JOURNAL OF APPLIED BEHAVIOR ANALYSIS 2006, **39,** 429–440 NUMBER 4 (WINTER 2006)

ESTABLISHING THE FEASIBILITY OF DIRECT OBSERVATION IN THE ASSESSMENT OF TICS IN CHILDREN WITH CHRONIC TIC DISORDERS

MICHAEL B. HIMLE

UNIVERSITY OF WISCONSIN–MILWAUKEE

SUSANNA CHANG

UNIVERSITY OF CALIFORNIA–LOS ANGELES SCHOOL OF MEDICINE

DOUGLAS W. WOODS

UNIVERSITY OF WISCONSIN–MILWAUKEE

AMANDA PEARLMAN AND BRIAN BUZZELLA

UNIVERSITY OF CALIFORNIA–LOS ANGELES SCHOOL OF MEDICINE

LIVIU BUNACIU

UNIVERSITY OF WISCONSIN–MILWAUKEE

AND

JOHN C. PIACENTINI

UNIVERSITY OF CALIFORNIA–LOS ANGELES SCHOOL OF MEDICINE

Behavior analysis has been at the forefront in establishing effective treatments for children and adults with chronic tic disorders. As is customary in behavior analysis, the efficacy of these treatments has been established using direct-observation assessment methods. Although behavior-analytic treatments have enjoyed acceptance and integration into mainstream health care practices for tic disorders (e.g., psychiatry and neurology), the use of direct observation as a primary assessment tool has been neglected in favor of less objective methods. Hesitation to use direct observation appears to stem largely from concerns about the generalizability of clinic observations to other settings (e.g., home) and a lack of consensus regarding the most appropriate and feasible techniques for conducting and scoring direct observation. The purpose of the current study was to evaluate and establish a reliable, valid, and feasible direct-observation protocol capable of being transported to research and clinical settings. A total of 43 children with tic disorders, collected from two outpatient specialty clinics, were assessed using direct (videotape samples) and indirect (Yale Global Tic Severity Scale; YGTSS) methods. Videotaped observation samples were collected across 3 consecutive weeks and two different settings (clinic and home), were scored using both exact frequency counts and partial-interval coding, and were compared to data from a common indirect measure of tic severity (the YGTSS). In addition, various lengths of videotaped segments were scored to determine the optimal observation length. Results show that (a) clinic-based observations correspond well to home-based observations, (b) brief direct-

This work was supported by collaborative research grants from the Tourette Syndrome Association to Douglas Woods and John Piacentini. We acknowledge the contributions of Jordan Bonow, Ryan Walsh, Brecken Gilbert, Araceli Gonzalez, and the children and families who participated in this research.

Address correspondence to Douglas W. Woods, Department of Psychology, University of Wisconsin–Milwaukee, 2441 E. Hartford Ave., Garland Hall 224, Milwaukee, Wisconsin 53211 (e-mail: dwoods@uwm.edu).

doi: 10.1901/jaba.2006.63-06

observation segments scored with time-sampling methods reliably quantified tics, and (c) indirect methods did not consistently correspond with the direct methods.
DESCRIPTORS: tics, Tourette's syndrome, behavioral assessment, direct observation

Tics are defined as "sudden, rapid, recurrent, nonrhythmic, stereotyped motor movements or vocalizations" (American Psychiatric Association, 2000, p. 100). The tics of any two individuals are likely to differ in terms of the body location, number, frequency, complexity, intensity or forcefulness, noticability, and resulting social consequences. Even within individuals, multiple tics may be present, with each tic varying considerably in frequency, complexity, topography, and appearance (Leckman, King, & Cohen, 1999).

In terms of frequency, intertic intervals can range from seconds to hours or even days. The forcefulness with which a tic is performed can range from slight and barely noticeable to intense and obvious. Over time, the frequency and intensity of tics may wax and wane and may be influenced by a variety of internal and external stimuli including private events, contextual variables, and social reinforcement contingencies (Carr, Taylor, Wallander, & Reiss, 1996; O'Connor, Brisebois, Brault, Robillard, & Loiselle, 2003; Piacentini et al., in press; Silva, Munoz, Barickman, & Friedhoff, 1995; Watson & Sterling, 1998; Woods, Watson, Wolfe, Twohig, & Friman, 2001).

Tics can also vary considerably in their complexity. Tics that involve the contraction of a single muscle group are typically referred to as simple tics, and those that involve the contraction of multiple muscle groups are typically considered to be complex. Simple tics are typically of very short duration (<1 s) and include such behaviors as eye blinking; jerking of the face, head, torso, or limbs; coughing; sniffing; throat clearing; and making single-syllable sounds. Complex tics are often sustained for longer durations or occur in paroxysms and can include virtually any orchestrated pattern of behavior otherwise meeting the definition of a tic. Common examples include picking, tapping, gesturing, mimicking the gestures of others (echopraxia), repeating one's own speech (palilalia), mimicking the speech of others (echolalia), and the production of inappropriate words or sentences.

