edu).

doi: 10.1901/jaba.2010.43-195

Signal-detection theory (SDT; Green & Swets, 1966) may provide a useful framework for further analysis of observer accuracy. SDT was developed to examine the behavior of an observer in the presence of ambiguous stimuli. The task of the observer is to discriminate the presence versus absence of a stimulus (i.e., detect a signal against a background of noise). In classical signal-detection experiments, the observer either responds ''yes'' or ''no'' regarding the presence of the signal on each trial. Correctly indicating that a stimulus is present is called a *hit*, and correctly indicating that a stimulus is absent is called a *correct rejection*. Indicating that a stimulus is absent when it is actually present is called a *miss*, and indicating that a stimulus is present when it is actually absent is called a *false alarm*.

According to SDT, the behavior of an observer in this type of situation has at least two dimensions. One dimension is determined by the sensory capability of the observer and the actual ambiguity of the stimulus and is called the *sensitivity* of the observer (i.e., how well the observer discriminates the signal from the noise). A second dimension is the proclivity of the observer to judge in one direction as opposed to the other (e.g., to indicate that the signal is present rather than absent), referred to as the observer's *response bias*. Research on SDT indicates that response bias is affected by a number of variables, including the consequences for each outcome of judgment, the a priori probability of each option, the decision rule that influences the observer, and instructions about how to make the observations (Green & Swets, 1966). Sensitivity, on the other hand, is usually affected only by operations that change the amount of ambiguity in the stimulus situation. SDT provides a way to evaluate the effect of factors on sensitivity and response bias separately.

Methods based on SDT have been applied across a variety of disciplines (e.g., medicine, industry, psychiatry, engineering) to evaluate decision making, including clinical diagnosis and assessment (see McFall & Treat, 1999; Swets, 1988, 1996, for reviews). In the experimental analysis of behavior, signal-detection methods have been used to study stimulus control and reinforcement effects in choice situations (e.g., Alsop & Porritt, 2006; Davison & McCarthy, 1987; Nevin, Olson, Mandell, & Yarensky, 1975).

The concepts of SDT also could be extended to the direct observation of behavior in research and clinical settings. Any behavior that should be recorded by observers is analogous to the signal in SDT. All other behaviors are analogous to the noise in SDT. Correctly recording that a behavior has occurred is analogous to responding, ''Yes, the signal is present'' (i.e., a hit). Correctly refraining from recording a behavior that does not meet the definition of the target response is analogous to responding, ''No, the signal is absent'' (i.e., a correct rejection). Failing to record a behavior that has occurred is analogous to responding incorrectly in the presence of the signal (i.e., a miss), whereas recording a behavior that did not occur is analogous to responding incorrectly in the presence of noise (i.e., a false alarm). Research on SDT indicates that observer error may reflect problems with sensitivity (i.e., discriminating the target behavior from other behaviors) or response bias (i.e., the criterion used by the observer to determine whether a behavior should be recorded). SDT also suggests that problems with sensitivity and bias are more likely to occur when the observer encounters ambiguous samples of the targeted behaviors.

Thus far, research on observer accuracy in behavioral assessment has not differentiated between sensitivity and bias or considered the role of ambiguous behavioral samples when examining factors that may influence accuracy or reliability. Various types of ambiguous behavioral samples might arise during naturalistic observation. Three types will be described

for illustrative purposes. First, a particular event may possess a subset (rather than all) of the criteria specified in the defined response class. For example, the definition of a tantrum might be "screaming and falling to the floor," such that both responses must be present for a tantrum to be scored. A sample consisting of falling to the floor in the absence of screaming might be associated with inconsistent or erroneous data collection (i.e., a false alarm). Second, a particular event may possess all of the criteria specified in the behavioral definition but include other elements that differentiate the sample from other members of the defined response class. For example, a behavioral sample consisting of screaming and laughing while falling to the floor may appear ambiguous to the observer who is scoring tantrums as defined above. Such a sample may increase the likelihood of inconsistent or erroneous data collection (i.e., a miss). Finally, ambiguity may arise when the nature of the behavior makes it difficult to specify samples that should be included and excluded in the behavioral definition. For example, *inappropriate vocalizations* might be defined as vocalizations unrelated to the topic being discussed or to stimuli in the environment (e.g., DeLeon, Arnold, Rodriguez-Catter, & Uy, 2003). Such a definition permits some degree of subjective interpretation. For example, disagreement might occur between two observers when an individual states, "I love marshmallows," after hearing someone say, "There are some beautiful, fluffy clouds in the sky today."

Factors that influence observer accuracy may be particularly problematic when an observer is faced with these types of ambiguous samples. Nonetheless, in prior research on observer accuracy and reliability, the nature of the behavioral samples (i.e., clear vs. ambiguous) was uncontrolled. Detailed analyses of observer errors (e.g., percentage of misses vs. false alarms) also have rarely been conducted. These gaps limit our knowledge about how and why different variables affect the accuracy and consistency of behavioral measurement. For example, several studies have shown that the presence of another observer (or general knowledge of monitoring) improves observer accuracy (Repp et al., 1988). It is possible that this factor influences the observer's criterion for a hit (e.g., the observer adopts a more conservative criterion for the target behavior) and, thus, is less likely to record false alarms. Alternatively, the observer may attend more closely to the behavior samples (i.e., increase vigilance without changing the criterion), thereby decreasing the number of misses. Similar changes in overall accuracy would occur but for very different reasons.

Analyses of measurement based on SDT might better differentiate among factors that influence response bias and sensitivity, leading to a greater understanding of factors that influence measurement and, thus, better solutions for rectifying these problems. For example, response bias would be indicated if a factor produces similar effects on both hits and false alarms. A simultaneous increase (or decrease) in hits and false alarms suggests that the observer has altered his or her criterion for scoring a behavior.

The identification of stimuli and conditions that reliably bias observers' responding as predicted by SDT would be useful for the further study of clinically relevant variables that may affect the accuracy of observations in the natural environment. The purpose of this study was to develop and test a procedure for evaluating factors that may influence observer accuracy and bias in behavioral assessment. As a first step, we sought to determine whether we could reliably bias responding in the laboratory by having individuals score videotaped segments of simulated child–teacher interactions with clear and ambiguous samples of designated behaviors. Based on previous research on SDT, we hypothesized that (a) hits and false alarms would increase when observers were given brief

feedback and told that they would receive monetary points for each hit, (b) hits and false alarms would decrease when observers were given brief feedback and told that they would lose points for each false alarm, and that (c) changes in hits and false alarms would be more likely to occur with samples designated as ambiguous rather than as clear.

After evaluating the utility of our procedures in Experiment 1, we conducted two additional studies on factors that might alter observer bias in the presence of ambiguous events. Experiment 2 focused on the specificity of the operational definition. We hypothesized that this variable might decrease the likelihood of bias under conditions found to bias responding reliably (i.e., the factors manipulated in Experiment 1). The relation between response bias and two other clinically relevant factors that have been shown to alter observers' recordings—social consequences and information about expected behavior change—were examined in Experiment 3.

