Educ Asse Eval Acc (2013) 25:281–302
DOI 10.1007/s11092-013-9166-x

Development and testing of a direct observation code training protocol for elementary aged students with attention deficit/hyperactivity disorder

Naomi J. Steiner · Tahnee Sidhu · Kirsten Rene · Kathryn Tomasetti · Elizabeth Frenette · Robert T. Brennan

Received: 29 August 2012 / Accepted: 11 April 2013 /
Published online: 7 May 2013

Abstract Observational measures can add objective data to both research and clinical evaluations of children's behavior in the classroom. However, they pose challenges for training and attaining high levels of interrater reliability between observers. The Behavioral Observation of Students in Schools (BOSS) is a commonly used school-based observation instrument that is well adapted to measure symptoms of attention deficit/hyperactivity disorder (ADHD) in the classroom setting. Reliable use of the BOSS for clinical or research purposes requires training to reach reliable standards (kappa≥0.80). The current study conducted training observations in one suburban and one urban elementary school in the Greater Boston area. To enhance interrater reliability and reduce training time, supplemental guidelines, including 30 additional rules to follow, were developed over two consecutive school years. The complete protocol was then used for training in the third school year. To reach sufficient interrater reliability (kappa≥0.80) during training, 45 training observations were required in the first year while, in the third year, only 17 observations were required. High interrater reliability was sustained after training across all three school years, accumulating a total of 1,001 post-training observations. It is estimated that clinicians or researchers following this proposed protocol, who are naive to the BOSS, will require approximately 30 training observations to reach proficient reliability. We believe this protocol will make the BOSS more accessible for clinical and research usage, and the procedures used to obtain high interrater reliability using the BOSS are broadly applicable to a variety of observational measures.

N. J. Steiner (✉) · T. Sidhu · K. Rene · K. Tomasetti · E. Frenette
The Floating Hospital for Children at Tufts Medical Center, Boston, MA, USA
e-mail: nsteiner@tuftsmedicalcenter.org

R. T. Brennan
Harvard School of Public Health, Cambridge, MA, USA

Keywords Direct observation codes · Behavioral Observation of Students in Schools (BOSS) · Attention deficit/hyperactivity disorder (ADHD) · Classroom observation · Behavior observation

1 Introduction

Students with academic or behavioral concerns in school are assessed using a wide range of instruments such as behavior rating scales, semi-structured interviews, and direct observations. Direct observation methods are demanding in that they require trained field researchers to carry them out. Training methods are not always well documented, and techniques for obtaining high levels of interrater reliability, a critical characteristic of any observational technique, may be elusive. In this paper, the research team presents our 3-year Behavioral Observation of Students in Schools© (BOSS) (Shapiro 2011) training and administration data that supports the enhanced protocol described here. The BOSS is a direct systematic observation tool used by researchers and clinicians that assesses both on-task and off-task student behaviors in the classroom environment through an interval recording format system; it can also be used to assess children with attention deficit/hyperactivity disorder (ADHD) (Vile-Junod et al. 2006). The BOSS is being utilized more frequently among researchers and clinicians; however, data and information to guide training and reliable usage of the tool is lacking in the literature. Thus, we demonstrate methods for training field researchers to administer an observational tool, such as the BOSS, working toward a high level of interrater agreement using the robust kappa statistic. Our methods include close study of formal definitions, coding of practice video and in-class observations, rapid quantitative analysis of observation data through a programmed Interrater Agreement Calculator, and immediate post-observation discussion.

In general, direct observations are used for a number of assessment tasks such as screening for emotional and behavioral problems and monitoring interventions, thus serving as an objective measure for both educators and researchers (Riley-Tilman et al. 2005). Direct observation of children in their classroom is one of the most common assessment strategies used by school psychologists (Shapiro and Heick 2004; Wilson and Reschly 1996) as it allows for measurement of a behavior as it occurs naturally, without bias of a parent or teacher. Direct observation assessments can be separated into two categories: naturalistic observations with anecdotal descriptions or systematic observations with interval coding. Naturalistic observations do not allow for standardization or psychometric testing of the assessment tool (Hintze and Matthews 2004; Hintze et al. 2008). In contrast, systematic observations use an interval recording format and provide quantitative data, which can help reduce observer bias. Furthermore, the interval recording format is useful for systematic training of observers using a detailed protocol and checking for reliability between observers.

This article exclusively addresses the training of systematic direct observations of children in their classroom. Although systematic observations use an interval recording format, observer bias can still occur and affect the accuracy and objectivity of a direct observation. Observer bias refers to the tendency of an observer to consistently view and record observed behaviors in a particular way (either negatively or

positively), possibly because the observer: has been trained to categorize behaviors in such a way, has developed this habit over time, or even has a personal bias toward the child being observed (Merrell 1999). A bias in student observations from a clinician could lead to faulty recommendations. For instance, if an observer perceives a child to be very distracted and frequently off-task, that child could be recommended for unneeded preferred seating. A biased researcher could lead to systematic errors and faulty research conclusions.

Students with ADHD are frequently observed in the classroom setting for multiple reasons including diagnostic assessments or monitoring of a behavioral plan (Volpe et al. 2005). ADHD is defined by the current Diagnostic and Statistical Manual (American Psychiatric Association 2000) as a disorder in which an individual suffers from inattention and/or hyperactivity–impulsivity that is more acute and recurrent than typical for the individual's stage of development. The lifetime prevalence of ADHD for children and adolescents is estimated at 9 % (Merikangas et al. 2010).

High rates of off-task, disruptive, noisy, and rule-violating behavior displayed by children with ADHD often occur in school settings. A study by Vile-Junod et al. (2006) found that children with ADHD from ages 6 to 10 years exhibited significantly decreased rates of academic engagement than their comparison peers when directly observed. More specifically, children with ADHD are observed to display higher rates of: physical and verbal aggression, seeking attention, out-of-chair time, and noncompliance (Vile-Junod et al. 2006).

The serious nature of ADHD symptoms in the classroom and their negative effects on academic progress (Mautone et al. 2005) indicates the need of a precise assessment of student behavior to allow for the development of accurate treatment plans. The accepted battery of assessment tools for diagnostic or follow-up evaluations of ADHD includes behavior rating scales, semi-structured interviews, and direct observations (Atkins and Pelham 1991; DuPaul et al. 1992; DuPaul and Stoner 1994; Montague et al. 1994). Assessments from a subjective party, such as a parent or teacher, may be biased in terms of the types of behaviors they report. A study by Hartman et al. (2007) found that parents can be more biased than teachers when rating ADHD symptoms. Teachers generally have access to many normal students with whom they can compare a target student. On the other hand, parents commonly have only their own child(ren) with whom to compare the target child. Furthermore, ADHD symptoms may be more apparent to teachers as the school environment requires more sustained attention, greater independent completion of tasks, and higher activity and impulse control than the home environment (Hartman et al. 2007).