Behavior analysts have been at the forefront in developing and evaluating nonpharmacological treatments for tics (Miltenberger & Fuqua, 1985; Miltenberger, Fuqua, & McKinley, 1985; Woods, Miltenberger, & Lumley, 1996; Woods, Twohig, Flessner, & Roloff, 2003). As expected in behavior-analytic research, direct observation has been the preferred method for quantifying tic severity. However, researchers in disciplines outside of behavior analysis (e.g., psychiatry, neurology, and even the broader field of behavior therapy) have preferred indirect measures, such as clinical impression, self-report inventories, and clinician-rated scales. The most commonly cited reasons for not using direct observation include concerns about the generalizability of observations made in clinic or research settings to other relevant settings, such as home or school (Goetz, Leurgans, & Chmura, 2001; Nolan, Gadow, & Sverd, 1994) and disagreement about the best methods for collecting and scoring direct observation data (Chappell et al., 1994; Leckman et al., 1989; Nolan et al.; Walkup, Rosenberg, Brown, & Singer, 1992). Although the empirical basis for these concerns is not firmly established, acquisition of data supporting the use of direct observation methods may encourage those outside behavior analysis to use direct observation as a primary assessment method rather than relying on potentially biased verbal self-reports.

The current multisite study had three specific goals. The first goal was to evaluate the generalizability of clinic-based observations to

other settings by comparing clinic-based samples of tics to home-based samples. The second goal of the study was to evaluate the degree to which indirect measures correspond to direct observation by comparing the Yale Global Tic Severity Scale (YGTSS), the most commonly used indirect measure of tic severity, to direct observation. The third goal was to outline a feasible direct-observation methodology capable of being used in research and practice. We adopted the rationale that direct observation segments should be as brief and undemanding as possible while still providing a reliable and accurate sample of tics. We compared brief samples of direct observation to longer samples and also compared event-frequency coding to a less arduous time-sampling method (i.e., partial-interval coding). Although both methods have been used to evaluate outcomes in tic research (e.g., Azrin & Peterson, 1988; Peterson & Azrin, 1992; Woods et al., 1996, 2003), one might argue that partial-interval coding is more user-friendly because it does not require the observer to record each occurrence of the tic; thus, it might be preferred over the event-frequency method. However, partial-interval coding cannot be recommended as an alternative if it does not yield a reliable measure of the behavior. Because simulation studies have suggested that partial-interval coding may underestimate the frequency of high-rate short-duration responses, especially if they occur in rapid succession or as bouts (as is the case with many tics; Harrop & Daniels, 1986; Repp, Roberts, Slack, Repp, & Berkler, 1976), the two methods were examined in a population of individuals with tics.

METHOD

Participants

This study was conducted concurrently at the University of Wisconsin–Milwaukee (UWM) and the University of California–Los Angeles School of Medicine (UCLA) with approval of the respective institutional review boards. A total of 43 children, ages 8 to 17 years, participated (23 from UWM and 20 from UCLA). Participants were recruited through print advertisements, physician referrals, and referrals from the UWM Tic Disorders Specialty Clinic and the UCLA Child OCD, Anxiety, and Tic Disorders Program. To qualify for participation, children were required (a) to be generally healthy and between the ages of 8 and 17 years; (b) to have a diagnosis of Tourette's syndrome or chronic tic disorders (either motor or vocal); (c) to exhibit intellectual functioning in the low-average range or above as indicated by a score greater than 75 on a brief intelligence test; and (d) to have no self- or parent-reported recent (i.e., in the last 4 weeks) or planned initiation or change of dosage in medication or participation in behavioral treatment for their tics during the course of the study. Prior to participation, each child and parent signed appropriate informed consent documents and received a comprehensive assessment to determine study eligibility. The assessment consisted of a structured diagnostic interview, a brief assessment of intelligence, a videotaped observation, and a variety of self-, parent-, and clinician-rated instruments designed to assess the presence and severity of tic disorders and common comorbid conditions. Children received monetary compensation for their participation.

