*The Journal of General Psychology*, 2013, *140*(4), 269–281

# Behavior Description Effect on Accuracy and Reliability

G. DAVID SMITH
*GDS Behavioral Consulting*

JOSEPH V. LAMBERT
ZACHARY MOORE
*The University of the Sciences*

ABSTRACT. The clinical and scientific efficacy of behavioral analysis is dependent upon interveners' accurate and reliable detection and measurement of target behaviors. This study compared the accuracy and reliability of observers' detection and recording of a designated target behavior when different forms of a target behavior description were used. Using an intra-subject design, undergraduate college students were asked to count the number of target behaviors depicted on a videotape under each of two conditions. Conditions differed only to the extent that each contained a different description of the target behavior. Results showed that participants' detection and recording of the target behavior was more accurate and reliable when the target behavior description used a verb (in the present tense, active voice) depicting an action with an observable and discrete beginning and end and omitted modifiers requiring observers to make subjective or relative judgments. Analysis of the data using methods developed by Signal Detection Theory demonstrated the potential utility of this approach for studying observer detection of target behaviors.

Keywords: accuracy, description, gold standard, reliability, Signal Detection Theory, target behavior

THE CLINICAL AND SCIENTIFIC EFFICACY OF BEHAVIOR ANALYSIS is dependent upon accurate and reliable detection and measurement of behavior (Baer, Wolf, & Risley, 1968). Applied behavior analysts (hereinafter referred to as "practitioners") focus their attention on socially significant behaviors ("target" behaviors) and design interventions to modify them. They typically name and describe target behaviors to facilitate detection, measurement, and treatment. In practice, measurement (data collection) and treatment (intervention) are seldom performed by practitioners but by others, referred to as "interveners." Interveners

*Address correspondence to Dr. G. David Smith, GDS Behavioral Consulting, P. O. Box 694, Camp Hill, PA 17001-0694, USA; gdsconsult@verizon.net (e-mail).*

frequently include family members, friends of the client, teachers, and direct-service staff who often lack formal training and experience in behavior analysis.

By naming and describing target behaviors, practitioners aim to influence the behavior of interveners (Baer, Wolf, & Risley, 1987) so that interveners "... capture everything the researcher intended and nothing the researcher did not intend" (Baer, Harrison, Fradenburg, Peterson, & Milla, 2005, p. 443). The target behavior description provides criteria to help the intervener detect a designated target behavior and to respond to it as the practitioner directs. Reliable measurement and faithful application of treatment procedures (Fryling, Wallace, & Yassine, 2012) are critical to the success of behavioral intervention. Both are dependent upon an intervener's accurate detection of target behaviors. Clinical outcomes are jeopardized when interveners do not consistently detect target behaviors and, thus, fail to faithfully deliver prescribed interventions.

With some notable exceptions (Kubina & Yurich, 2012; Kunzelmann, Cohen, Hulten, Martin, & Mingo, 1970; Lindsley, 1964, 1991; McGreevy, 1981; White & Haring, 1980), published reports do not specify a common form or method for describing target behaviors. Despite specifying ways of describing target behaviors, the above authors provided no empirical support for their suggestions. Though it is logical that accurate detection and reliable measurement are dependent upon well-conceived target behavior descriptions, there seems to be no empirical evidence linking characteristics of target behavior descriptions to the accuracy of detection, reliability of measurement or efficacy of treatment.

Skinner (1938) aptly defined behavior as "... what an organism ... is observed by another organism to be doing" (p. 6). Consistent with this definition, Baer, Harrison, and colleagues (2005) make the case that the target behavior is what the code-writer (practitioner) decides is important and nothing else. Accordingly, the "true value" of a target behavior is what the practitioner has observed the client doing. The practitioner's task, then, is to formulate a description of the target behavior so that interveners will respond only to those client behaviors the practitioner has observed and considered important.

Investigators (Miller, 1997) routinely verify the reliability of measurement by computing interobserver agreement (IOA). In applied research a minimum of 80% IOA (Kennedy, 2005) is typically considered an acceptable indicator of reliable measurement. While acknowledging widespread reliance on IOA as a measure of reliability and accuracy, Johnson and Pennypacker (2009) warn against equating IOA with either accuracy or reliability. Rather, these authors assert that IOA enhances "... believability of data ... but provides no information about accuracy or reliability" (p. 148).