GENERAL METHOD EXPERIMENTS 1 AND 2

Participants and Setting

Graduate students enrolled in a learning principles course and undergraduate students enrolled in a research and statistics course were recruited for Experiments 1 and 2. Nineteen graduate students (22 years to 53 years old; M = 30 years) and 18 undergraduate students (20 years to 59 years old; M = 33 years) participated in one of two experiments (none participated in both). The graduate students received extra credit in their learning principles course for participating. The undergraduate students fulfilled a course requirement in their research and statistics class. (Students could choose among a variety of available experiments or select a nonresearch option to satisfy these course requirements.) Because similar results were obtained for the graduate and undergraduate participants, information about these two groups is not differentiated on this basis in the remainder of the paper. All participants (20 women, 17 men) completed an interview with the experimenter regarding previous experience with data collection and direct observation of behavior and prior course work in behavior analysis, behavior therapy, or behavior modification. The participants reported little or no previous relevant data-collection experience or course work (beyond a learning principles course). All sessions took place in a laboratory room that contained a table, chairs, TV/VCR, and handheld computer. In addition to these participants, an individual with a doctoral degree in behavior analysis and 8 years of experience collecting direct-observation data for both research and clinical purposes was recruited to participate in the experiments as an expert data collector (see further explanation below). She was naive to the purpose of the study.

Materials

A series of eight vignettes were developed and videotaped prior to the study. Each vignette showed a teacher instructing a student on a particular task (e.g., sweeping the floor, sorting objects at a table, playing with leisure materials) for approximately 4 min. The same two actors appeared in each vignette, and the total video lasted 33 min. The actors followed prepared scripts such that each vignette consisted of two or three clear samples of the target behavior (aggression, defined as "hitting the teacher, kicking the teacher, and throwing objects at the teacher"), two or three ambiguous samples of the target behavior, and five ambiguous nonexamples of the target behavior. The 33-min segment used in Experiments 1 and 2 contained a total of 20 clear samples of the target behavior, 20 ambiguous samples of the target behavior, and 40 ambiguous nonexamples of the target behavior (i.e., 40 possible hits and 40 possible false alarms [if, in fact, false alarms involved only ambiguous samples]). When creating the scripts for the vignettes, the experimenters classified samples as clear or

Table 1
Examples of Clear and Ambiguous Samples and Nonexamples of Aggression Shown in the Videotaped Segments

	Clear	Ambiguous
Samples (signals)	Hitting the teacher on the shoulder with a closed fist while screaming "no," kicking the teacher when she begins to provide a physical prompt, throwing task materials at the teacher when given an instruction.	Hitting the teacher's hand while reaching for task materials, hitting the teacher's shoulder while smiling, hitting the teacher on the arm with an object while playing with the object.
Nonexamples (noise)	Complying to an instruction, engaging in a task.	Screaming at the teacher, hitting own head, throwing objects in the opposite direction of the teacher; swinging an arm towards teacher without making contact.

ambiguous on the basis of expected outcomes. Examples of samples from each category are shown in Table 1. The remaining interactions consisted of clear nonexamples of the target behavior (e.g., compliance to instructions). Each behavioral sample of the student was separated by at least 10 s from the previous sample, so that observer accuracy could be determined (see below for further description). Three different versions of the 33-min segment were constructed by altering the order of the vignettes such that no vignette occurred contiguous with another vignette in more than one version. A voice stating "one, two, three, start," was inserted prior to the first vignette of each segment to cue initiation of the data-collection program. Observers used handheld PCs with external keyboards (Dell Axim) and data-collection software (Instant Data) to collect data. The software generated a text record of each key pressed and the moment that it was pressed for the entire observation.

To generate "gold standard" data records for each version of the 33-min videotaped segment, a group of six trained observers used the same handheld PCs and software to score occurrences of clear and ambiguous samples of aggression as well as ambiguous nonexamples of aggression. Prior to and during the scoring, the observers had access to the written scripts of the vignettes but did not interact with each other during the scoring. Different key codes were used for each of these three response categories. Working in pairs, the observers compared their data records following the scoring of an entire 33-min segment. If the precise timing of any event differed by more than 3 s between the observers, the segment was rescored until no such discrepancies occurred (this rarely happened). One of the observers' records was selected randomly to serve as the gold standard data record for each version of the videotape. To verify the accuracy of the gold standard data, the first author compared these records to the written scripts and the written scripts to the videotaped vignettes.

Response Measurement and Interrater Agreement

The primary measures of interest were the number of hits and false alarms in each scoring session. A hit was defined as the observer scoring the occurrence of aggression (designated as an "a" on his or her data record) within ± 5 s of the occurrence of aggression on the gold standard data record. A false alarm was defined as the observer scoring the occurrence of aggression more than ± 5 s from the occurrence of aggression on the gold standard data record. Each participant's hit rate was determined by dividing the total number hits by the total number of clear and ambiguous samples of aggression (40). The false alarm rate was determined by dividing the total number of false alarms by the total number of ambiguous nonexamples of aggression (40). All of the data records were scored independently by two experimenters, and the results were compared. The experimenters rescored any data records

that contained one or more disagreements. The results were again compared, and interrater agreement of 100% was obtained.

Procedure

Participants were escorted into the lab and given the following written information and instructions, which the experimenter read aloud:

> In my field (applied behavior analysis), direct observation of behavior is critical to both research and practice. We rely on observers to accurately record the occurrence of behaviors in field settings, such as homes, schools, and clinics. We must train observers to score data. The purpose of this study is to examine efficient ways to train observers and to evaluate the benefits of using handheld computers to collect data in field settings.
>
> We are simulating the procedures typically used to train observers who score behavior in the field. You will be viewing a series of videotaped segments with two actors who are pretending to be a teacher and a student with developmental disabilities. Each segment consists of a series of 2- to 5-min clips; you should score all of the clips in a segment as one continuous observation. First, you will be scoring some segments as practice sessions. During these practice sessions, you may discuss the scoring with me and ask questions. Once you feel comfortable with the scoring, you will score three additional 33-minute videotaped segments as though you are actually out in the field. Once you have begun this scoring, I will not be able to answer any questions about the data collection because we will be pretending that you are out in the field. Just do the best that you can.
>
> You will be scoring the following behaviors by pressing the appropriate key, as shown in parentheses: Aggression (a): Score whenever the student hits the teacher, kicks the teacher, or throws objects at the teacher. Praise (c): Score whenever the teacher delivers a statement that includes praise, such as "Nice working," "I like that," "Good job," "That was a good one," etc. Tone of voice is not relevant.
>
> It is important to score the data as accurately as possible. Press the appropriate key as soon as the behavior occurs. Any keys that are pressed more than 5 seconds after the behavior occurs will not be considered correct. Do not score additional occurrences of a behavior if they follow the initial behavior by less than 5 seconds. For example, if the student hits the teacher three times very quickly, only score the first behavior of this "burst" or "episode." If the teacher delivers multiple praise statements following the student's behavior, only score the first statement.

Depending on the condition, the participant also was told about the possibility of earning points for scoring correctly (see further description below). Participants were told to score praise in addition to aggression (the true target) to increase the demand of the observation and to reduce possible reactivity.