The BOSS is one example of an objective observation system for coding classroom behavior (Shapiro 2011) and has been found reliable between observers (Volpe et al. 2005). In addition, the BOSS has also been shown to differentiate between children with ADHD and their typically developing peers (i.e., children who were reported by their teachers to be average in terms of classroom behavior and academic achievement). Thus, the BOSS is well adapted to measure ADHD symptoms in the classroom setting.

However, even when using objective assessments, errors can still occur. Merrell (1999) reported six categories of coding errors that can threaten the validity of observation data:

1. Poorly defined behavior categories: occurs when categories are not defined specifically enough, which can lead to observer hesitation and mistakes. For instance, if there are no written guidelines in the training manual for children reading and tapping their pencils at the same time, an observer may not know whether this behavior should be categorized as on-task or off-task.
2. Low interrater reliability: occurs when observers code a behavior differently, or may document the behavior in a different interval (a few seconds apart), leading to coding differences.
3. Observee reactivity: occurs when students notice they are being observed and they alter their behavior.
4. Situational specificity of target behaviors: occurs when a certain behavior only occurs during particular events (e.g., if a child only interrupts the teacher during math, then an observation during reading might not reveal a full behavioral profile).
5. Inappropriate code selection: occurs when an observer miscodes an observed behavior.
6. Observer bias: occurs when an observer continually sees and records observed behaviors in a certain way (e.g., overly positive or overly negative). In other words, an observer's expectation of a specific behavior to occur may result in a biased observation.

We believe coding errors in these six categories can be reduced through a precise training protocol. Such a protocol is also imperative to track and reach high levels of interrater reliability during training. Requiring high levels of interrater reliability assures that the data collected from a direct observation does not depend on who conducted the observation but on the predefined standardized coding scheme. Interrater reliability can be calculated as the agreement between two observers using two calculation methods: percent agreement and kappa score.

Percent agreement is calculated as:

$$\text{Percent Agreement} = \frac{\#\text{of agreements}}{\#\text{of agreements} + \#\text{of disagreements}} * 100$$

A percent agreement of 100 % signifies perfect agreement, whereas a percent agreement of 0 % signifies that the observers did not agree on any of the observations. In general, a percent agreement of 80 % represents reasonable reliability (Miles and Huberman 1994). For example, in a study done by Mautone et al. (2005), a video tape of a classroom was used to train observers on the BOSS until they reached a percent agreement of 80 % in each domain. After the training, percent agreement was calculated for 19.5 % of the observation sessions with a maintained mean percent agreement of 94.8 % and a range between 86.7 % and 99.0 %.

While percent agreement is an accepted calculation for interrater reliability, a more robust calculation is the kappa score (Hintze 2005). The kappa score is considered to be more robust because it also considers and corrects for the random chance that two observers would code the same behavior by accident. For instance, two observers could have a percent agreement score of 98.5 % but a kappa score of 0.00 because a behavior happened during one interval of the 15-min observation and they did not agree on that one observation interval. Therefore, it is more difficult to reach a high kappa score than it is to have a high percent agreement. Hintze (2005) states a kappa score is calculated as:

$$K = \frac{P_o - P_c}{1 - P_c}$$

where P_o is the ratio of agreement between observers for occurrences and nonoccurrences, and P_c is the ratio of expected agreements based on chance. P_c is calculated as:

$$P_c = \frac{\left(\begin{array}{c}\#\text{Occurrences} \\ \text{for Observer1}\end{array}\right)\left(\begin{array}{c}\#\text{Occurrences} \\ \text{for Observer 2}\end{array}\right) + \left(\begin{array}{c}\#\text{Nonoccurrences} \\ \text{for Observer 1}\end{array}\right)\left(\begin{array}{c}\#\text{Nonoccurrences} \\ \text{for Observer2}\end{array}\right)}{(\#\text{Intervals})^2}$$

Kappa scores can range from −1.00 to +1.00. A large positive kappa indicates that observers agree more commonly then would be expected by chance, thus indicating high interrater reliability (Hintze 2005). A kappa of 1.00 signifies perfect agreement, whereas a kappa of 0.00 signifies agreement equivalent to chance, and a kappa less than 0.00, which happens very rarely, signifies agreement less than expected by chance. In general, a kappa from 0.21 to 0.40 is considered fair, 0.41 to 0.60 moderate, 0.60 to 0.80 substantial, and finally 0.81 and above as almost perfect (Gelfand and Hartmann 1975; Landis and Koch 1977). The BOSS has been found to yield high kappa scores during training as compared with other direct observation measures. DuPaul et al. (2004) achieved kappa scores on the BOSS between 0.93 and 0.98 in a comparative study of children with ADHD to those without (N=136).

The BOSS developer (Shapiro 2011) reports training for proficiency takes 10 to 15 h, which is shorter when compared with other measures such as the Classroom Observation Code, which takes 50 h of training to develop proficiency (Volpe et al. 2005). Although a short manual has been published, a systematic stepwise training protocol is not currently available. The BOSS was used in the current study to observe students with ADHD in the context of a school-based intervention; therefore, the observations were in the same context as that of many school psychologists evaluating children with ADHD. However, it was found that training by using the BOSS Manual (Shapiro 2011) with video and in-class practice observations alone did not enable the research team to reach high interrater reliability on training during the first school year. Supplemental specifications in addition to the BOSS Manual behavior criteria were needed in order to avoid coding errors associated with inadequate training and to achieve high reliability indices. Therefore, Supplemental Guidelines for training were created at the beginning of the first and second school year, followed by the use of this completed BOSS training protocol for training in the beginning of the third (last) school year. This paper describes an in-depth, standardized, and user-friendly BOSS training protocol that has been developed by this research team to serve as a training guide, complementing the published BOSS Manual (Shapiro 2011), for clinicians and educational researchers. For the current research group, the use of this protocol led to high interrater reliability scores (percent agreement and kappa scores; see "Results" section) and successful standardization for the continued use of the BOSS.

2 Methods

2.1 Participants

A total of six observers were trained to administer the BOSS at the beginning of three consecutive school years, hereafter known as Round One, Round Two, and Round Three. Training included two steps: a practice video and further practice in a classroom setting. During training Round One, two school psychology graduate students who had already received formal training on the BOSS through review of the BOSS Manual (Shapiro 2011) and conducting practice video observations in a course instructed by Robert Volpe became observers for our study. A third observer, chosen to be the lead observer for logistical reasons, completed an in-depth literature review of the BOSS and its training in Round One, and became the BOSS trainer. Through discussions with other observers about repeated coding discrepancies due to ambiguous guidelines, the lead observer compiled the Supplemental Guidelines and trained further observers for Round Two and Round Three. All of the subsequent trainee observers were naive to the BOSS method; however, all observers had prior experience working with children. Some observers had past experience working with children with ADHD, while others did not. Practice training observations were conducted in randomly selected elementary school classrooms in one suburban (Round One) and one urban (Round Two and Round Three) school in the Greater Boston area. Observers were well-trained in efforts to avoid bias. They randomly selected target students to observe; this means they were not informed whether children they were observing had been diagnosed with ADHD or any other disorder. Furthermore, they were instructed to avoid selecting students from any particular demographic group (e.g., race or ethnicity). The development of this enhanced, more accessible, and clarified BOSS training protocol was carried out over 3 years for a larger study.