Despite these admonitions, neither researchers nor practitioners routinely use direct and independent measures of accuracy (Mudford, Martin, Hui, & Taylor, 2009). Instead, accuracy is typically inferred from high IOA which may be considered a necessary but not sufficient condition for determining accuracy. To address this deficiency, some researchers (e.g., Mudford, Zeleny, Fisher, Klum, & Owen, 2011) have used a "gold standard," method to establish the accuracy of measurement. When using a "gold standard," investigators independently establish the presence of a designated target behavior ("true" value) which is then used as a point of reference for measuring the degree to which these independently defined events are detected by observers. Numerous methods for establishing the "gold standard" have been documented. These include: electromechanical recording (Kapust & Nelson, 1984), predetermined scripted performances with scripts acting as true records (Lerman et al., 2010, Powell, Martindale, Kulp, Martindale, & Bauman, 1977), repeated viewing of video records until consensus is achieved (Boykin & Nelson, 1981, Mudford et al., 2009), and reliance upon expert, highly trained observers (Sanson-Fisher, Poole, & Dunn, 1980; Wolfe, Cone, & Wolfe, 1986). Though useful in research, these methods are, for the most part, unwieldy and difficult to adapt for use in applied clinical settings.

Practitioners commonly report (Vollmer & Sloman, 2008) that, under the pressure of clinical need, they seldom have adequate time to work out precise target behavior descriptions, let alone to verify the accuracy with which interveners detect target behaviors. These authors also point out that practitioners are seldom able to obtain independent observers in addition to designated interveners (usually direct-care staff) to corroborate behavioral observations and verify acceptable levels of IOAs. To address this shortcoming, Lerman et al. (2010) first applied Signal Detection Theory (SDT) to the study of interveners' accuracy in detecting target behaviors. They concluded that their findings support the viability of using SDT to evaluate variables influencing accurate target behavior detection. Signal Detection Theory posits that detection is based on a sensory process and a decision process. Green and Swets (1966) described a general theory of signal detection and refined methods for experimentally testing it. Signal Detection Theory predicts that different forms of a target behavior description are likely to affect the accuracy with which interveners detect target behaviors because descriptions influence the criteria that interveners (observers) use in deciding whether or not a target response is present.

The current study sought to determine whether or not the form and content of target behavior descriptions differentially affect the accuracy and reliability with which interveners detect target behaviors. If target behavior descriptions are found to differentially affect their detection, then SDT analysis can be used to further isolate the characteristics of target behavior descriptions that produce accurate detection. Ultimately this has the potential to improve the efficacy of treatment and increase desirable clinical outcomes.

# Method

## Participants

Eighteen undergraduate, psychology students, at the University of the Sciences in Philadelphia (19 years to 22 years) served as participants. Like most interveners they lacked formal training in behavior analysis. Participants were recruited by means of notices placed throughout campus and an oral invitation to participate delivered by the third author to a large Introduction to Psychology class. Those who volunteered and were available at the times scheduled for the study were selected. No other selection criteria were used. After obtaining informed consent, each participated in two (approximately 30 min.) experimental sessions and was paid $20 at the conclusion of the second session.

## Materials

The same 14 min color videotape of a 12-year old female student at a school for children with Autism was used in all conditions. Permission for use of the student's image in this study was obtained from her parents. The student was the focus of clinical attention because she "hit" her head frequently, sometimes forcefully, creating the risk of physical injury. She "hit" her head with her own hand, with objects she held in her hand, and by bringing her head in contact with fixed objects (e.g. walls, desk, etc.).

Participants viewed the videotape on a desk top computer monitor screen measuring 43cm in diameter. The videotape was divided into 60, 10-s, segments. Fifty percent of segments showed the student hitting her head and 50% did not. The 10-s segments were numbered consecutively from 1 to 60 with a white numeral in the upper right-hand corner of the screen. Each 10-s segment was followed by a 5-s segment during which the screen was black and no images were visible. Participants used a scoring form consisting of a sheet of paper with 60 numbered spaces (each corresponding to a like-numbered 10-s video segment).

## Procedure

Target behavior description was the independent variable in the present study. Two separate descriptions addressing the same target behavior were presented. Participants viewed the same videotape under each of two target behavior description (TBD) conditions. One target behavior description, TBD1, was taken directly from the student's school record: "forcefully swings an open or closed fist in the direction of her head or chin, swings objects in the direction of her head/chin, or swings her head in the direction of a wall or desk with or without making contact."