After receiving the instructions, all participants viewed a 16-min practice video consisting of eight vignettes that resembled those of the test video. However, the practice video contained only clear samples and clear nonexamples of the target behavior. Following the first practice session, the experimenter discussed any errors in the participant's scoring. Additional 10-min practice sessions continued (using different versions of the practice video) until the participant made no errors in the scoring of aggression. The instructor then asked the participant if he or she was ready to begin the actual scoring (all participants indicated that they were ready). At this point, the written instructions were removed from the room. Across both experiments, participants required a mean of 20 min to meet the training criterion (range, 16 to 36 min).

Experiment 1

The purpose of Experiment 1 was to examine the effects of instructions that included information about contingencies and brief feedback on the rate of hits and false alarms. The goal was to determine if our procedures would be useful for the further study of observer accuracy and bias within the framework of SDT.

Procedure

Twenty participants scored three versions of the 33-min video. For 15 participants, factors intended to bias responding were manipulated for two of the three scorings. The remaining 5 participants (control group) were not exposed to the experimental manipulation. These participants simply scored the three videos to evaluate patterns of responding in the absence of any factors designed to alter criterion. The

participants were assigned randomly to the experimental and control groups. The expert data collector who was recruited for the study was asked to score one version of the 33-min video. She was not exposed to the experimental manipulation. Our intention was to use her scoring as a measure of the quality of the video. That is, we assumed that someone with expert levels of experience would capture most, if not all, instances of aggression (both clear and ambiguous) if these responses were clearly visible in the video. The data-collection expert was not given any feedback about her scoring.

In addition to the instructions shown above, all participants, with the exception of those in the control group and the data-collection expert, were told, ''I will tell you if you can earn points for recording data in the upcoming segment and how you can earn these points.'' These participants also were told that the points would be exchangeable for money at the end of the sessions and that they could earn up to $40 (graduate students) or $20 (undergraduate students). After completion of the practice sessions, all participants were exposed to Condition A (baseline). Participants in the control group were then exposed to two additional scoring sessions under Condition A. The remaining participants were exposed to either Condition B (consequences for hits) or Condition C (consequences for false alarms) for the second scoring, depending on the pattern of scoring during baseline (see further description below). The participant's performance during the second scoring determined whether Condition B or Condition C was implemented during the third scoring. Participants received a 10-min break between each scoring session.

Baseline (A). Prior to the start of the video segment, participants were told, ''For this segment, you will not have an opportunity to earn points. Just score as accurately as you can.'' (Participants in the control group were just told to score as accurately as they could.) During the session break, the experimenter compared the participant's data record to the gold standard data record. If the participant had more misses than false alarms, the participant was exposed to Condition B for the second scoring session. If the participant had more false alarms than misses, the participant was exposed to Condition C for the second scoring.

Consequences for hits (B). Prior to the scoring, participants were told,

> When downloading your data record, I noticed that you missed some of the aggressions in the last segment. It's really important to catch all of the aggressions that occur. Thus, for this segment, you will have an opportunity to earn points for correctly scoring aggression. You will earn one point for each aggression that you catch. The more points that you earn, the more money you will receive.

The statements above included a brief feedback statement because feedback is typically included in signal-detection procedures. Participants were told about the importance of hits to increase the saliency of the consequences for hits.[1] Following the scoring session, the experimenter compared the participant's data record to the gold standard data record. The participant remained in Condition B for the third scoring session as long as the frequency of hits was equal to or less than 37 (of a total possible 40). Otherwise, the participant was exposed to Condition C for the third scoring (this happened for 2 participants).

Consequences for false alarms (C). Prior to the scoring, participants were told,

> When downloading your data record, I noticed that you scored aggression at times when aggression did not occur. It is really important not to include things that aren't aggression. Thus, for this segment, you will start with a certain number of points, and you will lose one point each time that you score something as aggression when it is not. You will earn however many points are left at the end of the

[1] At the conclusion of this experiment, 10 additional participants were recruited. Half were exposed to feedback only and the other half were exposed to the point contingency only (combined with the statement about the importance of hits and avoiding false alarms; see also Condition C). Results suggested that both components were necessary to reliably bias responding. Data are available from the first author.

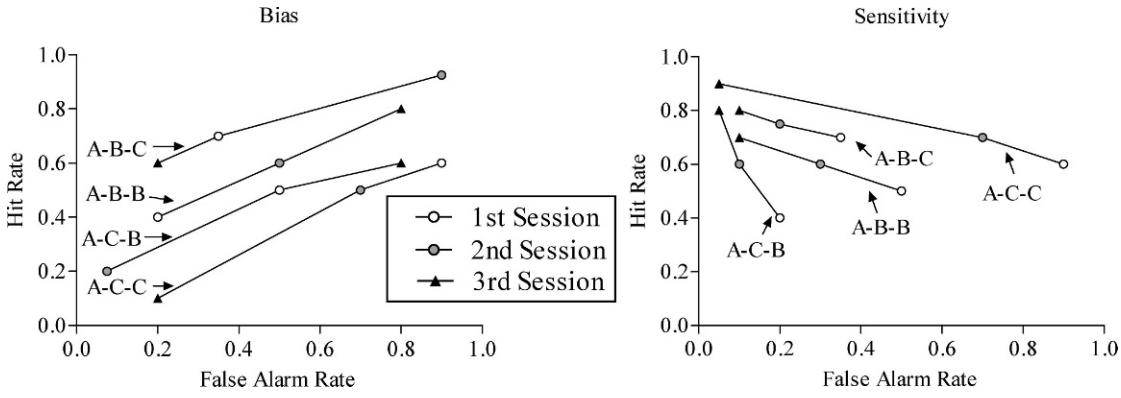

Figure 1. Response patterns across the three scoring sessions that would indicate manipulation of response bias (left) or changes in sensitivity (right) for the four possible sequences of conditions (A-B-B, A-C-C, A-B-C, A-C-B).

session. The more points that you earn, the more money you will receive.

As noted above, participants were given the brief feedback statement because feedback is typically included in signal-detection procedures. Participants were told about the importance of avoiding false alarms to increase the saliency of the consequences for false alarms. Following the scoring session, the experimenter compared the participant's data record to the gold standard data record. If the participant was completing the second scoring session in this condition, the participant remained in Condition C for the third scoring session as long as the frequency of false alarms was at or above three. Otherwise, the participant would have been exposed to Condition B for the third scoring (however, this did not happen for any participants).

Data Analysis

The performance of each participant was plotted on a graph displaying the hit rate (vertical axis) against the false alarm rate (horizontal axis) for each observation session. With this type of display, data points that appear close to the upper right corner of the graph indicate a high hit rate with a high false alarm rate (tendency to record events). Data points that appear close to the lower left corner of the graph indicate a low hit rate with a low false alarm rate (tendency to refrain from recording). Data points that appear closer to the upper left corner of the graph reflect higher levels of accuracy (i.e., a high hit rate with a low false alarm rate), whereas data points that appear closer to the lower right corner of the graph reflect lower levels of accuracy (i.e., low hit rate with a high false alarm rate). Data for the three observation sessions were plotted on the same graph for each participant. Different symbols were used to indicate the session number, and the three points were connected by a line for visual inspection purposes.

