Available online at www.sciencedirect.com
SCIENCE DIRECT®
Research in Developmental Disabilities
25 (2004) 99–118
Research in Developmental Disabilities

# Comparison of direct observational methods for measuring stereotypic behavior in children with autism spectrum disorders☆

Nicole Ciotti Gardenier[a,b,*], Rebecca MacDonald[a], Gina Green[a,c]

[a]*New England Center for Children, 33 Turnpike Road, Southborough, MA 01772-2108, USA*
[b]*Northeastern University, Boston, MA, USA*
[c]*Shriver Center, University of Massachusetts Medical School, Waltham, MA, USA*

Received 29 October 2001; received in revised form 17 April 2003; accepted 1 May 2003

**Abstract**

We compared partial-interval recording (PIR) and momentary time sampling (MTS) estimates against continuous measures of the actual durations of stereotypic behavior in young children with autism or pervasive developmental disorder—not otherwise specified. Twenty-two videotaped samples of stereotypy were scored using a low-tech duration recording method, and relative durations (i.e., proportions of observation periods consumed by stereotypy) were calculated. Then 10, 20, and 30 s MTS and 10 s PIR estimates of relative durations were derived from the raw duration data. Across all samples, PIR was found to grossly overestimate the relative duration of stereotypy. Momentary time sampling both over- and under-estimated the relative duration of stereotypy, but with much smaller errors than PIR (Experiment 1). These results were replicated across 27 samples of low, moderate and high levels of stereotypy (Experiment 2).


*Keywords:* stereotypy; direct observational methods; measurement; autism

☆This report is based on a thesis completed by Nicole Ciotti Gardenier in partial fulfillment of requirements for a master of science degree in applied behavior analysis from Northeastern University. Portions of this research were presented at the annual convention of the Association for Behavior Analysis, May 1999 and May 2000.

* Corresponding author. Tel.: +1-508-481-1015x3105; fax: +1-508-485-3421.
*E-mail address:* ngardenier@necc.org (N.C. Gardenier).

0891-4222/$ – see front matter 
doi:10.1016/j.ridd.2003.05.004

## 1. Introduction

Stereotypic behavior, or stereotypy, is generally defined as repetitive motor or vocal responses that serve no obvious adaptive functions (LaGrow & Repp, 1984). It is a key diagnostic feature of autism spectrum disorders (American Psychiatric Association, 1994; Lewis & Bodfish, 1998), but also occurs frequently in people with sensory impairments and mental retardation, less frequently in typically developing children and adults (Rojahn, Matlock, & Tasse, 2000). Many different topographies of stereotypy have been described in the research literature, including body rocking, pacing, posturing, vocalizing, sniffing, facial grimacing, nonsocial laughing, manipulating objects, and repetitively moving various body parts, such as hands, arms, legs, and eyes (LaGrow & Repp, 1984; Lewis & Bodfish, 1998; Rojahn et al., 2000).

Engagement in stereotypy has been found to interfere with both the acquisition of new skills and the performance of established skills (e.g., Epstein, Doke, Sajwaj, Sorrell, & Rimmer, 1974; Koegel & Covert, 1972; Morrison & Rosales-Ruiz, 1997). It can also be stigmatizing (Jones, Wint, & Ellis, 1990), and may be a precursor of self-injurious behavior (Guess & Carr, 1991; Schroeder, Rojahn, Mulick, & Schroeder, 1990). For these and other reasons, reducing stereotypy is often a high priority for intervention. Additionally, stereotypy has been the topic of extensive research by investigators who are interested in its pathophysiology and in discovering what a better understanding of stereotypy may reveal about the underlying neurophysiology of disorders like autism (LaGrow & Repp, 1984; Lewis & Bodfish, 1998; Rojahn et al., 2000).