School behavior analysts had determined that a "swing" was a reliable antecedent of a "hit." The target behavior was defined in this manner to prevent injury to the child during assessment and to allow for an intervention to be delivered early in a chain of responses before the student could perform a potentially injurious "hit."

The second target behavior description, TBD2, described the same behavior, and was composed by the first two authors by using the present tense, active voice of a verb describing movement associated with the socially significant behavior, plus an object naming the receiver of the action. The resulting TBD was: "the child touches any part of her head with her hand or an object she is holding in her hand or moves her head so that it touches an object."

We did not use the word "hit" because of its broad connotations which likely elicit subjective observer judgment. As an alternative we used the verb "touch" because we judged it to be a necessary component of "hitting" and to have fewer connotations. Furthermore, we reasoned, that "touching" is a discrete action with an easily discernible beginning and end. We intentionally omitted qualifiers (e.g., like "forcefully") that did not have readily identified, observable and concrete physical properties.

Participants were instructed to view each 10-s segment and, during the 5-s black screen segment, record the number of times the student performed the designated target behavior in the corresponding section of the score sheet. If they did not observe the student performing the target behavior during the segment, participants were instructed to record a "0" on the data sheet. The dependent variable for this study was the number of responses recorded by each participant during each10-s segment.

All 18 participants viewed the video tape using both TBDs. Nine participants viewed the video tape using TBD1, first, and the other nine participants viewed the video tape using TBD2, first. There was at least a one day interval between the first and second viewing for each participant.

## Gold Standard

Studiocode (Studiocode Version 4; Sportstec, Ltd., Sydney, NSW, Australia, 2006) is video editing and analysis software that allows a user to permanently annotate videotape to keep track of designated events. This software permits precise, time-sensitive counting and recording of videotaped events and allows for viewing in real time or as slowly as needed for precise determination of an event.

A "gold standard" was established for each definition so that the accuracy of participants', detection could be measured. Authors achieved perfect agreement by jointly viewing the designated events (in real time and as slowly as needed) to precisely determine that the event had occurred. This established a "true" value for the frequency of target behaviors for each segment of the video tape. In each case the "gold standard" was the behavior that the researchers intended for the participants to detect, either a "touch" or a "forceful swing." The first two authors produced a permanent record of target behaviors by marking each "forceful swing" of the student's fist, object or head (TBD1) and each "touch" (TBD2) that they agreed was depicted on the videotape.