For the purposes of illustration, response patterns that indicate manipulation of response bias are shown in the left panel of Figure 1 for the four possible sequences of conditions (A-B-B, A-C-C, A-B-C, A-C-B). Manipulation of response bias as predicted by SDT is indicated by a positive relation between the hit rate and false alarm rate across the three sessions, with the direction of the relation determined by the conditions experienced by the participant. Data points move towards the upper right corner of the graph from one session to the next session given a greater tendency to record events (i.e., an increase in the rate of hits and false alarms). Data points move towards the lower left corner of the graph from one session to the next session given a greater tendency to refrain from recording events (i.e., a decrease in the rate of hits and false alarms).

Response patterns that reflect systematic changes in sensitivity across the three sessions are shown in the right panel of Figure 1. A negative relation between the hit rate and false alarm rate (i.e., data points moving towards the upper left corner or the lower right corner of the graph) indicates a change in sensitivity. For this particular study, we would expect to see an increase in the hit rate with a corresponding decrease in the false alarm rate (i.e., an increase in accuracy) if the manipulated contingencies altered sensitivity for any participants. Thus, in Figure 1, the pattern of responding across the three sessions is identical regardless of the conditions experienced by the participant.

The mean percentages of clear samples of the target behavior, ambiguous samples of the target behavior, and ambiguous nonexamples of the target behavior scored by participants also were calculated to determine if errors were more likely to occur with samples considered to be ambiguous than with those considered to be clear.

Results and Discussion

Plots showing the hit rates and false alarm rates for the individuals who participated in Experiment 1 are displayed in the left column of Figure 2. The data are grouped according to the sequence of conditions experienced by the participant (i.e., A-B-B, A-C-C, A-B-C, or A-A-A). The hit rate and false alarm rate for the session scored by the data-collection expert are plotted in the top left panel (displayed with an asterisk). Results for all participants (with the exception of the expert and those in the control group) showed a positive relation between the hit rate and false alarm rate across the three sessions. Data from the second and third sessions moved in the expected direction with respect to the previous sessions, given the consequences that were experienced (i.e., monetary points earned for hits vs. monetary points lost for false alarms). That is, Condition B (monetary points earned for hits) was always associated with an increase in both hits and false alarms relative to the previous session. Condition C (monetary points lost for false alarms) was always associated with a decrease in both hits and false alarms relative to the previous session. These findings indicate that the procedures reliably biased responding. It should be noted that the outcomes for the 2 participants who experienced the A-B-C sequence of conditions deviated slightly from the predicted response pattern (see Figure 1) in that the data point for the third session (Condition C) did not fall to the left of the first session (Condition A or no consequences). That is, although the participants' hit and false alarm rates were lower than in the previous session, they did not fall below baseline rates (i.e., Condition A). This may have occurred due to their initial low false alarm rates under Condition A or to their exposure to Condition B in the prior session (i.e., sequence effects).

The hit rate for the data-collection expert, who missed five instances of aggression (all categorized as ambiguous), was equal to or greater than that obtained for any participant under Condition A (first scoring session). Although this might suggest that some instances of aggression were not clearly visible in the video, a large number of participants exceeded her hit rate under other conditions. As such, it seems likely that the expert's misses were determined by her criteria for scoring an event as an instance of aggression (rather than her ability to detect the event). This possibility was further explored in Experiment 2.

The mean percentage of clear samples of the target behavior, ambiguous samples of the target behavior, and ambiguous nonexamples of the target behavior scored by participants are shown in Figure 3. The participants scored nearly all clear samples of aggression, regardless of the condition. On the other hand, the mean percentage of ambiguous events scored by the participants depended on the condition, with a higher percentage scored in Condition B than in the other two conditions.

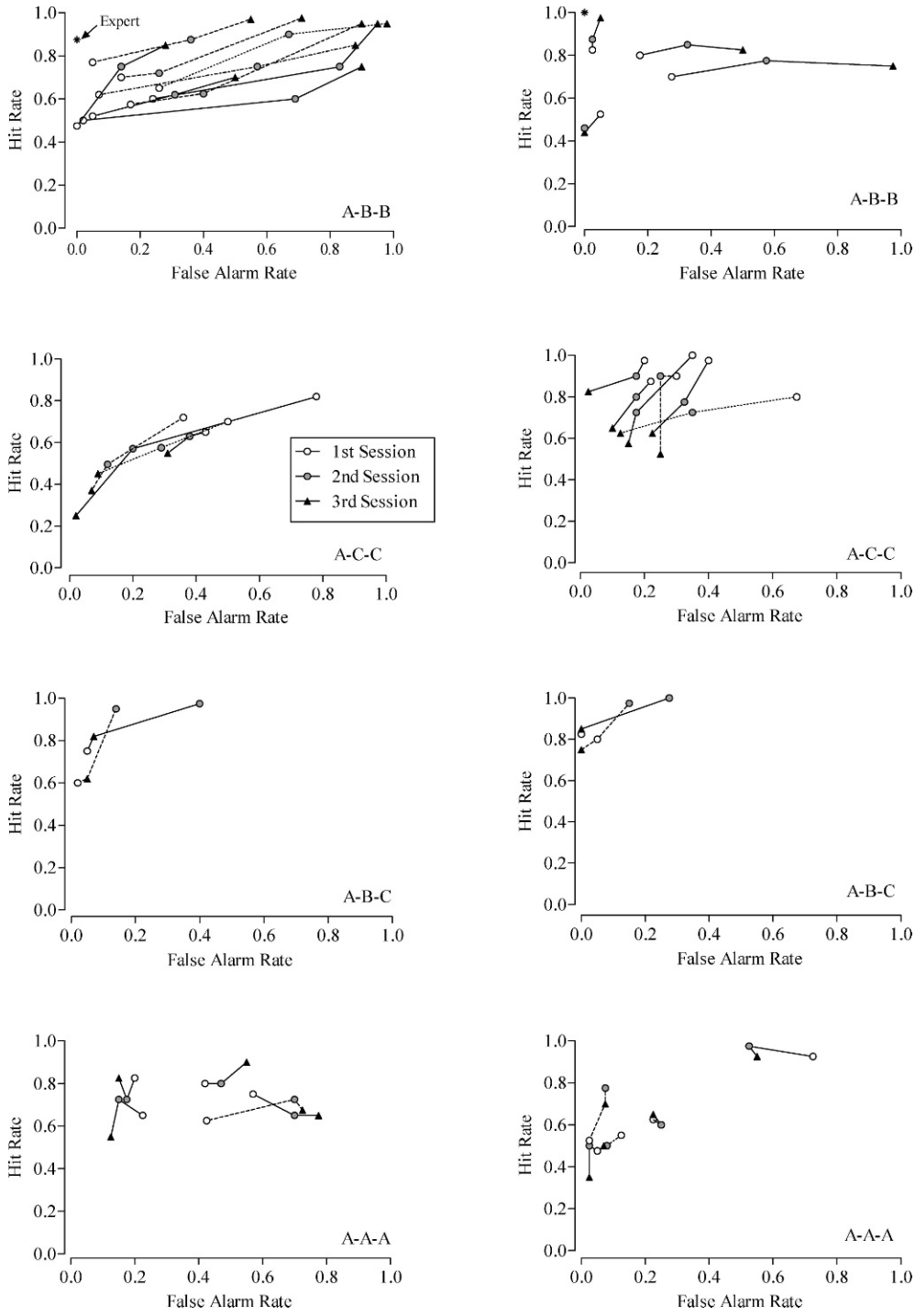

Figure 2. Hit rates and false alarm rates for the participants in Experiment 1 (left) and Experiment 2 (right) across the three scoring sessions. Data are grouped according to the sequence of conditions experienced. Different line patterns are used for some data sets to assist with visual inspection. The asterisks show the expert's hit and false alarm rates in Experiment 1 and Experiment 2.