Precise measurement of stereotypy is essential to both of these enterprises. Practitioners are typically concerned with the stereotypic behavior exhibited by an individual—its topography, how often it occurs, how much time it consumes, and the conditions under which it occurs and does not occur. Detailed and accurate information about all of those dimensions is critical to determining the baseline or pre-treatment level of stereotypy, designing intervention programs, and evaluating the effects of intervention for that individual. Frequent direct observation and recording of the behavior in various settings is most likely to provide that kind of information. Some researchers studying the phenomenology or prevalence of stereotypy in populations of people with disabilities or group treatment outcomes have used behavior rating scales to measure stereotypy. From a measurement standpoint, rating scales have many well-documented problems (e.g., high subjectivity, low inter-rater agreement, unknown accuracy; Johnston & Pennypacker, 1993, p. 127). Nevertheless, several such scales for measuring stereotypy have been developed and evaluated; their psychometric properties and utility vary (Lewis & Bodfish, 1998; Rojahn et al., 2000). Other researchers, recognizing that rating scales often do not capture inter-individual variations in the topographical, numerical, and temporal dimensions of stereotypy, have used direct observational measurement methods (e.g., McEntee & Saunders, 1997; Pyles, Riordan, & Bailey, 1997).

Direct observational measurement of stereotypy presents several challenges, in part because of the characteristics of this category of behavior. Stereotypic responses are by definition repetitive actions that often occur in rapid succession, making it very difficult to identify discrete starting and ending points for each action, which in turn makes it impossible to record response frequencies. On the other hand, stereotypic responses typically occur in bouts, or episodes. Episodes do have clear-cut beginnings and endings, so their frequency can be measured readily, but they are extended in time, so the duration of each episode is also of interest. Additionally, it is often important to determine how much time overall is consumed by stereotypy, so the relative duration of stereotypy during each observation period is often calculated as well. Continuous observation, timing, and recording of the duration of each episode of stereotypy require a dedicated observer, that is, one who can devote his or her undivided attention to observing and recording the behavior. Such observers are scarce in many intervention programs. Even in research settings, observers may be required to record several behaviors concurrently. This again can pose logistical and measurement challenges when stereotypic behavior is of interest.

The difficulties inherent in conducting continuous direct observational recording of stereotypy has led many practitioners and researchers to adopt sampling methods like partial-interval recording (PIR) and momentary time-sampling (MTS). These discontinuous recording methods necessarily yield inaccurate estimates of the actual level of stereotypy, because they do not allow detection and recording of every occurrence of the behavior. Further, the degree of inaccuracy cannot be evaluated because these methods do not produce a complete record of responding (Johnston & Pennypacker, 1993). Despite these shortcomings, PIR and MTS methods recording methods continue to be used widely in applied research (Johnston, 1996).

Obviously, it is important to know how accurately PIR and MTS methods estimate the actual levels of stereotypy and other behaviors with similar characteristics. A number of investigators have compared estimates produced by MTS and PIR. For example, Powell, Martindale, and Kulp (1975) measured the duration of the in-seat behavior of a secretary from 20 min videotaped samples. Partial- and whole-interval as well as MTS estimates were then derived from the duration data, using intervals of 10–120 s for PIR and whole-interval recording (WIR), 10–600 s for MTS. Results indicated that PIR consistently overestimated the actual duration of the behavior, while WIR underestimated duration. Measurement error increased with increasing interval length. MTS with interval lengths between 10 and 120 s both over- and under-estimated the duration of in-seat behavior, but with considerably less error than either PIR or WIR. Overall, MTS methods yielded comparable and reasonably accurate estimates of actual duration, with measurement error increasing with increasing interval length.