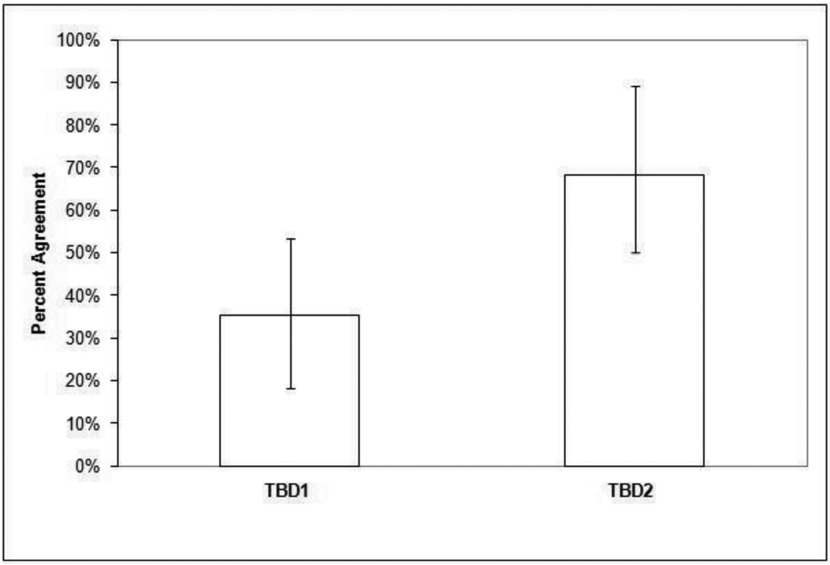

**FIGURE 1. Mean and range of percent agreement with "gold standard" under each TBD condition.**

## Results

Figure 1 shows the mean and range of percent agreement obtained when the frequency of target behaviors recorded by participants under each instructional condition was compared to the "gold standard" on a segment-by-segment basis. This measure reflects the accuracy with which participants detected the presence of a target behavior. Only those segments in which the target behavior was present (as per the "gold standard") or a participant recorded a target response were used to calculate percent agreement. Agreement was scored for segments in which at least one target behavior was present (as per the "gold standard") and participants counted at least one target behavior. Disagreement was scored for segments in which there was no target behavior (as per the "gold standard") and participants counted a target behavior and for segments in which participants failed to count a target behavior when one was present (as per the "gold standard"). Percent agreement was computed by dividing the total number of agreements by the sum of agreements and disagreements for each participant in each condition. The average percent agreement for participants across all segments obtained for TBD1 was 35% (range = 18% to 53%) and 68% (range = 50% to 89%) for TBD2. This

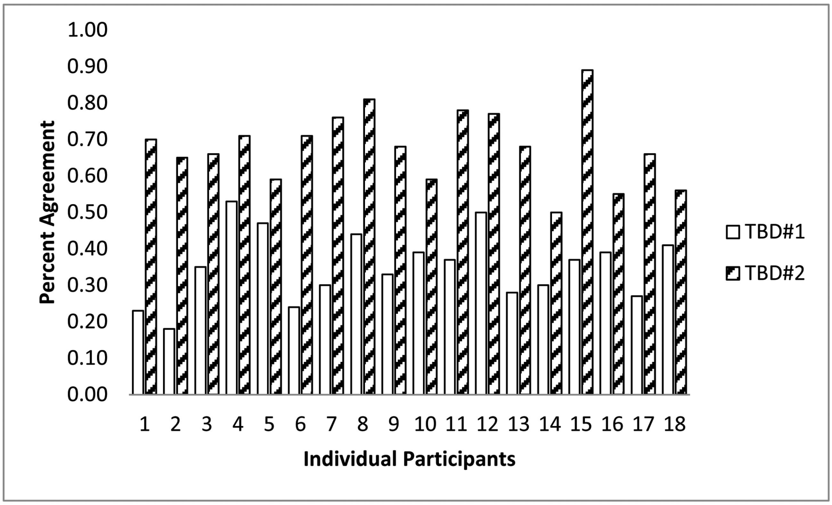

**FIGURE 2. Percent agreement versus the "gold standard" for each participant under each TBD condition.**

difference in average percent agreement was found to be statistically significant, t (17) = 10.6, $p < .001$.

Figure 2 compares percent agreement versus the "gold standard" for each participant under each instructional condition. Eighteen participants (100%) obtained a higher percent agreement when TBD2 was used. Inspection of data revealed that the order of presentation had no effect on percent agreement. Paired t-tests conducted on IOAs versus the gold standard for comparisons between first and second exposures to TBD 1 (t = –.412) and TBD 2 (t = .696) were non-significant.

Table 1 summarizes several measures of variability among participants for each instructional condition. These measures reflect the differences between the

**TABLE 1. Comparison of Variability Measures for Participant Behavior Counts for Each TBD Condition**

| Reliability Measure | TBD1 | TBD2 |
|---|---|---|
| Average range | 2.13 | 1.27 |
| Average standard deviation | .89 | .46 |
| Average variability coefficient | .29 | .14 |

frequencies of target behaviors recorded by each participant on a segment-by-segment basis under each TBD condition without regard to the "gold standard." By comparing the performance of each participant to the others in this manner, the reliability of the measurements they produced was quantified. The first measure, "Average range," depicts the average of the differences between participants' lowest and highest frequency counts for each segment (excluding those segments in which no events were recorded). "Average standard deviation" is the mean of standard deviations computed for each segment (excluding those segments in which no events were recorded). A "variability coefficient" was also computed. This measure was derived by dividing the standard deviation by the mean frequency recorded by participants for each segment. All of these measures indicate less variability for TBD2 compared to TBD1. Taken together, the data presented in Table 1 suggest that participants' observations were less variable and, by inference, more reliable when instructions included TBD2 than when they included TBD1.