Participant characteristics are provided in Table 1. Given the two samples were collected from different types of health care facilities (outpatient psychiatry clinic at UCLA; community psychology clinic at UWM) located in different geographic regions of the country, appropriate analyses were conducted to determine cross-site differences on a variety of subject characteristics including age, IQ, YGTSS total severity scores, tic disorder diagnosis (e.g., Tourette's syndrome or chronic tic disorder), and comorbid status (comorbid psychiatric disorder present or absent). The samples were identical with respect to age,

Table 1
Sample Description by Site

	UWM (n = 23)	UCLA (n = 20)	Total (N = 43)
Mean age (*SD*)	10.9 (2.6)	10.7 (1.9)	10.8 (2.3)
IQ (*SD*)	112 (17.3)	103 (12.8)	108 (15.9)
Gender			
Male	21	18	39
Female	2	2	4
YGTSS (intake)			
Motor total (*SD*)	14.3 (4.1)	16.7 (4.1)	15.4 (4.2)
Vocal total (*SD*)	7.7* (6.4)	12.7* (6.8)	10.0 (7.0)
Overall total (*SD*)	21.7* (9.4)	29.7* (9.8)	25.4 (10.3)
Tic diagnosis			
Tourette's syndrome	21	20	41
Chronic tic disorder	2	0	2
Comorbidity status			
Yes	18	16	34
No	5	4	9

* $p < .05$.

gender, IQ, and comorbid status. Children at UCLA had greater vocal tic and overall tic severity than children at UWM, as measured by the YGTSS. This may not be surprising given that UCLA is an outpatient psychiatry clinic and UWM is a community psychology clinic.

Materials

Yale Global Tic Severity Scale. The YGTSS (Leckman et al., 1989) is a clinician-completed rating scale used to rate tic severity along several dimensions based on parent and child reports and clinician observations during the interview. Each dimension is represented by a subscale designed to quantify the number, frequency, duration, intensity, and complexity of both motor and vocal tics. Each subscale includes several descriptions to help the clinician make his or her ratings. Guided by these descriptions, each subscale is issued a rating between 0 and 5, with higher scores indicating greater severity.

Examples of descriptions included on the number subscale are "single tic," "multiple discrete tics," and "multiple discrete tics plus several orchestrated paroxysms of multiple simultaneous or sequential tics where it is difficult to distinguish discrete tics."

Examples of items on the frequency subscale are "rarely—specific tic behaviors have been present during the previous week; these behaviors occur infrequently, often not on a daily basis; if bouts of tics occur, they are brief and uncommon," and "always—specific tic behaviors are present virtually all the time."

Examples of items on the intensity subscale include "minimal intensity—tics not visible or audible (based solely on patient's private experience) or tics are less forceful than comparable voluntary actions and are typically not noticed because of their intensity" and "severe intensity— tics are extremely forceful and exaggerated in expression; these tics call attention to the individual and may result in risk of physical injury because of their forceful expression."

Examples on the complexity subscale include "borderline—some tics are not clearly 'simple' in character" and "severe—some tics involve lengthy bouts of orchestrated behavior or speech that would be impossible to camouflage or successfully rationalize as normal because of their duration or extremely unusual, inappropriate, bizarre, or obscene character."

Examples of interference items include "minimal—when tics are present, they do not interrupt the flow of behavior of speech" and "severe—when tics are present, they frequently disrupt intended action or communication."

Finally, examples of items on the impairment subscale include "minimal—tics associated with subtle difficulties in self-esteem, family life, social acceptance, or school or job functioning" and "severe—tics associated with extreme difficulties in self-esteem, family life, social acceptance, or school or job functioning."

The five subscales are rated separately for motor and vocal tics. The motor subscales are then summed to produce an overall motor tic severity rating, and the vocal tic subscales are summed to provide an overall vocal tic severity rating; each ranges from 0 to 25. The motor and vocal tic severity ratings are then summed

to produce an overall tic severity score that ranges from 0 to 50. In a recent normative clinical sample of 28 children with Tourette's syndrome, the mean total tic score of the YGTSS was 17.5 (SD = 11.70; Storch et al., 2005). Studies have shown the YGTSS total tic score to have acceptable internal consistency, good interrater reliability, and acceptable convergent and divergent validity in samples of adults and children (Leckman et al., 1989; Storch et al.).

Procedure

Setting and data collection. Each child was observed on six occasions spanning 3 weeks. Three of the observations were conducted at specialty clinics housed within the Psychology Department at UWM and the Department of Psychiatry at UCLA (clinic observations). The other three observations were conducted in the child's home (home observations). Clinic and home observations were conducted on consecutive days such that home visits occurred on the day either immediately before or after the clinic visit. The order of observation setting (home–clinic or clinic–home) was chosen randomly for each child but remained consistent across the 3 weeks. Each set of observations was separated by 1 week (for a total of 3 weeks). During each clinic visit, clinicians completed the YGTSS with the child and his or her parents and the child was observed for 30 min using the procedures described below. During home visits, a researcher visited the child's home and conducted a 30-min observation.

Observations were conducted by trained graduate or undergraduate research assistants and were conducted identically across the two sites. Clinic observations were conducted in observation rooms equipped with one-way observation mirrors and a videocamera that that recorded audio and video. Home observations were conducted in the child's home with a portable videocamera. During all observations, the child was seated in a chair while he or she watched a video or television. To control for any effects of the video or television content, each child watched the same video or television program across all observation sessions. During all observations, a camera was placed in front of the child oriented such that the child could be recorded without obstruction. The camera was in plain sight, and the child knew that he or she was being videotaped. The child was alone in the room while being videotaped.