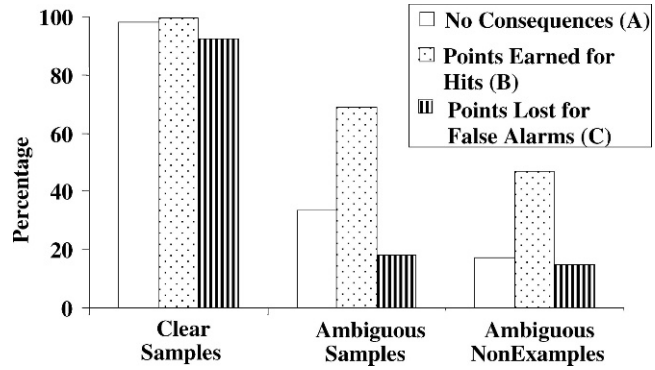

Figure 3. Mean percentage of clear samples of aggression, ambiguous samples of aggression, and ambiguous nonexamples of aggression scored under each condition in Experiment 1.

To summarize, results of Experiment 1 indicated that we could reliably bias observers' scoring of aggression by providing feedback and points when ambiguous stimuli were present. These findings are important because observers are likely to encounter ambiguous events in the environment and because a variety of factors may influence observers' decision rules when scoring these events. Thus, strategies are needed for preventing or reducing this problem among observers. The laboratory stimuli and methods examined in Experiment 1 could be used to further evaluate factors that might alter response bias. Identifying factors that decrease the likelihood of response bias under conditions that have been found to reliably bias responding may lead to strategies for improving the accuracy of data collected in the natural environment.

One such factor is the specificity of the response definition. Numerous authors have suggested that operational definitions should be objective (i.e., refer to only observable features of responding), clear (i.e., provide unambiguous descriptions), and complete (i.e., differentiate between responses that should and should not be considered an occurrence (e.g., Cooper, Heron, & Heward, 2007). Nonetheless, surprisingly few studies have examined the role of this factor on the accuracy and reliability of behavioral observation. In Experiment 1, observers were given a relatively general definition of aggression (i.e., hitting, kicking, or throwing items at the teacher). It should be noted that this level of specificity was selected based on a review of articles on aggression published in the *Journal of Applied Behavior Analysis*. Providing observers with more specific definitions of aggression might result in less flexible criteria for determining whether events should be scored. In other words, by reducing the amount of ambiguity in the stimuli via more precise response definitions, observers may be less susceptible to factors that alter bias. In fact, the conditions found to produce bias in Experiment 1—receiving brief feedback and information about contingencies—may improve accuracy (i.e., discriminability or sensitivity) among observers who have more specific definitions on which to base their criteria.

On the other hand, one assumption of SDT is that sensitivity and bias are independent. Factors that influence sensitivity (e.g., specificity of the behavior definition) should not influence response bias and vice versa. Nonetheless, some research findings have been inconsistent with this assumption. For example, in Alsop and Porritt (2006), pigeons' responding on a discrimination task showed less sensitivity to changes in reinforcement magnitude (i.e., less response bias) as the discrimina-

bility of the stimuli was increased. The interaction between sensitivity and bias was explored in Experiment 2 by giving observers more precise definitions.

Experiment 2

The purpose of Experiment 2 was to determine whether providing observers with more specific operational definitions of aggression would moderate the effects of feedback and points on response bias, as well as lead to higher levels of accuracy than those obtained in Experiment 1.

Procedure

Seventeen participants scored the same three versions of the 33-min video used in Experiment 1. For 12 participants, factors intended to bias responding were manipulated for two of the three scorings. The remaining 5 participants (control group) were not exposed to the experimental manipulation. The data-collection expert also participated in the experiment by again scoring one version of the 33-min video (she had not received any feedback about her previous scoring). All procedures were identical to those described for Experiment 1, with the exception of the definition for aggression. Participants were given the following definition of aggression, which closely described the topographies of hitting, kicking, and throwing that appeared in the video:

> (a) Hitting the teacher, defined as any forceful contact between the student's hand or arm and any part of the teacher; (b) kicking the teacher, defined as any forceful contact between the student's foot or leg and any part of the teacher; and (c) throwing objects at the teacher, defined as any time the student released an object from her hand and it made contact with any part of the teacher.

In addition to this definition, participants were given the following additional instructions (orally and in writing) about the scoring: ''Any response that meets the definition of aggression should be scored, even if the response appears to be accidental. Other behaviors that are aggressive in nature should not be scored if they do not meet the definition of hitting, kicking, or throwing objects at the teacher.''

Results and Discussion

The hit rates and false alarm rates for the individuals in Experiment 2 are displayed in the right column of Figure 2. The data are grouped according to the sequence of conditions experienced by the participant (i.e., A-B-B, A-C-C, A-B-C, or A-A-A). The hit rate and false alarm rate for the session scored by the data-collection expert are plotted in the top right panel (displayed with an x). Of the 12 participants exposed to the experimental manipulation of bias, 7 consistently showed a positive relation between the hit rate and false alarm rate across the three sessions (5 of the 6 participants in the A-C-C group and both participants in the A-B-C group). For these 7 participants, Condition B (monetary points earned for hits) was always associated with an increase in both hits and false alarms relative to the previous condition experienced, and Condition C (monetary points lost for false alarms) was always associated with a decrease in both hits and false alarms relative to the previous condition experienced. For the remaining 5 participants who were exposed to the experimental manipulation, the points and feedback did not bias their responding consistently. Furthermore, none of these participants showed a systematic increase in accuracy, which would have been indicated by an increase in hits or a decrease in false alarms. However, it should be noted that the high levels of accuracy under Condition A for 1 participant in the A-B-B group left little room for improvement. The expert did not miss any instances of aggression, and, as expected, the control participants did not show evidence of bias across the three scoring sessions.

A comparison of the findings from Experiments 1 and 2 suggests that the type of definition received (general vs. specific) influenced the outcomes. Providing a more specific definition appeared to reduce the likelihood of bias among participants in Experiment 2, given

that all of the participants in Experiment 1 showed bias. If so, the variable produced a relatively small effect, however, because more than half of the participants in Experiment 2 still showed evidence of bias. This was somewhat unexpected because the observers had been given definitions and scoring instructions that were more detailed and complete. For example, each form of aggression was explicitly defined, and observers were told to score anything that met this definition even if the response seemed unintentional.