2.2 Materials

This enhanced BOSS training protocol simplifies the flow of the training and includes the use of the following materials:

2.2.1 Manual for the Behavioral Observation of Students in Schools (BOSS)

The BOSS Manual (Shapiro 2011) offers a solid foundation for BOSS training. It provides the rationale for using the observational tool and a description of the five behavioral categories for observation (Table 1) and the teacher and peer coding (Table 2). The BOSS defines classroom engagement as the desired behavior, represented by on-task active (AET) and on-task passive (PET) behaviors. Impulsivity and hyperactivity are measured by off-task motor (OFT-M) and verbal (OFT-V) behaviors, while inattention is quantified by the frequency of off-task passive (OFT-P) behaviors (Table 1). In addition, the BOSS Manual provides instruction in the completion of the BOSS Observation Form (Shapiro 2011) and the development of a coding interval audiotape. The BOSS Observation Form assesses the number of on-task and off-task behaviors a child exhibits over a 15-min observation period. The observations can be done for up to 30 min; however, most published papers report the

Table 1 Description of BOSS behavior categories and classroom settings categories (Shapiro and Heick 2004)

Category	Description
On-task (momentary time sampling)	
Active engaged time (AET)	Student is actively attending to assigned work, e.g., writing, reading aloud, or raising hand.
Passive engaged time (PET)	Student is passively attending to assigned work, e.g., listening to a lecture, looking at an academic worksheet, or silently reading assigned material.
Off-task (partial-interval method)	
Off-task motor (OFT-M)	Any instance of motor activity that are not directly associated with the assigned academic task, e.g., engaging in any out-of-seat behavior (defined as buttocks not in contact with the seat), aimlessly flipping the pages of a book, or manipulating objects not related to the academic task (e.g., playing with a paper clip, throwing paper, twirling a pencil, folding paper).
Off-task verbal (OFT-V)	Any audible verbalizations that are not permitted and/or are not related to an assigned academic task, e.g., making any audible sound, such as whistling, humming, forced burping, talking to another student about issues unrelated to an assigned academic task, or talking to another student about an assigned academic task when such talking is prohibited by the teacher.
Off-task passive (OFT-P)	Times when a student is passively not attending to an assigned academic activity for a period of at least three consecutive seconds. Includes when a student quietly waits after the completion of an assigned task, but is not engaged in an activity authorized by the teacher, e.g., sitting quietly in an unassigned activity, looking around the room, or staring out the window.
Classroom settings	
Independent seatwork: Teacher present (ISW:TPsnt)	Target child is doing work by him/herself. The teacher is either doing work individually at a desk or rotating around the classroom.
Independent seatwork: Teacher small group (ISW:TSmGp)	Target child is doing work by him/herself. The teacher is working with children in a small group (eight children or fewer).
Small group: Teacher present (SmGp:TPsnt)	Target child is doing work in a small group (eight children or fewer). The teacher or an assistant teacher may be working with the group or just be present in the classroom.
Large group: Teacher present (LgGp:TPsnt)	Target child is doing work in a large group (i.e., whole classroom or a group with nine or more children). The teacher is present in the room instructing the class.

use of a 15-min observation period (Volpe et al. 2005). To our knowledge, there are no convergent validity studies of the BOSS with another measure. However, there are data supporting the ability of the BOSS to discriminate between children with ADHD and typically developing children (DuPaul et al. 2004).

2.2.2 BOSS Observation Form

The BOSS Observation Form (Shapiro 2011) is included in the BOSS Manual (Shapiro 2011) and is used to record the occurrence of the five behavior categories

Table 2 BOSS coding instructions (Shapiro and Heick 2004)

Aspect of the BOSS	Description
Observation time	Either 15 (60 intervals) or 30 (120 intervals)-min periods.
Peer coding	Code for a peer in the classroom every fifth interval (e.g., interval 5, 10, 15) using the behavior coding used for the target child (see Table 1). An observer can either pick three peers to alternate between ahead of time or go in order around the classroom observing peers.
Teacher-directed instruction (TDI)	Code for TDI every fifth interval (same interval classroom peer is coded). Coded for if the teacher, or an assistant teacher, is giving academic instruction. This can be in a large group setting, small group setting, or one-on-one.

(Table 1). Before an observation begins, the observer fills out all observation information at the top of the BOSS Observation Form, including participant identification information, date, time, observer identification information, academic subject (math or language arts), and classroom setting, such as students working independently (Table 2). There are 60 intervals over the 15-min observation, and each interval contains five boxes, corresponding to the five behavior categories, to record on-task and off-task behaviors. Once the observation begins, when an on-task or off-task behavior is observed, a check is placed in the designated box for that interval. Both engaged behaviors (AET and PET) are scored at the beginning of each 15-s interval; this is referred to as momentary time sampling. During the remainder of each interval, off-task behaviors are recorded if the event occurs within that interval; this is referred to as partial-interval method. At every fifth interval, teacher-directed instruction and classroom peer behaviors are coded instead of the target child's behaviors (Table 2). Amendments were made to this BOSS Observation Form to include additional helpful information (see "Procedure").

2.2.3 Practice video

Robert Volpe provided a DVD with a 15-min training video of seven students and one teacher in an urban classroom setting. The video was used to train observers concerning the development of: familiarity with the five behavior categories, fluency on the BOSS Observation Form, and familiarity with the Supplemental Guidelines. The video shows one teacher providing direct instruction for the entire class and time for independent seatwork.

2.2.4 Coding interval audio

Robert Volpe provided the coding interval audio, developed following the directions in the BOSS Manual (Shapiro 2011), which is an mp3 file that prompts observers every 15 s through an ear piece by calling out the number of the observation ("Observation 1," "Observation 2," etc.), so that observers do not to have to simultaneously look at a watch. This allows observers the ability to focus on classroom activities and coding. The audio is also required to synchronize coding for a session with multiple observers.

2.2.5 Supplemental guidelines

These guidelines (Table 3) were developed to elucidate particular unclear or vague behaviors that were noted to cause repeated coding mistakes. They increase the standardization of the BOSS training and therefore increase interrater reliability and decrease length of training. These guidelines are broken down into general information guidelines and guidelines for each category of the BOSS.