Cross-site procedural training. Prior to the beginning of the study, a face-to-face meeting between personnel for both sites was held to review the standardized observation protocol and to conduct training on YGTSS administration and scoring. A uniform manual (available from the corresponding author) was used to describe the direct-observation data collection and scoring protocol. Sample tapes of children with tics were used to conduct cross-site YGTSS training and direct observation coding. Tapes included an interview and YGTSS administration conducted by the primary investigators with a child and his or her parents, along with a 10-min direct observation segment of the child. YGTSS training continued until the clinicians obtained agreement of at least 90% on the training tapes. Videotape coders were trained by the primary investigators. During training, the coders and primary investigators collectively reviewed and scored 10-min videotaped samples until each coder was competent in the procedures. Coders were considered competent when they (a) demonstrated the ability to follow the set-up procedures outlined in the scoring manual, (b) demonstrated the ability to use the coding materials, and (c) reached 90% agreement with the primary investigators when scoring the sample videotapes. Disagreements during training were resolved by discussion between the primary investigators and coders and recoding of videotapes until agreement criterion was reached.

Direct observation scoring. Based on information obtained during the initial assessment and

review of the initial 10-min videotapes, a list of operationally defined tics was created for each participant. The six observation segments were then randomly ordered for each participant and were scored for tic occurrence by the coders who were blind to segment number. Each 30-min segment was scored separately for motor tics, vocal tics, and total tics. To score the videotapes, a research assistant watched each segment on a television. Each tic was recorded using a simple computer program capable of recording and time-stamping computer keystrokes. As the researcher watched the videotape, he or she indicated the presence of each tic by pressing a designated key on a laptop computer. After the segment had been scored, the computer provided an output listing each keystroke (corresponding to each tic) and the precise time at which the keystroke was made, sensitive to a 10th of a second. A separate key was used to designate the onset and offset of any moments during which the child was not in full camera view (e.g., turned away from the camera, left the chair, etc.). Using the output from this program, tics were scored using two different methods. For the event-frequency method, each occurrence of the tic was counted and the total number of tics was divided by the number of minutes the child was visible, resulting in an index of tics per minute. For the partial-interval method, the 30-min videotape segment was divided into 180 10-s segments, and the observer recorded the presence or absence of motor or vocal tics in each interval. Although any length of observation segment could have been used, 10 s was chosen because it is the most commonly used interval length in studies of tics (e.g., Woods et al., 1996, 2003). Any intervals during which the child was not completely visible were excluded from the analysis. The primary dependent variable from the partial-interval method was the percentage of intervals during which one or more tics were observed; this was calculated by dividing the number of intervals in which a tic was observed by the total number of intervals in which the child was visible and multiplying by 100%.

Within-site and cross-site reliability. Two observers at each site, blind to segment order, independently scored a randomly selected 24% of total observation segments (clinic and home). Agreement for the partial-interval scoring method was calculated by dividing the number of intervals in which observers agreed on the presence or absence of a tic by the total number of intervals and multiplying by 100%. Overall within-site agreement for the partial-interval method was acceptable (M = 88%, range, 68% to 98%). Agreement for the event-frequency scoring method was calculated using a frequency-within-interval method (e.g., Sharenow, Fuqua, & Miltenberger, 1989). Each 30-min observation segment was first divided into consecutive 10-s intervals. Agreement within each interval was calculated by obtaining the number of tics recorded by each observer, dividing the higher number of tics observed by the lower number of tics, and multiplying by 100%. These scores were then averaged across the entire 30-min segment to yield an overall agreement score. Overall within-site agreement was acceptable (M = 76%, range, 58% to 99%).

To assess agreement across sites, 16% of segments collected at each site (clinic and home) were selected at random and scored by the other site. Agreement was calculated in the same way as the within-site agreement described above. Cross-site observer agreement was found to be acceptable for both the partial-interval method (M = 86%, range, 72% to 99%) and the event-frequency method (M = 72%, range, = 57% to 95%).

RESULTS

Generalization of direct observation from home to clinic. To evaluate the degree to which clinic-based observations generalized to the home setting, data from clinic observations were compared to home observations for the three

Table 2
Correlations (*r*) Between Clinic EF and PI and YGTSS Scores at Weeks 2 and 3

	Week 2 YGTSS Week 1 clinic direct observation	Week 3 YGTSS Week 2 clinic direct observation
EF (motor) – YGTSS motor frequency subscale	.07	.40*
PI (motor) – YGTSS motor frequency subscale	.15	.31*
EF (vocal) – YGTSS vocal frequency subscale	.64*	.62*
PI (vocal) – YGTSS vocal frequency subscale	.53*	.64*
EF (motor and vocal) – YGTSS severity score	−.12	−.04
PI (motor and vocal) – YGTSS severity score	−.15	−.14