Another feature of the data, however, suggests that the extent of the participants' bias was related to the type of definition received. The amount of change in the hit rate and false alarm rate across the three sessions was typically smaller for participants who received the specific definition than for those who received the general definition. This was particularly noticeable for the false alarm rates. The difference between the highest and lowest hit rate across the three scoring sessions for each participant was averaged across the participants who received the specific definition; a mean difference score of .20 resulted. The same calculation was performed with false alarm rates for participants who received the specific definition and was .23. Mean difference scores among hit rates and mean difference scores among false alarm rates also were calculated for participants who received the general definition (those from Experiment 1) and were .29 and .51, respectively. Two-tailed independent *t* tests were used to compare the mean difference scores for the hit rates (i.e., comparing .20 from Experiment 2 participants and .29 from Experiment 1 participants) and the mean difference scores for the false alarm rates (i.e., comparing .29 from Experiment 2 participants and .51 from Experiment 1 participants). Results were statistically significant for the experimental group differences in the false alarm rate scores, $t(25) = 3.13$, $p = .004$, but results were not statistically significant for the experimental group differences in the hit rate scores, $t(25) = 1.99$, $p = .057$. This suggests that the type of definition moderated the effects of the feedback and points on response bias.

The type of definition also appeared to have a small influence on overall accuracy, particularly the false alarm rate. Overall, the mean hit rate for participants in Experiment 2 who were exposed to the experimental manipulation was slightly higher than the mean hit rate for those in Experiment 1 (*M*s = .8 and .7, respectively). The participants in Experiment 2 also had a lower false alarm rate than those in Experiment 1 (*M*s = .2 and .4, respectively). These differences, however, were not statistically significant, as indicated by the results of two-tailed independent *t* tests for the hit rates, $t(25) = -2.04$, $p = .052$, and false alarm rates, $t(25) = 2.02$, $p = .054$. Although participants in the control group who received the specific definition had a slightly lower hit rate than those who received the general definition (*M*s = .6 and .7, respectively), they had a lower false alarm rate (*M*s = .2 and .4, respectively). Finally, the data-collection expert had a much higher hit rate when given the specific definition (1.0) than when given the general definition (.87). Her false alarm rate remained unchanged (one false alarm).

Together, results of Experiments 1 and 2 suggest that the type of definition provided to observers may alter the likelihood of bias. The consequences manipulated in these experiments, however, were somewhat contrived for the purpose of conducting laboratory research on observer bias. This could potentially limit the generality of these laboratory findings to real-world settings. In the next experiment, the materials and procedures developed in Experiment 1 were used to evaluate factors that may be more likely to occur in research and clinical settings.

Experiment 3

Kazdin (1977) suggested that providing information about possible changes in behavior

during upcoming observations might influence the accuracy of data collection if observers also receive feedback about their recording. In one of the few studies to examine these variables, observers recorded four target behaviors exhibited by 2 children on videotapes (O'Leary, Kent, & Kanowitz, 1975). An experimenter told observers that two of the behaviors were expected to decrease during treatment sessions relative to baseline sessions. In actuality, levels of responding were similar under both conditions. Observers also received social consequences in the form of approval or disapproval for their scoring. Specifically, the experimenter expressed approval (e.g., "Dr. O'Leary will be pleased to see the drop in the level of —"; "These tokens are really having an effect on —.") if the observers' data showed a reduction in responding during treatment sessions relative to baseline sessions. Social disapproval was provided (e.g., "You don't seem to be picking up the treatment effect on —.") if the data showed no reduction or an increase in behavior. Patterns of scoring across the two baseline and two treatment sessions suggested that these factors altered the accuracy of observation.

Nonetheless, it is possible that the social consequences, independent of information about expected behavior change, might have altered the observers' accuracy. It is also unclear whether information about expected behavior change contributed to the outcomes. Furthermore, neither factor has been examined while controlling the ambiguity of the behavioral samples or evaluating the effects on hits versus false alarms.

Thus, the purpose of Experiment 3 was to extend O'Leary et al. (1975) by examining the separate effects of social consequences and information about expected behavior change on response bias using the videotaped samples developed for Experiments 1 and 2. Unlike O'Leary et al., however, the social contingency was based on inaccurate scoring of hits rather than on changes in the amount of behavior scored. This modification was made to simulate more closely the social consequences that might occur when training observers or checking reliability in field settings. That is, a researcher or supervisor who sees evidence of errors might inform observers about the errors (e.g., "It looks like you missed some occurrences of self-injury") and encourage them to score more accurately in the future (among other things). This consequence was similar to that manipulated in Experiments 1 and 2, except that the points were replaced by social consequences only.

Participants and Setting

A total of 32 undergraduate students enrolled in a research and statistics course were recruited for the study. None had previously participated in any experiments conducted by the authors. The students fulfilled a course requirement in their class as a result of their participation. (Students could choose from among a variety of available experiments or select a nonresearch option to satisfy this course requirement.) Of these 32 participants, 8 met the exclusion criteria in the first condition (see description below), resulting in 24 individuals (19 women, 5 men) who completed the study. The mean age of the participants was 30 years (range, 20 years to 53 years).

Materials

Results of the scoring in Experiment 1 were used to create three different videotaped segments from the 33-min video. Each segment contained six clear samples of aggression, six ambiguous samples of aggression, six clear nonexamples of aggression, and six ambiguous nonexamples of aggression. None of the samples appeared in more than one segment. Three different videos were necessary because, for one condition, participants would be told to expect higher levels of problem behavior in the upcoming session. To equate the level of ambiguity across the three video segments, the specific samples for each segment were selected by examining the accuracy of scoring across all

participants and conditions in Experiment 1. Clear and ambiguous samples that were associated with similar levels of accuracy were assigned randomly to Segments 1, 2, and 3. Gold standard data records were established for each segment using the same procedures as those described previously.

Procedure

Participants were escorted to the lab and given the same initial written and verbal instructions as the participants in Experiment 1 except that they were told that each behavioral sample lasted 12 s (rather than 2 min to 5 min) and that each video segment lasted 5 min (instead of 33 min). Practice sessions also were conducted as previously described. All participants were exposed to Condition A (baseline) after completing the practice sessions. Participants then were assigned randomly to either Condition B (social consequences) or Condition C (information about expected behavior change) in the second scoring and to the remaining condition in the third scoring. The three video segments were assigned randomly to each condition across participants. Participants received a brief break between each scoring session.

Baseline (A). Prior to the start of the video segment, participants were told to score as accurately as they could. During the session break, the experimenter compared each participant's data record to the gold standard data record. If the participant scored at least five of the six instances of ambiguous aggression or at least five of the six instances of ambiguous noise (aggression nonexamples), they were excluded from the remainder of the study. These exclusion criteria were used because both of the subsequent conditions were expected to increase the number of hits and false alarms. The 8 participants who met the exclusion criteria were thanked and told that they did not need to complete the remainder of the scorings.

Social consequences (B). Prior to scoring, participants were told, "When downloading your data record, I noticed that there were not as many aggressions scored as there should have been. It is really important that you score all the aggression that occurs and that you try harder on the next video."

Information about expected behavior change (C). Prior to scoring, participants were told, "These clips resemble the actual person's behavior on a day when there was more aggression because the behavior change intervention was not being implemented." No feedback was given to the participants about their scoring in the previous session.