2.2.6 Interrater Agreement Calculator

We developed the Interrater Agreement Calculator in Microsoft Excel to use in the training and fidelity procedure. After completing an observation, observers enter data from the BOSS Observation Forms into the Excel file, which takes less than 1 min. This serves two purposes. First, a formula in the Excel file highlights coding boxes that differ between observers, allowing observers to easily identify and discuss discrepancies. This is an essential step in reducing repeated coding errors and improving scores on the next observation. The second purpose is to calculate both percent agreement and kappa scores in three domains: total engaged behavior (AET and PET), off-task motor/verbal behavior (OFT-M and OFT-V), and off-task passive behavior (OFT-P). (Interrater Agreement Calculator Excel file is available upon request with corresponding author).

2.3 Procedure

2.3.1 Development of new standardized material

An amended BOSS Observation Form, as well as both the Supplemental Guidelines and the Interrater Agreement Calculator, were developed during training Rounds One and Two by the research team. The completed materials were used in training Round Three. The following materials are explained below:

1. Amended BOSS Observation Form: The scoring section at the bottom of the observation form was removed, and its' function was replaced by the Interrater Agreement Calculator. Four sections that were found helpful when conducting an observation were then added. First, three additional sections were added at the top of the observation form to include (1) the name of the target student's school, (2) the name of the target student's teacher, and (3) the observation number of each target student, which is necessary if each student is observed more than once at a specific time point (e.g., if it was the third observation on the same child, a '3' was written in the space). Lastly, (4) a 'notes' section was added to the bottom of the amended BOSS Observation Form (in place of the scoring section) for space to record activities unrelated to the five behavior categories (e.g., if the child goes to the bathroom, the observation interval the child left the classroom, and how long the child was gone for could be recorded). Furthermore, this space can be used to record changes in the academic subject or classroom setting (see Table 2) and at which interval this took place.

Table 3 Supplemental guidelines for the BOSS

1) General information	
•Record the observation number if observing the same target student more than once.	
•Record if there is a change of setting (e.g., large group teacher present to student in independent seatwork).	
•Record if there is a change in academic subject.	
•Only observe academic time, not circle time or a game, etc.	
•Moment observation: Begin observation after coding audio finishes saying observation number (i.e., at the end of 14 for observation 14).	
•Only mark TDI when teacher is giving academic instruction; do not mark TDI when teacher is disciplining.	
•All teachers in the classroom can be included when coding TDI (e.g., assistant teacher, student teacher), however one- to-one teacher aids should not be included.	
2) On-task behavior	
Active engaged time (AET)	•Erasing answers to a question (not the desk or something they doodled, etc.) is AET.
	•If teacher asks a question and child nods head in response during momentary scoring, child should be marked as AET.
	•If child is writing and has a *very* momentary pause, or is going from erasing to writing, this is AET.
Passive engaged time (PET)	•If child has a worksheet in front of them, and teacher is giving instructions, then child can look at either teacher or worksheet and be scored as PET. If teacher gives specific instructions to "look at me" or "look at what I am pointing to on the worksheet," do not mark child as PET if still looking at worksheet and not teacher.
	•Following non-academic instruction (i.e., passing papers, walking back to desk) is considered PET.
	•Flipping a page while reading a book is PET (unless child is aimlessly flipping pages, then it is OFT-M).
	•If a child is standing while doing work this is PET and not OFT-M *unless* the teacher tells the child to sit down, then it is OFT-M.
	•If child could be looking at paper, give them the benefit of the doubt and mark as PET.
	•If child ignores a peer who is talking to him/her when he/she is not supposed to be, do not penalize the target child, and mark as PET.
3) Off-task behavior	
General off-task	•Hand-raising behavior:
	◦If child raises hand in response to teacher asking for class participation (e.g., to give an answer or ask a question), child is AET and *not* off-task if hand is raised for an extended period of time (can be longer than three seconds as long as teacher is still accepting answers/questions).
	◦If child raises hand unprompted (to ask a question, etc.) for longer than 3 s, this is considered OFT-P.
	◦If child raises hand unprompted (to ask a question, etc.) and keeps waving hand for longer than 3 s, this is considered OFT-M.
	•If child is working or listening to the teacher and fidgeting at the same time, do not count as off-task.
Motor (OFT-M)	•If child is paying attention/working and then starts being OFT-M (e.g., playing with shirt) and looks at the shirt while playing with it for at least 3 s then mark the child as OFT-M.
	•If child deliberately turns around (not a quick glance behind them), this is considered OFT-M.
	•Sharpening a pencil is not on-task or off-task unless child is repeatedly sharpening pencil for no reason. Then it is considered OFT-M.

Table 3 (continued)

	•When child is on-task, then becomes OFT-M (i.e., stands up), they have to be OFT-M for at least 3 s to be counted as OFT-M. However, if child is already off-task and then is OFT-M (i.e., talking and then stands up), it does not have to be for 3 s to be counted as OFT-M.
Verbal (OFT-V)	•IF a child is talking to a peer *and* playing with something at the same time, then both OFT-M and OFT-V should be checked off.
Passive (OFT-P)	•If child is making a repetitive motion (e.g., bouncing legs, playing with hair, biting nails, etc.) while staring off then child is both OFT-P and OFT-M.
	•If teacher gives specific instructions such as "stop reading and put away your books" and child continues to read, mark child as OFT-P.
	•If the observer can tell the student is reading the wrong page or book, this is considered OFT-P.
	•If target child looks at another student's work when they are not supposed to, it is not off-task unless they do it for more than 3 s, then it is considered OFT-P.
	•Watching the teacher discipline or talk with another child is considered OFT-P.

2. Supplemental Guidelines: During training Round One, observers had difficulty achieving a 0.80 kappa score, due to repeated coding mistakes, when following the BOSS Manual (Shapiro 2011). A review of BOSS Observation Forms was conducted to analyze where repeated errors were the most frequent. This review showed that off-task behaviors were the least specific and the hardest to code accurately. Based on these findings, a list of 30 additional guidelines pertaining to general information as well as the five behavioral categories was compiled (Table 3) to remedy these recurrent unclear areas. These additional guidelines decreased training time in Round Two, where some further clarifications were also added. In Round Three, however, no further changes were required.
3. Interrater Agreement Calculator: The Interrater Agreement Calculator was created in Microsoft Excel during Round One to compare observations between observers and to calculate percent agreement and kappa scores. It takes approximately 1 min to enter each observation. (See section on "Interrater Agreement Calculator").