In a follow-up to the study by Powell et al. (1975), PIR and WIR estimates of the duration of scheduled in-seat behavior (i.e., programmed to occur during 20, 50, or 80% of videotaped sessions) were derived using interval lengths of 5–300 s, and MTS estimates were derived using interval lengths of 5–900 s. Overall, WIR

was found to underestimate the true duration of the behavior, PIR overestimated, and MTS both over- and under-estimated. At 5 s intervals, all three methods produced similar estimates, but as interval length increased, MTS consistently yielded more accurate estimates of duration than either PIR or WIR. As interval lengths increased, however, MTS began to yield inconsistent results, varying between over- and under-estimations. The degree of measurement error varied with the relative duration of the behavior. When in-seat behavior occurred for 80% of an observation period, PIR overestimated actual relative duration by no more than 20%, but when the actual relative duration was 20% of the observation period, PIR overestimated by as much as 80%. The authors noted that such an inconsistent error pattern with changes in the level of a behavior over time would render PIR invalid for evaluating intervention effects (Powell, Martindale, Kulp, Martindale, & Bauman, 1977). A similar study found that MTS methods produced more accurate estimates of the relative duration of hair twisting that was programmed to occur for 20, 50, or 75% of on observation period than did PIR methods, which consistently overestimated the duration of the behavior (Green, McCoy, Burns, & Smith, 1982).

Two studies compared the accuracy with which PIR and MTS methods estimated the duration of simulated, rather than live, behavior. The term ''simulated behavior'' refers to response patterns generated by a computer or some mechanical means, as opposed to behavior emitted by living organisms. Repp, Roberts, Slack, Repp, and Berkler (1976) used electromechanical equipment to generate simulated responses of different rates (0.1, 1 and 10 responses per minute) and patterns (constant responding or bursts of responding). Two MTS procedures were used to estimate rates. With the first, observation occurred for 5 s at the end of each 10 min interval. If at least one response was observed during the 5 s, an occurrence was scored. With the second method, observation occurred for 10 s at the end of each 10 min interval, and an occurrence was scored if the response occurred at least once during the 10 s. The same sessions of simulated behavior were also scored with two PIR procedures, described as 10 s observe, 5 s record, and 10 s observe, 0 s record, respectively. An occurrence was scored if the response was observed at least once during the observe portion of each interval. Results were reported in terms of the percent deviation from the PIR 10 s observe, 0 s record method, as this was the only one that included responding from all the intervals. The authors reported that estimates produced by the PIR 10 s observe, 5 s record method most closely matched those produced by the PIR 10 s observe, 0 s record method. Both MTS methods were found to produce errors.

The Repp et al. (1976) study is difficult to interpret for several reasons. First, the procedures described as MTS were not consistent with standard MTS procedures, which involve observing behavior briefly at the very end of a specified interval, usually for only 1–2 s (Barton & Ascione, 1984; Cooper, 1987). Observing for 5 or 10 s at the end of each interval and recording only one occurrence regardless of how many times the behavior was observed during the interval is tantamount to PIR recording and has the same limitations. Further, the 10 min intervals used by Repp et al. (1976) were considerably longer than what is

generally recommended for interval-based recording, given that measurement errors have been found to occur at unacceptable levels with interval durations that exceed 30 s (Powell et al., 1977). Finally, Repp et al.'s (1976) comparison of the data yielded by MTS and PIR methods to those produced by another PIR method provided no information about the extent to which any of those estimates corresponded to the true rate or duration of the simulated behavior.

In another study involving simulated behavior, Harrop and Daniels (1986) used a computer to create samples of responding with consistent bout durations of 1, 5, 10 or 20 s, as well as samples of low-to-medium and medium-to-high rates. Rate was defined in terms of the probability of the onset of a response, which was 1:180–1:30 for low-to-medium rates, and 1:30–1:5 for medium-to-high rates. One-hour samples were divided into 15 s intervals and measured using both MTS and PIR methods. For MTS, the behavior was observed for 1 s at the end of each interval. For PIR, the behavior was observed for 10 s, followed by a 5 s recording period. Results were reported as the mean percentage error relative to the absolute level for both the rate (response per unit of time) and duration (total time the behavior occurred) of the simulated behavior. Overall, PIR was found to provide more accurate estimates of rate than MTS, but MTS more accurately estimated duration. PIR consistently overestimated the duration of the behavior, while no systematic error was found with MTS methods. Both methods produced more errors with samples containing short bouts at low rates of responding.