Data Analysis

The total number of hits and false alarms in each condition was calculated for each participant. The number of hits and false alarms in Conditions B and C was then subtracted from the number of hits and false alarms in Condition A. The change in hits and false alarms in each condition was examined for each participant to determine (a) the number of participants who showed evidence of response bias (i.e., an increase in both hits and false alarms), (b) the number of participants who showed an increase in sensitivity (i.e., an increase in hits or a decrease in false alarms), and (c) the number of participants who showed a decrease in sensitivity (i.e., a decrease in hits or an increase in false alarms). These results were examined separately for each participant and for the two groups of participants who were exposed to the two conditions in a different order (B-C vs. C-B). The latter analysis was conducted to identify possible sequence effects.

Results and Discussion

Figure 4 shows the change in the number of hits and false alarms during Conditions B (top) and C (bottom) relative to Condition A. Data are grouped according to the sequence of conditions experienced by the participants. Data for the 12 participants who experienced Condition B prior to Condition C are displayed on the left, and data for the 12 participants who experienced Condition C prior to Condition B

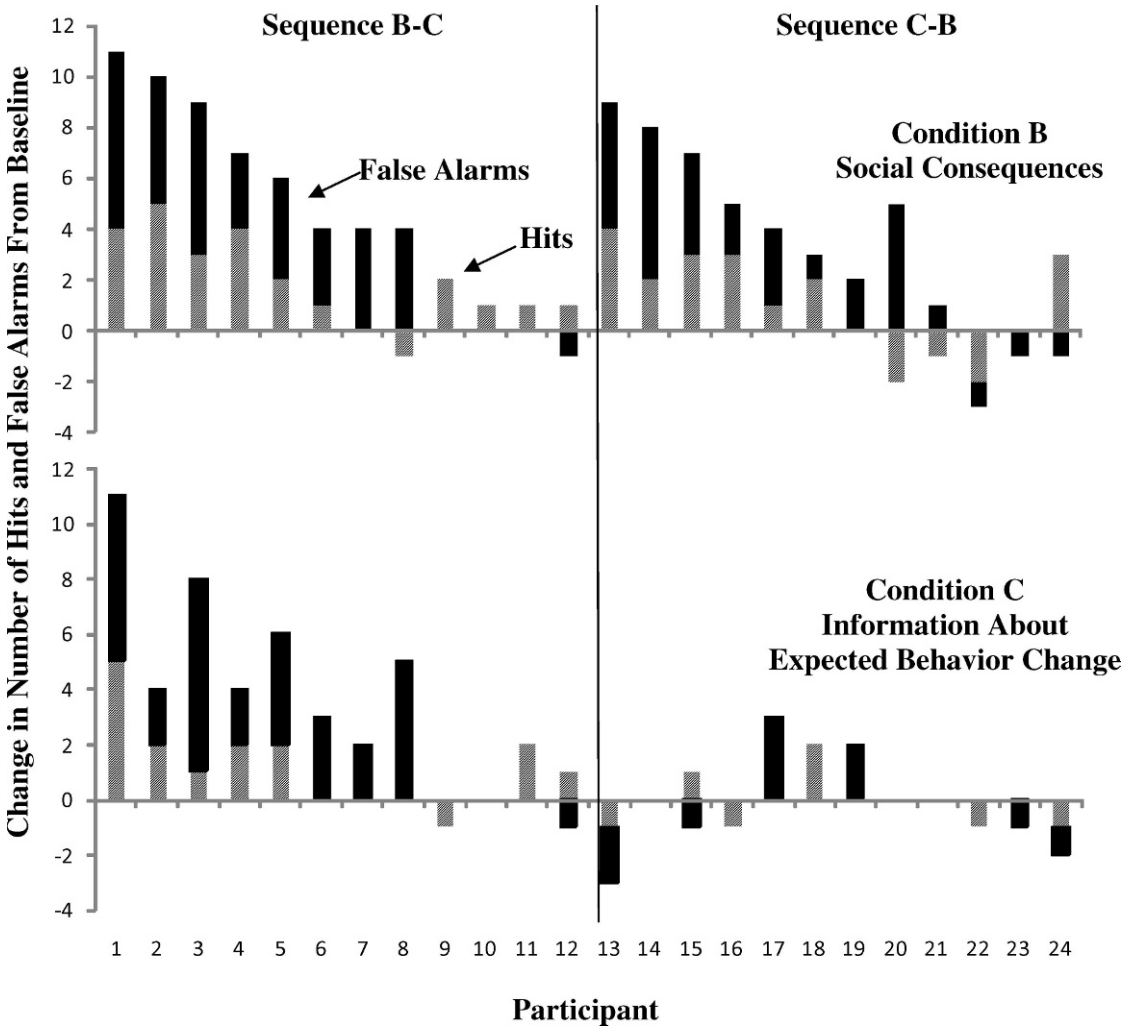

Figure 4. Change in the number of hits and false alarms under Condition B (social consequences) and Condition C (information about expected behavior change) relative to Condition A (baseline) for each participant.

are displayed on the right. Under Condition B, 12 of the 24 participants showed an increase in both hits and false alarms, indicating that social consequences biased responding. One participant showed a decrease in hits and false alarms, an outcome that is consistent with response bias but opposite to the direction expected. An additional 5 participants showed an increase in false alarms only (2 participants) or an increase in false alarms combined with a decrease in hits (3 participants). These patterns suggest a decrease in sensitivity. For the remaining 6 participants, outcomes suggested an increase in sensitivity, with 3 participants showing an increase in hits only, 1 participant showing a decrease in false alarms only, and 2 participants showing a decrease in false alarms combined with an increase in hits. Similar results were obtained regardless of whether the participant was exposed to Condition B before or after Condition C.

When exposed to Condition C, 5 of the 24 participants showed an increase in both hits and false alarms. These 5 participants had been

exposed to Condition B before Condition C and had shown evidence of bias under Condition B. Among the 12 participants who were exposed to Condition C before Condition B, none showed evidence of bias. Three participants showed no change in responding, and 2 participants showed evidence of response bias but opposite to the direction expected (i.e., a decrease in both hits and false alarms). The remaining response patterns (i.e., increase or decrease in sensitivity) were similar to those for participants who had experienced Condition B prior to Condition C.

Together, these results showed that social consequences but not information about expected behavior change altered response bias among some observers who were scoring clear and ambiguous events in a laboratory setting. Sequence or interaction effects also appeared to influence the outcomes for observers who were first exposed to social consequences. That is, the effects of social consequences either carried over into the next scoring session or altered the participants' responses to the experimenter's statement about expected behavior change.

GENERAL DISCUSSION

Preliminary findings support the viability of a procedure based on SDT for evaluating variables that may influence observer accuracy and bias in behavioral assessment. In Experiment 1, individuals who collected data on clear and ambiguous samples of a common target behavior and ambiguous nonexamples of the behavior exhibited predictable patterns of responding. Consistent with previous research on SDT, response bias occurred when observers received brief feedback about their performances and consequences for either hits or false alarms. Changes in scoring were more likely to involve samples designated as ambiguous rather than as clear, providing some support for the designations. The effects of the experimental procedure were robust, in that a consistent relation occurred across participants despite varying initial levels of hits and false alarms. This approach extends that used in previous research on observer accuracy by controlling the nature of the behavioral samples and examining the sources of error obtained (i.e., random vs. nonrandom error).