2.3.2 Observations and post-observation discussions

During training, two or more observers participate simultaneously for all training observations (double observations), where the lead observer is always present (for seating arrangements, see "In-class practice observations"). Reliability scores are calculated between the lead observer and each of the trainee observers. Double observations are used throughout research and clinical observations to ensure continued reliability. Each training observation includes the 15-min observation along with a 15-min post-observation discussion, which is just as important for training as the observation itself. After the development of new standardized materials, the four observation procedures below were followed in Round Three:

1. Practice video observations: The goals of the practice video observations are to become familiar with the flow of the BOSS observation, obtain fluency with the BOSS guidelines and the five behavioral categories of the BOSS Observation Form, and follow the coding interval audio rhythm and the pace of the coding, which can initially seem rushed. This procedure decreases the total number of in-class training observations needed, which can be a burden to the school system, classrooms, and teachers. Before beginning the practice video observations, observers review the BOSS Manual (Shapiro 2011) and the Supplemental Guidelines. They then learn how to record information on the BOSS Observation Form and how to code the five behavioral categories. Observers select a specific target student in the video to observe and select the peer students to be observed every fifth interval (Table 2). For each subsequent practice video observation, the same video is used, yet a different target student and different peers are observed each time. While watching the video and listening to the coding interval audio, observers fill out the BOSS Observation Form. As stated above, observers should separately enter the BOSS Observation Form into the Interrater Agreement Calculator and review which observations they have coded differently. It is possible to review discrepant intervals of the video during the post-observation discussion, which can be helpful in the beginning. Particular attention to these discussions, which take an average of 15 min, will ultimately lead to an increased understanding of the coding categories and to higher kappa scores on subsequent observations. This process continues until a kappa score of around 0.70 is reached on all three kappa calculations, when competency and fluency in filling out the BOSS Observation Form is achieved. Observers then proceed with practice in-class observations.
2. Practice in-class observations: The goals of the practice in-class observations are to precisely record behaviors in a variety of classroom activities and to obtain a kappa score of 0.80 or higher. After observers reached kappa scores of 0.70 on the practice video observations, a nearby school was contacted and asked if observers could complete practice observations in their classroom. Since observing in schools may be seen as a burden to some teachers or school officials, if a specific teacher is willing to have multiple observations in his or her classroom, the observers should opt to train in a classroom that is amenable to their presence and select a different target student for each observation. In this case, the diversity of settings and grades for practice observations comes secondary to the observations going smoothly, as discussing the coding discrepancies between observers is the most beneficial piece of training, regardless of the classroom setting. These in-class observations give observers the opportunity to apply the knowledge and experience they gained through coding and discussing the practice video to a live classroom setting, also giving observers the opportunity to refine their skills and become familiar with what a real classroom environment is like. Upon entering the classroom, observers select a student at random for their target child and pick the best vantage point possible, which is an extremely important step. It is ideal for observers to select a position that allows a full view for all observers of the child's face, arms, legs, and eyes and is not blocked by the teacher moving around. Additionally, a position not directly in front of the target child is recommended, so that the student is not aware of being observed, even

subtly. When there are three observers, the lead observer sits between the other two, to obtain a vantage point that is most similar to each of the other two observers. Observers then fill out the identifying information on the BOSS Observation Form, recording the child's gender and color of clothing in place of child ID. Peer selection is important in order to ensure that the observers code for the same peer child for the peer comparison even if the children get up and move during the observation. Thus, observers typically choose three easily visible peers and rotate between them in the same order for every fifth observation. Observers then start the observation, listening to the coding interval audio on an mp3 device with an earpiece in only one ear, enabling observers to listen to both the audio and to what is happening in the classroom. During observations with two observers, each observer listens to one of the two earpieces to ensure that they are both coding in synch. An adapter that allows two sets of headphones to be plugged into the same device can be used for three observers. As part of the amended BOSS Observation Form described above, if the classroom setting changes within the 15-min period, the observation number at which the setting changed, and the setting the classroom was changed to is recorded (e.g., large group teacher present to independent seatwork teacher present at observation 15). Similarly, if the academic subject changes, the observation interval at which it changed, and the academic subject the classroom changed to is recorded. Furthermore, if there is a long transition time or if the target student leaves the classroom, the audio is paused, and observers record this in the notes section of the BOSS Observation Form. The observation continues once the classroom is back to academic work or the student returns to the classroom. Observers should be aware that, unlike for the practice video observations, there is no possibility to "go back" during the real life observations. Observers should write down comments regarding possible discrepancies next to the questionable box making it easier to review and discuss. Again, the discussion that follows is the cornerstone toward improvement. Repeated in-class training observations are continued until observers reach a kappa score of 0.80 on the Interrater Agreement Calculator in all three of the kappa calculations (engagement, OFT-MV, and OFT-P). Back-to-back observations may lead to observation fatigue and a dip in kappa scores, therefore training observations should contain at most two or three practices in a row. After the 0.80 kappa scores are reached, observers are considered fully trained and ready to conduct in-school observations for clinical or research purposes.

3. Clinical/research observations: Double observations (see section on "Observations and post-observation discussions") should be conducted and sustained throughout clinical or research observations. Double observations were carried out for one third of the total research observations (N=1,001) conducted over the 3 years. These double observations are the ones referred to in the "Results" section. For double observations, the process of selecting peer comparisons follows that of the double training observations. In single observations, peer selection rotates from the target student to the next seated student in the classroom at every fifth observation. Again, due to observation fatigue (see "In-class practice observations") consecutive observations should be limited to two, or at most three, consecutive observations. In order to observe a student in a variety of settings, each target child is observed three times at each time point in the study.

4. Refresher training observations: In order to account for skill decay, refresher training assures continued high fidelity if there is a lapse in time between conducting observations, such as school vacation or during standardized state competency testing. It is important for already trained observers who may conduct observations only every 3–4 months, to review the training protocol to assure that high kappa scores are maintained.

3 Results

3.1 Practice video observations

In order for trainee observers to become familiar with the BOSS, practice video observations were used for all three training rounds. A large decrease in sessions required to reach reliability over time was noted, with 20 training sessions in Round One, decreasing to 13 sessions in Round Two, and decreasing even further to 4 sessions in Round Three (Fig. 1). At the conclusion of the video observations, kappa scores reached above 0.70 in engaged behavior, off-task motor/verbal behavior, and off-task passive behavior, with kappa scores of 0.86, 0.93, and 0.93 respectively in Round Two, and 0.94, 0.79, 0.78 respectively in Round Three. (No kappa scores were available for the practice video observations in Round One.)

3.2 In-class practice observations

In order to refine observation skills and become familiar with a live classroom setting, in-class observations were used for all three training rounds. In-class observations were used for training 25 times in Round One, decreasing to 11 times in Round Two, and showing stability with 13 times in Round Three (Fig. 1). At the conclusion of the in-class observations, kappa scores in engaged behavior, off-task motor/verbal behavior, and off-task passive behavior reached 0.83, 0.95, and 0.79 in Round One, increasing to 0.92, 0.86, and 1.00 in Round Two, and again increasing to 1.00, 1.00, and 1.00 in Round Three, respectively.