* $p < .05$.
EF = event frequency; PI = partial interval; YGTSS = Yale Global Tic Severity Scale.

samples of behavior (Week 1, Week 2, and Week 3). (Sample description and comparisons between home and clinic observations first appeared in Piacentini et al., in press.) Reported and discussed in detail by Piacentini et al., statistical analyses showed that home-based and clinic-based observations were significantly and highly correlated at each of the three observation weeks, showing that the observed frequency of tics was consistent across settings. Follow-up analyses, conducted to analyze individual trends in clinic–home correspondence, showed that 15% of subjects had significantly more tics during the clinic observation (defined as a difference of 20% or greater change), 30% had significantly more tics during the home observation, and the majority (55%) showed no change or a variable pattern between the three home and clinic observations.

Correspondence of direct and indirect methods. To determine the correspondence between direct observation and YGTSS, Spearman's rho correlations were calculated between the clinic and home 30-min event-frequency and partial-interval data and the overall tic severity scores and tic frequency subscale scores of the YGTSS. Because the YGTSS is designed to measure tic severity retrospectively over the past week, YGTSS administration at Week 2 would be expected to correspond with observations at Week 1, and YGTSS administration at Week 3 would be expected to correspond with observations conducted at Week 2. As such, correlations were calculated between Week 2 YGTSS scores and Week 1 direct observation scores and between Week 3 YGTSS scores and Week 2 direct observation scores. Separate correlations were calculated for motor and vocal tics and for home and clinic observations. We were interested in two relations. First, we were interested in examining the correspondence between direct observation scores and YGTSS total severity scores (which are calculated by summing the five motor subscales and five vocal subscales of the YGTSS). Second, we were interested in examining the correspondence between direct observation scores and YGTSS frequency ratings (motor and vocal separately). Results are presented in Table 2 (clinic observations) and Table 3 (home observations). Overall, neither the event-frequency nor the partial-interval coding from either clinic or home observations correlated significantly with the YGTSS total severity scores (range, $r = -.04$ to $-.26$).

The second relation of interest was between direct observation and the specific subscales of the YGTSS. Given that the direct observation used in this study focused on tic frequency as the primary dependent variable, we were especially interested in the correlations between direct observation and the YGTSS frequency subscale scores. Correlations were conducted separately for motor and vocal tics and for home and clinic observations. Results are provided in Tables 2 and 3. For motor tics, correspondence between direct observation and YGTSS frequency scores was inconsistent across

Table 3
Correlations (*r*) Between Home EF and PI and YGTSS Scores at Weeks 2 and 3

	Week 2 YGTSS Week 1 clinic direct observation	Week 3 YGTSS Week 2 clinic direct observation
EF (motor) – YGTSS motor frequency subscale	.07	.26
PI (motor) – YGTSS motor frequency subscale	.08	.25
EF (vocal) – YGTSS vocal frequency subscale	.58*	.45*
PI (vocal) – YGTSS vocal frequency subscale	.58*	.44*
EF (motor and vocal) – YGTSS severity score	−.13	−.15
PI (motor and vocal) – YGTSS severity score	−.10	−.26

* $p < .05$.
EF = event frequency; PI = partial interval; YGTSS = Yale Global Tic Severity Scale.

administrations. At the second YGTSS administration (Week 2 YGTSS and Week 1 direct observation), correspondence between the YGTSS motor frequency score and direct-observation motor tic frequency was poor for observations conducted at both the clinic (r = .07 and .15 for event frequency and partial interval, respectively) and home (r = .07 and .08 for event frequency and partial interval, respectively). At the third YGTSS administration (Week 3 YGTSS and Week 2 direct observation), correspondence improved, especially for clinic observations (r = .40 and .31 for event frequency and partial interval, respectively; both correlations significant at $p >$ .05). For vocal tics, correspondence between direct observation and YGTSS frequency scores was good across both administrations. At both the second (Week 2 YGTSS and Week 1 direct observation) and third (Week 3 YGTSS and Week 2 direct observation) administrations, the YGTSS vocal tic frequency scores correlated significantly with direct-observation vocal tic frequencies from both clinic (r = .53 to .64; both correlations significant at $p >$ .05) and home (r = .44 to .58; both correlations significant at $p >$ .05) observations. Direct observation scores did not correlate significantly with any of the other YGTSS subscales (e.g., number, intensity, complexity, or interference).