Table 2 uses a SDT analysis to compare participant performance for each TBD condition. The SDT paradigm allows each observer's response to be categorized as a hit, correct rejection, miss, or false alarm. For purposes of this analysis, observers' responses were categorized and scored on a segment-by-segment basis. A "hit" was scored each time a participant recorded a target response that was present. A "correct rejection" was scored when a participant recorded a zero frequency and no target behavior was present. When one or more target behaviors were present and the participant failed to record it, a "miss" was scored for each target behavior that was not recorded. Finally, "false alarms" were scored when a participant recorded target behaviors that were not present in the segment.

**TABLE 2. Comparison of Signal Detection Analysis for Each TBD Condition**

| | TBD1 | TBD2 |
|---|---|---|
| Total accurate responses[a] | 75% | 85% |
| Correct rejections[b] | 71% | 82% |
| Hits[c] | 88% | 86% |
| False alarms per target behavior[d] | .7 | .1 |
| Populated segments | .2 | .1 |
| Unpopulated segments | .8 | .2 |

*Note.* [a]Sum of hits and correct rejections divided by the sum of target behaviors present and absent for each participant across all segments; [b]Sum of correct rejections divided by the total number of segments in which no target behavior was present; [c]Sum of hits divided by the total number target behaviors; [d]Sum of false alarms divided by the sum of target behaviors for all segments including those in which a target behavior was (populated) and was not (unpopulated) present.

Table 2 shows that participants' total percentage of accurate responses and percentage of correct rejections were greater when TBD2 was used while "hit" rates were approximately equal for each condition. The percentages of accurate responding differed significantly (t(17) = 4.08, $p < .001$) as did those for correct rejections (t(17) = 3.98, $p < .001$). When TBD1 was used, participants tended to report the presence of a target behavior, even when none was present.

Participants recorded many more false alarms with TBD1 than with TBD2 (t(17) = 12.54, $p < .001$). This difference was significant whether segments contained target behaviors (t(17) = 4.32, p < .001) or did not (t(17) = 11.78, $p < .001$). The false alarm per target behavior ratio was .7—almost one false alarm for every target behavior when TBD1 was used. When TBD2 was used, the false alarm per target behavior ratio was one false alarm for every 10 target behaviors.

## Discussion

Participants more accurately and reliably detected and measured "touching" (TBD2) than they detected and measured "swinging forcefully" (TBD1), thus demonstrating that the behavior of interveners was influenced by characteristics of a target behavior description. This finding is consistent with the SDT prediction that instructions to the observer affect the criteria observers use in their "decision" process (Sawin & Scerbo, 1995). Two measures of accuracy were used—percent agreement and the total percent accurate responding. Both measures documented that participants more accurately detected target behaviors when TBD2 was used. Average percent agreement reflected the extent to which participants detected target behaviors when they were present. Differences in percent agreement under each of the two instructional conditions demonstrated that TBD2 produced more accurate detection and measurement of events that were actually present (as per the gold standard) than did TBD1. This difference was statistically significant and inspection of individual percentage of agreement scores revealed very little overlap in these measures. Percent agreement with the "gold standard" for TBD2 was greater for every participant than percent agreement with the "gold standard" for TBD1. Total percent accurate responding represents a more complete measure of accuracy to the extent that it goes beyond percent agreement by accounting for participants' (correct) non-responding when no target behavior is present. When this measure was used, results showed that, though participants were equally likely to correctly detect a target behavior that was actually present, they differed greatly in the degree to which they tended to detect target behaviors not present and to withhold recording when none was present. Several measures of variability reflected more reliable responding among participants when TBD2 was used.

The clinical implications of these data are significant. Practitioners commonly direct interveners to deliver prescribed consequences (e.g., a reinforcer or a punisher) when they observe that a client has performed a target behavior. Treatment outcome is very sensitive to the accuracy and reliability with which consequent events are delivered (Pipkin, Vollmer, & Sloman, 2010). False alarm rates predict that interveners, using TBD1, would be almost as likely to deliver prescribed consequences for the wrong behaviors as they would for targeted behaviors. The probability that interveners would make this error is considerably less when TBD2 is used, TBD2 therefore produces greater treatment integrity than TBD1 creating the potential for better clinical outcomes when TBD2 is used.