To our knowledge, only one published study compared methods of estimating levels of stereotypy. Murphy and Goodall (1980) collected videotaped samples of the stereotypic behavior of children with mental retardation. Samples were categorized by bout length: short (2 s or less); medium (2–20 s); and long (35 s–2 min). An event recorder was used to measure the duration of each bout and the relative duration of stereotypy in each 5 min sample. The number of bouts and their mean duration were also calculated. Estimates were then derived from the event records with the following methods: WIR (10 s intervals), PIR (10 s observe, 10 s record and 2.5 s observe, 7.5 s record) and MTS (10 s intervals). The authors reported that MTS produced more accurate estimates of the true duration of the stereotypic behavior than PIR. PIR consistently overestimated duration, but with smaller errors at short intervals (2.5 s) than at longer intervals (10 s).

In summary, results of several studies have suggested that MTS methods produce more accurate estimates of the duration of responding than PIR methods. The majority of those studies, however, involved either human behavior that was programmed to occur at specified levels, or simulated behavior. Only Murphy and Goodall (1980) compared methods for measuring naturally occurring stereotypy. We sought to replicate and extend their study by asking the following questions: Does PIR or MTS produce more accurate estimates of the actual duration of the stereotypic behavior of a sample of young children diagnosed with autism spectrum disorders (Experiment 1)? Does the accuracy of estimates produced by PIR and MTS methods vary as a function of the overall level of stereotypy (Experiment 2)?

## 2. Experiment 1

This experiment compared the accuracy of MTS and PIR estimates with the continuously measured duration of stereotypic behavior.

### 2.1. Method

#### 2.1.1. Participants

Fifteen children diagnosed with autism or pervasive developmental disorder—not otherwise specified (PDD-NOS) participated. All were enrolled in an intensive behavior analytic preschool program. The children ranged in age from 3 years, 1 month to 5 years, 11 months. They are described in Table 1.

#### 2.1.2. Samples and setting

An operational definition was developed that included both motor and vocal stereotypy. This definition, which appears in Appendix A, was used for both Experiments 1 and 2. The definition was made purposely broad in order to encompass the full range of topographies of motor and vocal stereotypy exhibited by young children with autism spectrum disorders. A working draft was tested by having experienced observers use it to record data on nonexperimental samples of stereotypy, calculating indices of interobserver agreement among those observers, and revising the definition to remediate ambiguities or omissions.

Experimental samples of stereotypy were obtained from videotaped assessment sessions conducted with each child. A total of 22 samples were used in this experiment. All sessions took place in a research room containing a table and two chairs, as well as a bookshelf with an assortment of age-appropriate toys.

Table 1
Characteristics of participants in Experiment 1

| Participant | Gender | Chronological age | Diagnosis |
|---|---|---|---|
| CAZ | Male | 3 years, 10 months | Autism |
| CRN | Male | 3 years, 3 months | Autism |
| CRR | Male | 5 years, 11 months | Autism |
| DUL | Male | 3 years, 2 months | Autism |
| LBY | Male | 4 years, 4 months | Autism |
| NBE | Female | 3 years, 9 months | Autism |
| PCL | Female | 4 years, 4 months | Autism |
| RPO | Male | 4 years, 4 months | Autism |
| SRN | Male | 3 years, 1 month | Autism |
| BCD | Male | 3 years, 7 months | PDD-NOS |
| CAG | Male | 5 years, 9 months | PDD-NOS |
| JGN | Male | 4 years, 2 months | PDD-NOS |
| JLY | Male | 3 years, 1 month | PDD-NOS |
| PES | Male | 5 years, 1 month | PDD-NOS |
| WGN | Male | 3 years, 4 months | PDD-NOS |

*Note*. Age reflects participant's age at time of first sample.