3.3 Clinical/research observations

Observers accumulated 1,001 total observations of students in elementary schools using the BOSS. Of these 1,001 total observations, 701 were single observations, and 300 were double observations, (see, "Observations and post-observation discussions"). During these double observations, the lead observer's data were used to track reliability. On these 300 observations, percent agreement was found to be greater than 97.50 % in all three domains. Sustained high kappa scores were found in all three domains, with a score of 0.92 for engaged behavior, 0.92 for off-task motor/verbal behavior, and 0.92 for off-task passive behavior (Table 4). Kappa scores were high and did not differ for urban and suburban settings, classroom subject (math/science and language arts/social studies), classroom setting (teacher lead large

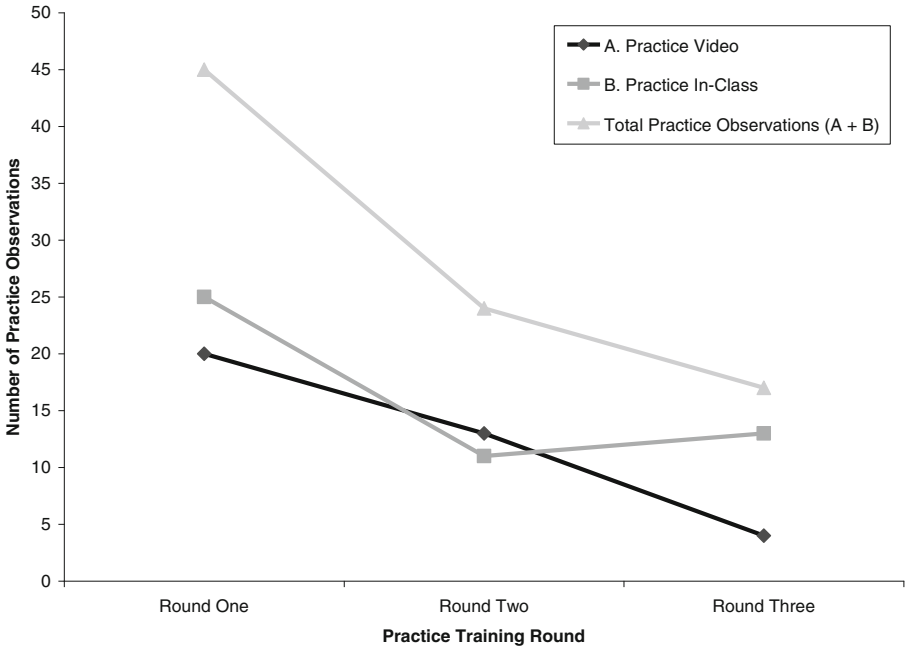

Fig. 1 Practice training observation counts required to meet kappa reliability scores over three training rounds

group, small group, and independent seatwork), and grade level (second through sixth) (Table 5).

3.4 Refresher training observations

At the start of refresher training, kappa scores had dropped significantly, to as low as 0.41 in Round One (Table 6). Video observations were used for refresher training an estimated 10 times in Round One, 3 times in Round Two, and 6 times in Round Three. In-class observations were used for refresher training 14 times in Round One, 4 times in Round Two, and 5 times in Round Three. Therefore, combined refresher training yielded 24 training sessions in Round One, 7 training sessions in Round Two, and 11 training sessions in Round Three in order to reach kappa scores comparable to the baseline training. Kappa scores for in-class refresher training for engaged behavior,

Table 4 Average percent agreement and kappa scores for the research observations (N=300)

BOSS category	Percent agreement (SD)	Kappa score (SD)
Active/passive engaged time	97.60 (2.90)	0.92 (0.11)
Off-task motor/verbal	97.70 (3.30)	0.92 (0.12)
Off-task passive	99.20 (2.00)	0.92 (0.19)

Table 5 Average percent agreement and kappa scores across BOSS categories by geographic location, academic subject, classroom setting, and grade for the research observations

		N	Percent agreement (SD)	Kappa score (SD)
Geographic location	Urban	96	98.00 (1.87)	0.95 (0.08)
	Suburban	204	98.60 (3.03)	0.91 (0.16)
Academic subject	Math/science	121	98.00 (2.83)	0.91 (0.15)
	LA/SS	168	98.30 (2.67)	0.93 (0.13)
Classroom setting	Large group	165	98.00 (2.80)	0.93 (0.12)
	Small group	49	98.50 (2.60)	0.91 (0.17)
	Independent seatwork	114	98.50 (2.60)	0.92 (0.15)
Grade	2nd	63	98.00 (2.40)	0.91 (0.15)
	3rd	28	98.90 (1.97)	0.94 (0.13)
	4th	140	95.50 (3.23)	0.93 (0.15)
	5th	67	98.90 (1.60)	0.95 (0.10)
	6th	2	98.30 (2.40)	0.95 (0.07)

Total does not add up to *N*=300 because of mixed academic subjects or classroom settings within one observation. LA=language arts; SS=social studies

off-task motor/verbal behavior, and off-task passive behavior reached 0.89, 0.84, and 1.00 in Round One, 0.97, 0.97, and 1.00 in Round Two, and 0.95, 0.93, and 1.00 in Round Three, respectively.

4 Discussion

Direct classroom observation measures are valuable tools for objective behavior assessment that can be used clinically in schools or as a research tool. The BOSS is an ideal direct observation measure for observing students with ADHD, whether it is

Table 6 Kappa scores at the start of the refresher training observations

Training round	Training method	Kappa score at start of refresher training observations		
		Engaged	OFT-MV	OFT-P
1	Video	x	x	x
	In-class	0.72	0.69	0.41
2	Video	0.80	0.66	0.65
	In-class	0.72	0.62	0.73
3	Video	0.84	0.66	0.85
	In-class	0.65	0.73	0.66

x represents data not available, *OFT-MV* represents student is either OFT-M or OFT-V at any interval

for research or clinical practice. Objective data enables an evaluator to discuss observational findings with the teaching team and parents, and it lends credence to an individual's recommendations, especially in the face of potentially conflicting opinions. This study used an iterative process that included repeated reviews of our BOSS data to optimize the described protocol, which will enable researchers and clinicians to conduct standardized reliable observations. We also adopted the more robust kappa statistic in place of the percent agreement score to achieve this. Thus, the purpose of this protocol is to assure accessibility and enable efficient training with high reliability for research and clinical settings.

In general, the number of observations required for training to reach the desired 0.80 kappa score decreased each subsequent round as the lead observer became more adept at teaching the BOSS and with the development of the training process described here. In Round One when the observers learned and trained on the BOSS, 45 total practice sessions were required, including both video and in-class practice. In Round Two, the total number of practice sessions decreased to 24, and, finally, in Round Three the total number of practice sessions required to reach the reliability threshold was just 17. Both the enhanced protocol and the lead observer's increased familiarity with the BOSS may explain this decrease in training time.