Though differences in accuracy and reliability have been demonstrated, this study does not allow us to identify the variables responsible for these differences. The "gold" standard analysis, however, suggests that the inclusion of the ambiguous adverb "forcefully," may have contributed to greater variability when TBD1 was used. Participants in that condition were required to make subjective judgments that were not required under TBD2. During data collection, when TBD1 was used, experimenters observed that participants tended to count swings resulting in touches even when swings were not "forceful" (as per the gold standard). Indeed, participants counted "forceful" swings in 56% of segments when TBD1 was used even when the gold standard indicated that none were present.

It is also likely that the nature of the verb selected to describe the target behavior affected the accuracy and reliability of measurement. It can be argued that "touches" is an action that has a more discrete and observable, beginning and end when compared to a "swing." Verbs that describe clearly observable movements, distinct environmental characteristics, or readily discernible changes in the position of the participant likely produce higher accuracy and reliability of measurement than verbs that describe ambiguous or complicated actions or infer outcomes (e.g. Mand, request, vocalize, make a sound).

The implications of the present study may be limited because of the use of untrained college students as participants. However, our participants do reflect characteristics of those who often act as interveners in applied settings. Interveners who typically lack the training or experience of the behavior analyst are directed to intervene when a target behavior is detected. Practitioners must designate the target behavior in a way that maximizes the likelihood that these relatively inexperienced and untrained interveners will detect the target response when emitted by the subject and take the prescribed action.

The present study and related analysis incorporate use of a "gold standard" and to some extent our finding are dependent upon the validity of this standard. Gresham (2003) has aptly stated the difficulty associated with establishing a "gold standard" against which to compare observer's recording of behavior. The "gold standard" employed in this study was based upon agreements between the first two authors and was obtained by procedures that were independent of and different from those that produced the data being evaluated. Future research should establish procedures for independently-deriving a "gold standard" which might include embedding predefined target behaviors into videotapes viewed by participants and employing a panel of neutral expert observers in place of the experimenters.

The findings presented here have implications for reliability of measurement, treatment integrity, and treatment efficacy. Treatment efficacy, whether or not a treatment produces a clinically significant and desirable result, is dependent upon the clinician's ability to measure change (to objectively establish "improvement") and the intervener's ability to detect the "target" behavior and to apply the prescribed treatment with integrity. Ultimately a useful description of behavior is one that produces agreement among observers and is consistent with the "true" value designated by the practitioner. The data from the present study suggest that the form and content of a description differentially affect this agreement. Specifically these findings suggest that interveners will more accurately and reliably detect and measure target behaviors when practitioners use descriptions of target behaviors that depict movement with a discrete beginning and end. These results also suggest that modifiers (e.g. adverbs like "forcefully") requiring interveners to make subjective judgments tend to diminish the likelihood of reliable and accurate target behavior detection.

Signal Detection Theory comprises a body of research going back to the late 1940s (Mackworth, 1948) and has significant implications for the study of intervener target behavior detection. It promises to provide a rich source of guidance and a basis for future research (Pastore & Scheirer, 1974). By categorizing responses as hits, correct rejections, misses, and false alarms, SDT offers a framework that facilitates precision in the description of interveners' target behavior detection. More work is needed to adapt the SDT experimental paradigm in which a the "signal" (target behavior) is superimposed on "noise" (non-target behaviors).

## AUTHOR NOTES

**G. David Smith** is a psychologist and certified behavior analyst providing training, consultation and direct services. **Joseph V. Lambert** is Professor of Psychology and has held teaching positions at Temple University, LaSalle University and St. Joseph's University, all in Philadelphia, PA and at De Sales University in Center Valley, PA. **Zachary Moore** holds a BS in Psychology from, and is pursuing a MS in Health Psychology, at the University of the Sciences.

## REFERENCES

Baer, D. M., Harrison, R., Fradenburg, L., Peterson, D., & Milla, S. (2005). Some pragmatics in the valid and reliable recording of directly observed behavior. *Research on Social Work Practice*, *15*, 440–451. http://dx.doi.org/DOI: 10.1177/1049731505279127

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, *1*, 91–97.

Baer, D. M., Wolf, M. M., & Risley, T. R. (1987). Some still-current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, *20*, 313–327.

Boykin, R. A., & Nelson, R. O. (1981). The effect of instructions and calculation procedures on observers' accuracy, agreement, and calculation correctness. *Journal of Applied Behavior Analysis*, *14*, 479–489.