The experimenter, an independent observer, and a child were present for each session. Sessions consisted of a series of trials designed to evaluate eye contact, imitation, and direction-following skills, as well as a 5 min free play period. During the free play period, the experimenter did not interact with the child. Each session lasted approximately 30 min, but only the first 10 min of each session were used for this experiment.

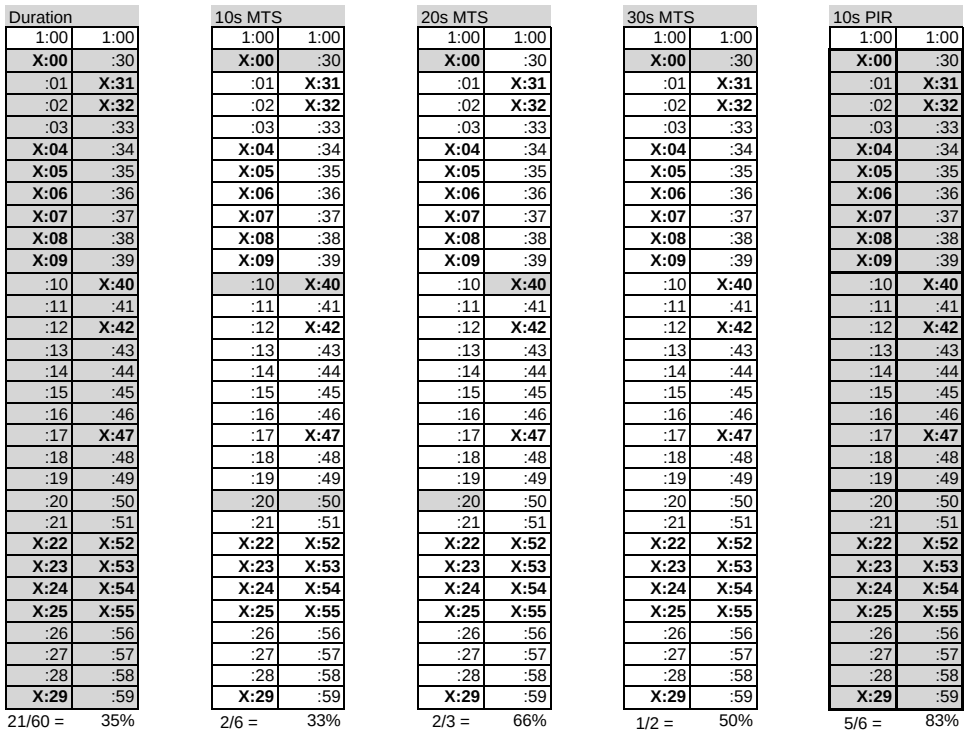
*2.1.3. Data collection procedures*

A television and videocassette recorder combination was used for all data collection (Symphonic brand, model #SC313A). A counter was displayed on the bottom of the TV screen, which indicated the time elapsed from the beginning of the videotape being played in seconds.

The duration of each episode of stereotypy occurring during each 10 min observation period was measured by using the second-by-second countup display on the television screen to record the time of onset and offset of each episode of stereotypy (Miltenberger, Rapp, & Long, 1999). When necessary, the observer rewound and re-played the tape to ensure accurate recording. This method yielded a continuous, complete record of all episodes of stereotypy occurring during each observation period. Relative duration was calculated by summing the durations of all episodes, dividing by 600 s (the duration of the observation period), and multiplying by 100%.

The relative duration of stereotypy in each observation period was estimated from the raw duration data by simulating two interval-based measurement methods. PIR was simulated by dividing the 0 min observation period into 10 s intervals, for a total of 60 intervals. An occurrence was scored if stereotypy occurred one or more times during each 10 s interval. A non-occurrence was scored if no stereotypy occurred during the interval. Relative duration was estimated by dividing the number of intervals in which occurrences were recorded by 60 and multiplying by 100%.