4.1 Implications for research and clinical practice

Researchers are frequently confronted with interrater reliability issues. Classroom observations are complex to code and are an example of where training has to reach high reliability so that data obtained by multiple observers can be reliably used. This challenge can become even greater if the research period is prolonged which can lead to weakening of the learned training methods. One implication of this study is our suggestion to conduct refresher training, which we found successful in maintaining high interrater reliability over time. In addition, there is frequent turnover of RAs within research teams. Another implication of this study extends to this research situation by suggesting the possibility of a replacement RA to follow this optimized training protocol, which leads to high interrater reliability making trained observers interchangeable.

Our scenario of a lead observer with an in-depth knowledge of the BOSS, training observers over three training rounds, mirrors that of a school psychologist training interns at the beginning of each semester or school year. The school psychologist could act as the lead observer, training interns after learning the BOSS him-or herself. The school psychologist would conduct the training and the trainees' reliability scores would be calculated by comparing observations with the school psychologist. Moreover, in addition to using the BOSS regularly, the school psychologist should double observe on the former trainee's BOSS observations at arbitrary points during the year to ensure that the observers are still following the protocol and high interrater reliability is continuously attained. An added benefit of reaching high interrater reliability on the BOSS is that observers are therefore considered equivalent, so the same observer does not necessarily need to observe the same student over time if multiple assessments are required (e.g., measuring response to intervention) (Lord et al. 2000).

Based on the experience and data of this research group, if a group that is completely naive to the BOSS is following this described protocol, it is estimated

that, similar to Round One, they will require around 15 video practice sessions and 15–20 in-class practice observations to yield reliable observations. If one person is familiar with and has trained using this BOSS protocol before, replicating a trained lead observer, it is estimated that they will need around 10 video and 10 in-class observations to train other trainee observers, mirroring Round Two. It is estimated that a school psychologist who has been trained in and is familiar with classroom observations using systematic direct observation forms will require training that is somewhat in between these numbers to become proficiently trained on the BOSS specifically using this protocol. For a school psychologist who trains interns on a yearly basis, it is anticipated that the amount of BOSS training time will decrease gradually.

4.2 Recommendations

4.2.1 Training observations

We recommend the use of the kappa statistic over simple percent agreement because the latter does not account for agreements that might occur by chance alone. We suggest training until an estimated kappa of at least 0.80 is achieved. Although kappa requires more calculation than simple agreement, developing a simple Interrater Reliability Calculator can mainstream and speed-up the process, specifically for a research team conducting training with multiple observers. The Interrater Reliability Calculator is also essential in the post-observation discussion as it shows intervals of agreement and disagreement, and also calculates the interrater reliability through both percent agreement and kappa scores. Otherwise, one would have to manually compare BOSS Observation Forms to determine when observers agree and disagree on coding a behavior for each of the five behavior categories. Then, once the number of agreements and disagreements were counted, the percent agreement and kappa score would have to be manually calculated. In comparison, data can be entered into the Interrater Reliability Calculator in under 1 min, and kappa scores are automatically calculated. Moreover, discrepant boxes between observers are automatically highlighted, allowing for quick comparisons after an observation and expedited discussion of these disagreements.

To decrease the number of training observations, it is essential to do in-depth post-observation discussions, which take around 15 min. For review of the training video sessions, observers have the ability to go back and watch the video for a specific interval and come to a concrete decision for how a behavior is supposed to be coded. This was an important aspect of achieving a kappa of at least 0.70 on the videos before moving on to training observations in a live classroom. With a live in-class observation, re-viewing an interval is not possible. To overcome this, observers tagged the questionable or hesitant coding box with a brief note, which facilitated later discussion during the post-observation review.

Throughout both training and clinical or research BOSS observations, it is essential to constantly prevent observers from possible observation fatigue. The current research group found that observers were less precise after three sequential 15-min observations. During training, observation fatigue can interfere with reaching reliability thresholds, and during observations for evaluation, fatigue can lead to invalid

observations. It is strongly suggested, for research and clinical purposes, that a maximum of three observations in a row is used to reduce the risk of possible observation fatigue, which would lead to a decrease in reliability.

During double observations, observers should position their observation forms out of view from the other observer and make a concerted effort not to look at each other's observation form or interact in any manner during the observation. However, on an estimated 1.00 % of the intervals, observers did confer about the child's behavior at that moment if their line of sight to the child was blocked for one or both observers. This was considered acceptable, as the observer could not see the child.

Achieving a similar viewpoint for observing can be challenging if there are more than two observers training at a time. The current research team used three observers, including the lead observer, for each training round. This can be both more disruptive for the classroom and make it more difficult for the observers to have the same viewpoint. For instance, the observers on the right and in the middle might see a child's foot tapping, while staring off for over 3 s and mark the child as OFT-P and OFT-M, while the third observer on the left might not be able to see the child's foot or lower body. As a result, this third observer would only mark the child as OFT-P and the reliability between observers would be off due to the observers' viewpoints, not their knowledge on how to code the BOSS. It is therefore recommended that observers be extremely careful when choosing viewpoints and make sure to pause the audio and adjust seating arrangements when necessary.

We found that age affects BOSS coding. Younger children are harder to observe and code for as they are off-task more and often do not have their behavior redirected as much by the teacher. In fact, this in-depth protocol, which further specifies behaviors to standardize BOSS training, might only be appropriate for elementary school-aged children. It is strongly encouraged that BOSS observers following this proposed protocol practice training observations in classrooms of all ages they plan on observing.

4.2.2 Research and clinical observations

In addition to our training recommendations, we also have specific suggestions for research and clinical observations. We recommend observing more subjects than only math and language arts, as this provides more opportunities for observers to observe each child, therefore making it easier to schedule multiple children on the same day. For post-training research observations, we dichotomized academic subjects into math and science or language arts and social studies, due to the similar nature of the classes. We also observed each child three times in order to take into consideration variables such as academic subject, as well as type of instruction and time of day, which we also strongly encourage BOSS observers follow.

Double observations were continued in order to assure maintained high kappa scores. The data from the lead observer were only used for reliability purposes. Observers should take into account the subject (e.g., math/science or language arts/social studies) and type of instruction (large group, small group, or independent seatwork) when making observations. These variables will be

taken into account to interpret the findings, as a child's behavior may change in various settings. While these variables may affect the interpretation of the observation they are unrelated to observer reliability. Once an observer has completed BOSS training and has started clinical or research observations, it would be beneficial to observe a child at different times during the day and in various classroom settings in order to get a more complete profile of the child's behavior.

4.3 Limitations

Limitations of this BOSS training protocol include the focus on children with ADHD. Although the target child in the in-class training observations may or may not have ADHD, high reliability was still maintained throughout the study where all observed children had a diagnosis of ADHD. If a child does not have ADHD, their behaviors may be easier to code for, as they may not exhibit as many off-task behaviors as a child with ADHD would. This only increases the importance of thorough training so that observers know the BOSS guidelines in detail and can detect and code for on-task and off-task behaviors quickly and accurately.

Furthermore, the BOSS training protocol could be enhanced with a formal standardized answer key that corresponds to a practice video. An answer key would allow trainees to see exactly how a BOSS developer would code a specific behavior. This would ensure that observers would be trained exactly the same across time, as there would only be the one correct code for each interval.