Optimal duration of observation segment. To examine the feasibility of direct observation procedures, the correlations between various durations of observation segments were computed. Our assumption was that a feasible coding system should be as brief as possible while still capturing a representative sample of the behavior. As such, we attempted to determine the shortest duration of observation segment that was still highly related to the overall 30-min segment using both scoring methods. To calculate these relations, the 30-min segments were divided into four durations: 10 min, 5 min, 3 min, and 2 min. All durations correspond to the first *n* minutes of each 30-min observation segment, because it is most likely that the beginning segment of an observation would be used in a clinical setting. Correlations were calculated between the event-frequency and partial-interval results from each of the observation lengths and the overall 30-min segment for each of the three clinic observations. There was a progressive decline in the relation between the shorter segments and the entire 30-min segment (Table 4). However, even the 2-min segments of both coding methods were significantly correlated with their respective 30-min observations (r =

Table 4
Correlations Between 30-Min Segments and Respective 10-, 5-, 3-, and 2-Min Segments for EF and PI Methods

	10 min	5 min	3 min	2 min
30 min (EF)	.96*	.93*	.89*	.86*
30 min (PI)	.96*	.92*	.88*	.85*

* $p < .01$.
EF = event frequency; PI = partial interval.

Table 5
Correlations (*r*) Between EF and PI Methods at Each of the Clinic Observations

Observation	UWM	UCLA	Overall
1	.86*	.91*	.86*
2	.84*	.92*	.88*
3	.84*	.88*	.84*

* $p < .01$.

.86 and .85 for event frequency and partial interval, respectively), indicating that even brief observation periods provided information similar to longer segments in terms of tic frequency.

Convergence of event-frequency and partial-interval coding. To determine the convergence between the two coding methods, Pearson's *r* correlations were calculated between the the two types of data from the three clinic visits. Data are presented in Table 5. Across all three visits and both sites, event-frequency and partial-interval data correlated significantly ($r = .84$ to .88), indicating that the partial-interval method is highly related to direct frequency counts and hence is an acceptable method for quantifying tic frequency.

DISCUSSION

One of the hallmarks of behavior analysis is the use of direct observation to quantify behavior. This preference is based on the premise that direct observation is more objective than indirect methods such as self-report or clinician ratings of tic severity. Several studies, published in behavior-analytic outlets, demonstrate the value of using direct observation to quantify changes in tic frequency when evaluating behavioral treatments for tics (Azrin & Peterson, 1988; Peterson & Azrin, 1992; Woods et al., 1996, 2003). Still, many researchers outside behavior analysis have largely preferred indirect measures over direct observation (e.g., Bruun & Budman, 1996; Scahill et al., 2001; Silver et al., 2001). Among the foremost concerns raised by these researchers is that observations conducted within a clinical or research context may not generalize to other settings such as school or home. Generalization between settings is an important issue in both research and clinical practice. Indeed, it is not uncommon for parents to report that a child's tics are more or less severe while at the clinic compared to when the child is at home (Comings, 1990). Such reactivity to setting has been attributed to several factors including natural fluctuation, reinforcement contingencies, children's ability to volitionally suppress or temporarily withhold tics, reactivity to observation, and internal states such as anxiety (Comings; Himle & Woods, 2005). Regardless of the reason for contextual variation in tics, such fluctuations have important implications for the measurement of tics. If the scientific and clinical community is to have confidence in the results of behavior-analytic work utilizing direct observation methodology, observations conducted within a research setting must be generalizable to other settings. Results of this study (reported elsewhere, see Piacentini et al., in press) showed that clinic- and home-based observations were highly related, suggesting that, in general, clinic observations correspond well with home observations. However, examination of individual data shows that generalization should not necessarily be assumed; many children exhibited differential tic frequencies across the two settings, suggesting that, whenever possible, observations should be conducted in multiple settings.

Lack of consensus regarding the most reliable, valid, and feasible methods for collecting and coding direct observation data has also been cited as a reason for the preference of indirect measures over direct observation (Leckman et al., 1989). The second goal of the current study was to begin to develop a reliable, valid, and feasible direct observation protocol capable of being used in mainstream clinical and research settings. Logically, practitioners and researchers in disciplines outside behavior analysis may be more likely to use direct observation methods if

the effort associated with their use can be reduced, without any sacrifice of their validity and capacity to generate representative samples of target behaviors. In general, this study found that even short observation durations, coded with partial-interval methods, correlated well with longer durations scored with more effortful event-frequency methods.

The third goal of this study was to compare direct observation to the YGTSS, a commonly used indirect measure of tic severity. Results show that, overall, neither the clinic nor the home direct observation correlated significantly with the overall YGTSS scores. One might argue that this is not surprising given that the direct observation used in this study focused exclusively on tic frequency, whereas the YGTSS was designed to measure several dimensions of tic severity including number, frequency, complexity, intensity, and interference. However, correspondence between direct observation and YGTSS frequency subscale scores were also inconsistent across administrations. A conservative conclusion from these findings is that indirect measures of tic severity are inadequate and, as such, research that has relied exclusively on instruments designed for collection of indirect measures should be interpreted with caution. Furthermore, we believe that future research should, at minimum, supplement indirect measures with direct methods. Whether the YGTSS is capable of accurately capturing relevant dimensions of tic severity other than frequency is unknown and warrants further investigation.