Results of Experiments 2 and 3 further suggest that this procedure may be useful for advancing our knowledge of factors that influence response bias, especially when some degree of ambiguity is present in the situation. Numerous variables may affect the accuracy of data collected in naturalistic settings, such as the type of instructions received, presence of other observers, clarity of the behavioral definitions, probability of the behavior, and consequences for scoring (or not scoring) events as instances of the targeted response (for reviews, see Kazdin, 1977; Repp et al., 1988). In Experiment 2, providing observers with a more complete and detailed definition of aggression appeared to reduce the likelihood and amount of response bias, as well as decrease the incidence of false alarms. In one of the few previous studies to evaluate this variable, House and House (1979) examined the correlation between the clarity of response definitions and the level of reliability among pairs of observers who collected data on 27 behaviors of children and their parents. One hundred college students read the definitions and ranked them based on clarity. A median clarity rating was then determined for each definition. A Spearman rank-order correlation between these median values and mean reliability scores across the 27 behaviors was not statistically significant (0.25). The authors suggested that all of the definitions might have been reasonably clear due to previous refinements in the definitions. Furthermore, the observers had received extensive training.

Although results of Experiment 2 suggested that the clarity of the definition may influence observer accuracy, the overall impact on response bias was not as substantial as expected.

Nearly half of the participants still showed evidence of response bias. The degree of ambiguity present in the videos or the strength of the biasing factors may have weakened the effects of this variable on responding. Alternatively, the specific definition may not have been as clear as intended. Nonetheless, results suggest that observers may be more resistant to factors that produce response bias when they are given more detailed, specific definitions.

Experiment 3 built on these studies by examining other factors that may bias responding in clinical settings. Instead of combining the general feedback statement with points exchangeable for money, participants were urged to "try harder" after being told that they did not score accurately. This type of social consequence, although still somewhat contrived, may more closely approximate the consequences that occur in clinical situations. Participants' responses to this factor also provided a basis for examining the effects of another variable implicated in previous research on observer accuracy. As noted previously, Kazdin (1977) suggested that information about expected behavior change may influence observer accuracy when combined with feedback and social consequences. Results of Experiment 3, however, showed that the feedback and consequences were responsible for changes in response bias that occurred after observers received information about expected behavior change.

A similar feedback statement was used in Experiments 1 and 3. In both cases, the experimenter told participants that they did not score accurately and that it was important to do so. The participants in Experiment 3, however, were not offered the opportunity to earn points exchangeable for money. This may explain why only half of the participants showed evidence of bias under the social consequences condition in Experiment 3, whereas all of the participants did so in Experiment 1. Other differences in the methods and materials also could explain the discrepant results. It also should be noted that individuals with more extensive training in direct observation techniques may not be as susceptible to bias as the participants in this study. However, as noted previously, behavioral consultants often rely on data collected by caregivers, teachers, and others with limited training and experience.

The purpose of these experimental arrangements was to advance basic and applied research on observer accuracy. A similar procedure may be useful for evaluating other variables that may increase response bias or alter observer sensitivity. These factors include the probability of the behavior, type of instructions or training provided to observers, and the availability of concurrent consequences for the various response options (i.e., hits, correct rejections, misses, and false alarms). Depending on the experimental question and arrangement, it may be helpful to draw on other data analysis methods that are commonly used in SDT and behavioral detection theory research (e.g., receiver operating characteristics, indexes of discriminability and bias; see Irwin & McCarthy, 1998, for an overview of these methods). However, further research is needed to evaluate the utility of these methods for examining observer accuracy and bias in behavioral assessment.

Although the concepts and methods of SDT may prove to be useful to researchers, they have no obvious direct application to the assessment of observer accuracy in clinical settings. Nonetheless, research findings may lead to a greater understanding of observer behavior and, thus, ways to improve the performance of those who collect data in the field. For example, although the monetary points in Experiments 1 and 2 were contrived, it is likely that other types of reinforcement contingencies operate in the natural environment for scoring ambiguous events in one manner or another. Several putative forms of positive and negative rein-

forcement may be available if data gathered by caregivers and staff alter conclusions about the effectiveness of an intervention or the severity of a problem. For example, teacher-collected data that include numerous false alarms and few misses (i.e., inflated levels of problem behavior) might lead to (a) negative reinforcement in the form of removal of the student from the classroom or (b) positive reinforcement in the form of continued assistance and attention from a behavioral consultant. Alternatively, data that contain numerous misses and few false alarms (i.e., deflated levels of problem behavior) might lead to (a) negative reinforcement via removal of the consultant or (b) positive reinforcement in the form of approval and recognition from the consultant, superiors, and peers. Thus, these findings may have some generality to behavioral assessment in the natural environment.

REFERENCES

Alsop, B., & Porritt, M. (2006). Discriminability and sensitivity to reinforcer magnitude in a detection task. *Journal of the Experimental Analysis of Behavior*, *85*, 41–56.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Davison, M., & McCarthy, D. (1987). The interaction of stimulus and reinforcer control in complex temporal discrimination. *Journal of the Experimental Analysis of Behavior*, *48*, 97–116.

DeLeon, I. G., Arnold, K. L., Rodriguez-Catter, V., & Uy, M. L. (2003). Covariation between bizarre and nonbizarre speech as a function of the content of verbal attention. *Journal of Applied Behavior Analysis*, *36*, 101–104.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

House, B. J., & House, A. E. (1979). Frequency, complexity, and clarity as covariates of observer reliability. *Journal of Behavioral Assessment*, *1*, 149–165.

Irwin, R. J., & McCarthy, D. (1998). Psychophysics: Methods and analyses of signal detection. In K. Lattal & M. Perone (Eds.), *Handbook of research methods in human operant behavior* (pp. 291–321). New York: Plenum.

Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: The ABCs of reliability. *Journal of Applied Behavior Analysis*, *10*, 141–150.

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology*, *50*, 215–241.

Mudford, O. C., Martin, N. T., Hui, J. K. Y., & Taylor, S. A. (2009). Assessing observer accuracy in continuous recording of rate and duration: Three algorithms compared. *Journal of Applied Behavior Analysis*, *42*, 527–539.

Nevin, J. A., Olson, K., Mandell, C., & Yarensky, P. (1975). Differential reinforcement and signal detection. *Journal of the Experimental Analysis of Behavior*, *24*, 355–367.

O'Leary, K. D., Kent, R. N., & Kanowitz, J. (1975). Shaping data collection congruent with experimental hypotheses. *Journal of Applied Behavior Analysis*, *8*, 43–51.

Repp, A. C., Nieminen, G. S., Olinger, E., & Brusca, R. (1988). Direct observation: Factors affecting the accuracy of observers. *Exceptional Children*, *55*, 29–36.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*, 1285–1293.

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology diagnostics: Collected papers*. Mawah, NJ: Erlbaum.

Received February 16, 2009
Final acceptance August 21, 2009
Action Editor, Gregory P. Hanley