Fryling, M. J., Wallace, M. D., & Yassine, J. N. (2012). Impact of treatment integrity on intervention effectiveness. *Journal of Applied Behavior Analysis*, *45*, 449–453.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Reprint ed.). Los Altos Hills, CA: Peninsula Publishing.

Gresham, F. M. (2003). Establishing the technical adequacy of functional behavioral assessment: Conceptual and measurement challenges. *Behavioral Disorders*, 282–298.

Johnson, J. M., & Pennypacker, H. S. (2009). *Strategies and tactics of behavioral research* (3rd ed.). New York, NY: Routledge.

Kapust, J. A., & Nelso, R. O. (1984). Effects of the rate and spatial separation of target behaviors on observer accuracy and interobserver agreement. *Behavioral Assessment*, *6*, 253–262.

Kennedy, C. H. (2005). *Single-case designs for educational research*. New York, NY: Allyn Bacon.

Kubina, R. M., & Yurich, K. K. (2012). *The precision teaching book*. Lemont, PA: Greatness Achieved Publishing Company.

Kunzelmann, H. P., Cohen, M. A., Hulten, W. J., Martin, G. L., & Mingo, A. R. (1970). *Precision teaching*. Seattle, WA: Special Child Publications, Inc.

Lerman, D. C., Tetreault, A., Hovanetz, A., Bellaci, E., Miller, J., Karp, H., . . . Toupard, A. (2010). Applying signal-detection theory to the study of observer accuracy and bias in behavior assessment. *Journal of Applied Behavior Analysis*, *43*, 195–213.

Lindsley, O. R. (1964). Direct measurement and prosthesis of retarded behavior. *Journal of Education*, *147*, 62–81.

Lindsley, O. R. (1991). From technological jargon to plain English for application. *Journal of applied behavior analysis*, *24*, 449–458.

Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology*, *1*, 6–21. http://dx.doi.org/10.1080/17470214808416738

McGreevy, P. (1981). *Teaching and learning in plain English* (2nd ed.). Kansas City, MO: Plain English Publications.

Miller, L. K. (1997). *Principles of everyday behavior analysis* (3rd ed.). Pacific Grove, CA: Brookes/Cole.

Mudford, O. C., Martin, N. T., Hui, J. K., & 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.

Mudford, O. C., Zeleny, J. R., Fisher, W. W., Klum, M. E., & Owen, T. M. (2011). Calibration of observational measurement of rate of responding. *Journal of Applied Behavior Analysis*, *44*, 571–586.

Pastore, R. E., & Scheirer, C. J. (1974). Signal detection theory: Consideration for general application. *Psychological Bulletin*, *81*, 945–958.

Pipkin, C. S., Vollmer, T. R., & Sloman, K. N. (2010). Effects of treatment integrity failures during differential reinforcement of alternative behavior: A translational model. *Journal of Applied Behavior Analysis*, *43*, 47–70.

Powell, J., Martindale, B., Kulp, S., Martindale, A., & Bauman, R. (1977). Taking a closer look: Time sampling and measurement error. *Journal of Applied Behavior Analysis*, *10*, 325–332.

Sanson-Fisher, R. W., Poole, A. D., & Dunn, J. (1980). An empirical method for determining an appropriate interval length for recording behavior. *Journal of Applied Behavior Analysis*, *13*, 493–500.

Sawin, D. A., & Scerbo, M. W. (1995). The effects of instrcution type and boredom proneness in vigilance: Implications for boredom and workload. *Human Factors*, *37*, 752–765.

Skinner, B. F. (1938). *The behavior of organism*. New York, NY: Appleton-Century-Crofts, Inc.

Studiocode (Version 4). [Computer software]. (2006). Sydney, New South Wales, Australia: Sportstec, Ltd, DBA Studiocode Business Group.

Vollmer, T. R., & Sloman, K. N. (2008). Practical implications of data reliability and treatment integrity monitoring. *Behavior Analysis and Practice*, *1*, 4–11.

White, O. R., & Haring, N. G. (1980). *Exceptional teaching* (2nd ed.). Columbus, OH: Charles E. Merrill.

Wolfe, V. V., Cone, J. D., & Wolfe, D. A. (1986). Social and solipsistic observer training: Effects on agreement with a criterion. *Journal of Psychopathology and Behavioral Assessment*, *8*, 211–226.

*Original manuscript received December 26, 2012*
*Final version accepted June 19, 2013*