MTS was simulated using three interval lengths. Powell et al. (1977) demonstrated that with interval durations greater than 60 s, MTS often yields large measurement errors. Therefore, we used intervals of 10, 20 and 30 s for the MTS estimates. For MTS 10 s, 1 s at the end of every 10 s interval was ''observed'' by examining the raw duration data. An occurrence was scored if stereotypy occurred during that 1 s; a non-occurrence was scored if stereotypy did not occur during that 1 s. This yielded a total of 60 observations for the 10 min observation period. A similar procedure was used to simulate MTS 20 s: 1 s was observed every 20 s, for a total of 30 observations. For MTS 30 s, 1 s was observed every 30 s, for a total of 20 observations. Relative duration was estimated by dividing the number of occurrences by the total number of intervals (60, 30 or 20 s) and multiplying by 100%. Fig. 1 is a representation of how each of the estimation methods were simulated. Estimates produced by MTS and PIR methods were compared to the continuously recorded duration measures to evaluate the degree of error. Error was calculated by calculating the difference between the continuously measured, actual relative duration and the estimated relative duration.



| Duration | |
|---|---|
| 1:00 | 1:00 |
| **X:00** | :30 |
| :01 | **X:31** |
| :02 | **X:32** |
| :03 | :33 |
| **X:04** | :34 |
| **X:05** | :35 |
| **X:06** | :36 |
| **X:07** | :37 |
| **X:08** | :38 |
| **X:09** | :39 |
| :10 | **X:40** |
| :11 | :41 |
| :12 | **X:42** |
| :13 | :43 |
| :14 | :44 |
| :15 | :45 |
| :16 | :46 |
| :17 | **X:47** |
| :18 | :48 |
| :19 | :49 |
| :20 | :50 |
| :21 | :51 |
| **X:22** | **X:52** |
| **X:23** | **X:53** |
| **X:24** | **X:54** |
| **X:25** | **X:55** |
| :26 | :56 |
| :27 | :57 |
| :28 | :58 |
| **X:29** | :59 |
| 21/60 = | 35% |

| 10s MTS | |
|---|---|
| 1:00 | 1:00 |
| **X:00** | :30 |
| :01 | **X:31** |
| :02 | **X:32** |
| :03 | :33 |
| **X:04** | :34 |
| **X:05** | :35 |
| **X:06** | :36 |
| **X:07** | :37 |
| **X:08** | :38 |
| **X:09** | :39 |
| :10 | **X:40** |
| :11 | :41 |
| :12 | **X:42** |
| :13 | :43 |
| :14 | :44 |
| :15 | :45 |
| :16 | :46 |
| :17 | **X:47** |
| :18 | :48 |
| :19 | :49 |
| :20 | :50 |
| :21 | :51 |
| **X:22** | **X:52** |
| **X:23** | **X:53** |
| **X:24** | **X:54** |
| **X:25** | **X:55** |
| :26 | :56 |
| :27 | :57 |
| :28 | :58 |
| **X:29** | :59 |
| 2/6 = | 33% |

| 20s MTS | |
|---|---|
| 1:00 | 1:00 |
| **X:00** | :30 |
| :01 | **X:31** |
| :02 | **X:32** |
| :03 | :33 |
| **X:04** | :34 |
| **X:05** | :35 |
| **X:06** | :36 |
| **X:07** | :37 |
| **X:08** | :38 |
| **X:09** | :39 |
| :10 | **X:40** |
| :11 | :41 |
| :12 | **X:42** |
| :13 | :43 |
| :14 | :44 |
| :15 | :45 |
| :16 | :46 |
| :17 | **X:47** |
| :18 | :48 |
| :19 | :49 |
| :20 | :50 |
| :21 | :51 |
| **X:22** | **X:52** |
| **X:23** | **X:53** |
| **X:24** | **X:54** |
| **X:25** | **X:55** |
| :26 | :56 |
| :27 | :57 |
| :28 | :58 |
| **X:29** | :59 |
| 2/3 = | 66% |