4.4 Conclusion

The critical finding of this direct observation training protocol is that interrater reliability increased or stayed at similar levels across Rounds 2 and 3 compared with Round 1, and the number of training observations required to reach high interrater reliability (kappa$\geq$0.80) greatly decreased. Furthermore, it reflects well on the proposed training protocol that research observations showed continued high kappa scores after training, refresher training, and throughout the research observations. It was surprising to see a decrease in kappa scores after only a 4-month break in conducting observations, which underlines the importance of the short refresher training. These reasons illustrate how this enhanced BOSS protocol makes training more accessible, shorter, and leads to higher reliability.

The procedures we used can be applied to a variety of observational measures. Initial training using recorded performances allows for critical review so that the trainer can correct errors among the trainees and reinforce the definitions of the behaviors. Although movement to a live setting may preclude the ability to review video (in some settings with some measures, it may be possible to record behavior as well as observe it live), the transition to live observation with two observers is essential to developing the observational skills further in a "real-world" setting. We recommend using the kappa statistic to assure robust interrater reliability. It is critical that comparison of coding through analysis of observation data and post-observation discussion take place as rapidly as possible. Finally, if there are gaps between time

periods of observation, refresher training should take place to maintain desired levels of interrater agreement.

Acknowledgments This research was financially supported by the Institute of Education Sciences (Project: Computer Attention Training in Schools for Children with Attention Deficit/Hyperactivity Disorder (ADHD), Project No: R305A090100). We would like to thank the Newton and Boston Public Schools and acknowledge R Chris Sheldrick for his support in the development of the Interrater Agreement Calculator and the following graduate students for their hard work and support of this research study: Jessica Bennett and Jessica Chen.

References

American Psychiatric Association. (2000). *Diagnostic and Statistical Manual of Mental Disorders-Text revision (DSM-IV-TR)* (4th ed.). Washington, DC: American Psychiatric Association.

Atkins, M. S., & Pelham, W. E. (1991). School-based assessment of attention deficit-hyperactivity disorder. *Journal of Learning Disabilities, 24*, 197–205. doi:10.1177/002221949102400403.

DuPaul, G. J., & Stoner, G. (1994). *ADHD in the schools: Assessment and intervention strategies*. New York: Guilford Press.

DuPaul, G. J., Anastopoulos, A. D., Shelton, T. L., Guevremont, D. C., & Metevia, L. (1992). Multimethod assessment of attention deficit hyperactivity disorder: The diagnostic utility of clinic-based tests. *Journal of Clinical Child Psychology (now called Journal of Clinical Child and Adolescent Psychology), 21*(4), 394–402. doi:10.1207/s15374424jccp2104_10.

DuPaul, G. J., Volpe, R. J., Jitendra, A. K., Lutz, J. G., Lorah, K. S., & Gruber, R. (2004). Elementary school students with AD/HD: Predictors for academia achievement. *Journal of School Psychology, 42*, 285–301.

Gelfand, D. M., & Hartmann, D. P. (1975). *Child behavior analysis and therapy*. New York: Pergamon Press.

Hartman, C. A., Rhee, S. H., Wilcutt, E. G., & Pennington, B. F. (2007). Modeling rater disagreement for ADHD: Are parents or teachers biased? *Journal of Abnormal Child Psychology, 35*, 536–542. doi:10.1007/s10802-007-9110-y.

Hintze, J. M. (2005). Psychometrics of direct observation. *School Psychology Review, 34*, 507–519.

Hintze, J. M., & Matthews, W. J. (2004). The generalizability of systematic direct observations across time and setting: A preliminary investigation of the psychometrics of behavioral assessment. *School Psychology Review, 33*, 258–270.

Hintze, J. M., Volpe, R. J., & Shapiro, E. S. (2008). Best practices in systematic direct observation of student behavior. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (Vol. 2, pp. 319–335). Bethesda, MD: National Association of School Psychologists.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174. http://www.jstor.org/stable/2529310.

Lord, C., Risi, S., Lambrecht, L., Cook, E. W., Leventhal, B. L., DiLavore, P., et al. (2000). The autism diagnostic observation schedule-generic: A standard measure of social and communication deficits associated with spectrum of autism. *Journal of Autism and Developmental Disorders, 30*(3), 205–223.

Mautone, J. A., DuPaul, A. K., & Jitendra, A. K. (2005). The effects of computer-assisted instruction on the mathematics performance and classroom behavior of children with ADHD. *Journal of Attention Disorder, 9*, 301–312. doi:10.1177/1087054705278832.

Merikangas, K. R., He, J., Burstein, M., Swanson, S. A., Avenevoli, S., Cui, L., et al. (2010). Lifetime prevalence of mental disorders in U.S. adolescents: Results from the National Comorbidity Study-Adolescent Supplement (NCS-A). *Journal of the American Academy of Child and Adolescent Psychiatry, 49*(10), 980–989. doi:10.1016/j.jaac.2010.05.017.

Merrell, K. W. (1999). *Behavioral, social, and emotional assessment of children and adolescents*. Mawah, NJ: Lawrence Erlbaum Associates.

Miles, M. B., & Huberman, M. (1994). *Qualitative data analysis: A sourcebook of new methods* (2nd ed.). Beverly Hills, CA: Sage Publications.

Montague, M., McKinney, J. D., & Hocutt, A. (1994). Assessing students for attention deficit disorder. *Intervention in School and Clinic, 29*, 212–218.

Riley-Tilman, T. C., Kalberer, S. M., & Chafouleas, S.M. (2005). Selecting the right tool for the job: A review of behavior monitoring tools used to assess student response-to-intervention. The California School Psychologist, 10, 81–91. http://www.caspwebcasts.org/pdfs/JRNLv_081.pdf.

Shapiro, E. S. (2011). Behavior observations of students in schools. In E. S. Shapiro (Ed.), *Academic skills problems fourth edition workbook* (pp. 35–56). New York: Guilford Press.

Shapiro, E. S., & Heick, P. (2004). School psychologist assessment practices in the evaluation of students referred for social/behavioral/emotional problems. *Psychology in the Schools, 41*, 551–561. doi:10.1002/pits.10176.

Vile-Junod, R. E., DuPaul, G. J., Jitendra, A. K., Volpe, R. J., & Cleary, K. S. (2006). Classroom observations of students with and without ADHD: Differences across types of engagement. *Journal of School Psychology, 44*, 87–104.

Volpe, R. J., DiPerna, J. C., Hintze, J. M., & Shapiro, E. S. (2005). Observing students in a classroom setting: A review of seven coding schemes. *School Psychology Review, 34*(4), 454–474.

Wilson, M. S., & Reschly, D. J. (1996). Assessment in school psychology training and practice. *School Psychology Review, 25*, 9–23.