There are a number of limitations to the current study. First, the small sample size limited our capacity to examine the psychometric properties of direct observation and indirect measures for high- versus low-frequency tics. At least one previous study demonstrated that correspondence between direct observation scores and YGTSS ratings may be lower for low-frequency tics than for high-frequency tics (Nolan et al., 1994). Similarly, the small sample size limited our ability to examine whether event-frequency and partial-interval methods are equally effective for detecting complex versus simple tics. It is possible that the correspondence between the two methods will vary with tic complexity or duration. For example, it is unclear whether the two methods are equally as adept at capturing low-frequency complex tics that occur for long durations versus high-frequency tics that are very brief.

Another limitation of the current study is that the direct observation samples were coded only for tic frequency, which is just one of several dimensions along which tics may vary. This study used frequency as the primary dependent variable because frequency is the dimension of tics most commonly used in behavior-analytic research (Azrin & Peterson, 1988; Peterson & Azrin, 1992; Woods et al., 1996, 2003). Whether tic frequency is the most important dimension of tic severity (e.g., best predicts psychosocial functioning) is an empirical issue that warrants investigation. To explore this issue, investigators should evaluate methods capable of quantifying multiple dimensions of tics including overt physical dimensions (e.g., frequency, intensity, complexity), social dimensions (e.g., social reinforcement and punishment contingencies, functional interference), and the concomitant private dimensions commonly reported to accompany tics (e.g., sensory events). The research we envision will likely require novel direct observation techniques used in combination with other measurement methods (e.g., functional assessment, self-report, clinician ratings, social acceptability ratings, physiological measures, neuroimaging techniques, etc.) and research strategies (e.g., functional analysis, group research designs, inferential statistical analyses).

The use of nontraditional measurement techniques to complement direct observation is likely to increase in popularity within the broader field of clinical behavior analysis. Clinical researchers are increasingly concerning

themselves with the study of behavior that is complex, highly variable, and not easily accessible by traditional direct-observation techniques (e.g., the private behaviors of individuals who suffer from anxiety and mood disorders). If behavior analysts are to continue to be at the forefront for understanding and treating clinical problems (including tic disorders), they must systematically determine which dimensions of specific target behaviors are socially relevant and must be diligent not to restrict themselves by investigating only those aspects that are easily quantifiable with traditional direct observation methods (Baer, 1986; Baer, Wolf, & Risley, 1968, 1987). This will require researchers both to refine their current measurement techniques and to incorporate techniques that have not traditionally been employed in behavior-analytic research (e.g., clinician ratings, self-report, physiological and neuroimaging techniques, etc.). This is not to suggest that clinical behavior analysts abandon direct observation in favor of other measurement techniques. On the contrary, it is a call to behavior analysts to develop, investigate, and incorporate new direct and indirect measurement techniques that will enhance scientific investigation of the environment–behavior relations involved in clinical problems.

REFERENCES

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

Azrin, N. H., & Peterson, A. L. (1988). Behavior therapy for Tourette's syndrome and tic disorders. In D. J. Cohen, J. F. Leckman, & R. D. Bruun (Eds.), *Tourette syndrome and tic disorders: Clinical understanding and treatment* (pp. 237–255). New York: Wiley.

Baer, D. M. (1986). In application, frequency is not the only estimate of the probability of behavior units. In M. D. Zeiler & T. Thompson (Eds.), *Analysis and integration of behavioral units* (pp. 117–136). Hillsdale, NJ: Erlbaum.

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.

Bruun, R. D., & Budman, C. L. (1996). Risperidone as a treatment for Tourette's syndrome. *Journal of Clinical Psychiatry*, *57*, 29–31.

Carr, J. E., Taylor, C. C., Wallander, J. J., & Reiss, M. L. (1996). A functional analytic approach to the diagnosis of transient tic disorder. *Journal of Behavior Therapy and Experimental Psychiatry*, *27*, 291–297.

Chappell, P. B., McSwiggan-Hardin, M. T., Scahill, L., Rubenstein, M., Walker, D. E., Cohen, D. J., et al. (1994). Videotape tic counts in the assessment of Tourette's syndrome: Stability, reliability, and validity. *Journal of the American Academy of Child and Adolescent Psychiatry*, *33*, 386–393.

Comings, D. E. (1990). *Tourette syndrome and human behavior*. Durante, CA: Hope Press.

Goetz, C. G., Leurgans, S., & Chmura, T. A. (2001). Home alone: Methods to maximize tic expression for objective videotape assessments in Gilles de la Tourette syndrome. *Movement Disorders*, *16*, 693–697.

Harrop, A., & Daniels, M. (1986). Methods of time sampling: A reappraisal of momentary time sampling and partial interval recording. *Journal of Applied Behavior Analysis*, *19*, 73–77.