| 30s MTS | |
|---|---|
| 1:00 | 1:00 |
| **X:00** | :30 |
| :01 | **X:31** |
| :02 | **X:32** |
| :03 | :33 |
| **X:04** | :34 |
| **X:05** | :35 |
| **X:06** | :36 |
| **X:07** | :37 |
| **X:08** | :38 |
| **X:09** | :39 |
| :10 | **X:40** |
| :11 | :41 |
| :12 | **X:42** |
| :13 | :43 |
| :14 | :44 |
| :15 | :45 |
| :16 | :46 |
| :17 | **X:47** |
| :18 | :48 |
| :19 | :49 |
| :20 | :50 |
| :21 | :51 |
| **X:22** | **X:52** |
| **X:23** | **X:53** |
| **X:24** | **X:54** |
| **X:25** | **X:55** |
| :26 | :56 |
| :27 | :57 |
| :28 | :58 |
| **X:29** | :59 |
| 1/2 = | 50% |

| 10s PIR | |
|---|---|
| 1:00 | 1:00 |
| **X:00** | :30 |
| :01 | **X:31** |
| :02 | **X:32** |
| :03 | :33 |
| **X:04** | :34 |
| **X:05** | :35 |
| **X:06** | :36 |
| **X:07** | :37 |
| **X:08** | :38 |
| **X:09** | :39 |
| :10 | **X:40** |
| :11 | :41 |
| :12 | **X:42** |
| :13 | :43 |
| :14 | :44 |
| :15 | :45 |
| :16 | :46 |
| :17 | **X:47** |
| :18 | :48 |
| :19 | :49 |
| :20 | :50 |
| :21 | :51 |
| **X:22** | **X:52** |
| **X:23** | **X:53** |
| **X:24** | **X:54** |
| **X:25** | **X:55** |
| :26 | :56 |
| :27 | :57 |
| :28 | :58 |
| **X:29** | :59 |
| 5/6 = | 83% |

Fig. 1. Sample data sheet. Each panel displays 1 min of recording, X marks an occurrence of stereotypy for that second, and gray boxes indicate those seconds that were ''observed'' for each simulated estimate method. Leftmost panel depicts the continuously recorded duration, with the relative duration calculation at the bottom. Duration data are replicated across the other panels, depicting MTS 10 s, MTS 20 s, MTS 30 s, and PIR 10 s methods, respectively. The estimates produced by each of those measurement methods are represented at the bottom of each panel.

*2.1.4. Interobserver agreement and procedural integrity*

Observers who served as primary and secondary data collectors for the study were required to record data with high degrees (>85%) of accuracy and inter-observer agreement for three consecutive practice sessions with nonexperimental videotaped samples, using the final operational definition and the continuous recording method described previously, before formal data collection began. During the course of the study, if IOA fell below the 85% criterion, the observers were re-trained.

An independent observer scored the first 3 min of each experimental sample of stereotypy using the continuous recording (actual duration) method. A sliding rule was used to determine agreement between observers (R. Miltenberger, personal communication, April 6, 1999). An agreement was defined as identification of the onset and offset of an episode by both observers within 1 s of each other. All overlapping seconds were considered agreements. For example, if the primary observer recorded an episode occurring from 1:33 to 1:42 and the secondary observer recorded the episode occurring from 1:35 to 1:46, there were 4 s of disagreement (with a 1 s window at the beginning and end) and 5 s of agreement (1:35–1:42). Interobserver agreement (IOA) was calculated by dividing the number of agreements by the total number of agreements plus disagreements and multiplying by 100%. Across all observation periods and all samples, IOA ranged from 77 to 100%, with a mean of 93%. Interobserver agreement was also calculated with Cohen's kappa, which yielded a mean of .94 ($p < .001$), range .92–.98.