Himle, M. B., & Woods, D. W. (2005). An experimental evaluation of tic suppression and the tic rebound effect. *Behaviour Research and Therapy*, *43*, 1443–1451.

Leckman, J. F., King, R. A., & Cohen, D. J. (1999). Tics and tic disorders. In J. F. Leckman & D. J. Cohen (Eds.), *Tourette's syndrome—Tics, obsessions, compulsions: Developmental psychopathology and clinical care* (pp. 23–42). New York: Wiley.

Leckman, J. F., Riddle, M. A., Hardin, M., Ort, S. I., Swartz, K. L., & Stevenson, J., et al. (1989). The Yale global tic severity scale: Initial testing of a clinician-rated scale of tic severity. *Journal of the American Academy of Child and Adolescent Psychiatry*, *28*, 566–573.

Miltenberger, R. G., & Fuqua, R. W. (1985). A comparison of contingent versus non-contingent competing response practice in the treatment of nervous habits. *Journal of Behavior Therapy and Experimental Psychiatry*, *16*, 195–200.

Miltenberger, R. G., Fuqua, R. W., & McKinley, T. (1985). Habit reversal with muscle tics: Replication and component analysis. *Behavior Therapy*, *16*, 39–50.

O'Connor, K. P., Brisebois, H., Brault, M., Robillard, S., & Loiselle, J. (2003). Behavioral activity associated with onset in chronic tic disorder and habit disorder. *Behaviour Research and Therapy*, *41*, 241–249.

Peterson, A. L., & Azrin, N. H. (1992). An evaluation of behavioral treatments for Tourette syndrome. *Behaviour Research and Therapy*, *30*, 167–174.

Piacentini, J., Himle, M. B., Chang, S., Baruch, D. E., Buzzella, B., & Pearlman, A., et al. (in press). Reactivity of tic observation procedures to situation and setting: A multisite study. *Journal of Abnormal Child Psychology*.

Repp, A. C., Roberts, D. M., Slack, D. J., Repp, C. F., & Berkler, M. S. (1976). A comparison of frequency, interval, and time-sampling methods of data collection. *Journal of Applied Behavior Analysis*, *9*, 501–508.

Scahill, L., Chappell, P. B., Kim, Y. S., Schultz, R. T., Katsovich, L., & Sheperd, E., et al. (2001). A placebo-controlled study of guanfacine in the treatment of children with tic disorders and attention hyperactivity disorder. *American Journal of Psychiatry*, *58*, 1067–1074.

Sharenow, E. L., Fuqua, R. W., & Miltenberger, R. G. (1989). The treatment of muscle tics with dissimilar competing response practice. *Journal of Applied Behavior Analysis*, *22*, 35–42.

Silva, R. R., Munoz, D. M., Barickman, J., & Friedhoff, A. J. (1995). Environmental factors and related fluctuation of symptoms in children and adolescents with Tourette's disorder. *Journal of Child Psychology & Psychiatry*, *36*, 305–312.

Silver, A. A., Shytle, R. D., Philipp, M. K., Wilkinson, B. J., McConville, B., & Sanberg, P. R. (2001). Transdermal nicotine and haloperidol in Tourette's disorder: A double-blind placebo-controlled study. *Journal of Clinical Psychiatry*, *62*, 707–714.

Storch, E. A., Murphy, T. K., Geffken, G. R., Sajid, M., Allen, P., & Roberti, J. W., et al. (2005). Reliability and validity of the Yale Global Tic Severity Scale. *Psychological Assessment*, *17*, 486–491.

Walkup, J., Rosenberg, L. A., Brown, J., & Singer, H. S. (1992). The validity of instruments measuring the severity in Tourette syndrome. *Journal of the American Academy of Child and Adolescent Psychiatry*, *31*, 472–477.

Watson, T. S., & Sterling, H. E. (1998). Brief functional analysis and treatment of a vocal tic. *Journal of Applied Behavior Analysis*, *31*, 471–474.

Woods, D. W., Miltenberger, R. G., & Lumley, V. A. (1996). Sequential application of major habit reversal components to treat motor tics in children. *Journal of Applied Behavior Analysis*, *29*, 483–493.

Woods, D. W., Twohig, M. P., Flessner, C., & Roloff, T. (2003). Treatment of vocal tics in children with Tourette's syndrome: Investigating the efficacy of habit reversal. *Journal of Applied Behavior Analysis*, *36*, 109–112.

Woods, D. W., Watson, T. S., Wolfe, E., Twohig, M. P., & Friman, P. C. (2001). Analyzing the influence of tic-related talk on vocal and motor tics in children with Tourette's syndrome. *Journal of Applied Behavior Analysis*, *34*, 353–356.

Received April 18, 2006
Final acceptance July 5, 2006
Action Editor, Pat Friman