The accuracy of the experimenter's calculation of the data produced by each measurement method was examined for 6 of the 22 samples (27%) to evaluate procedural integrity. An independent observer went through each duration record and calculated the actual relative duration as well as the percent of intervals in which occurrences were recorded by each of the estimate methods. Procedural integrity was calculated by dividing the smaller estimate by the larger estimate and multiplying by 100%. Agreement ranged from 93 to 100%, with a mean of 99%.

*2.2. Results*

Fig. 2 depicts the relative duration of stereotypy yielded by each measurement method for all samples. As the figure shows, PIR consistently overestimated the duration of stereotypy. MTS both over- and under-estimated the actual duration, but to a much lesser degree than the overestimates produced by PIR.

The extent to which the MTS and PIR estimates deviated from the actual duration measure is shown as measurement error in Fig. 3. Measurement error was calculated as the difference between the actual duration and the MTS and PIR estimates (cf. Murphy & Goodall, 1980). For example, the duration of stereotypy for the first sample was 9%. As Fig. 3 shows, MTS 10 s overestimated the actual duration by 1%, while MTS 20 s underestimated the actual duration of stereotypy by 6%. MTS 30 s overestimated the actual duration of stereotypy by 11%, while PIR 10 s overestimated by 33%.

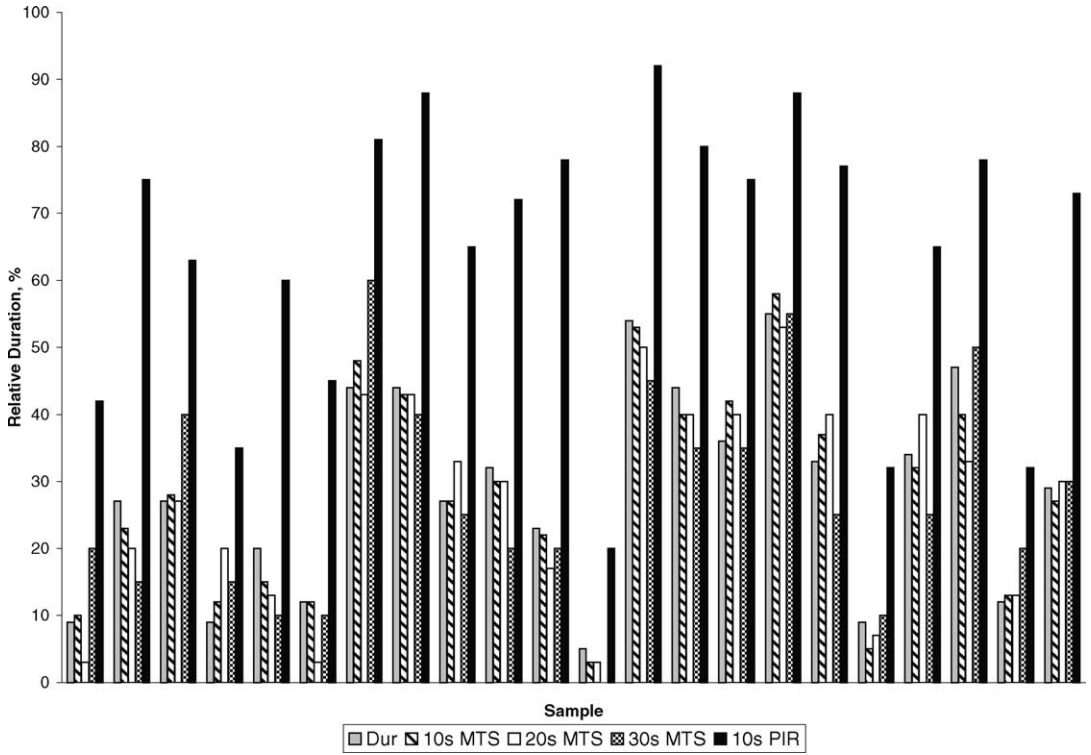



Fig. 2. Duration (gray bars) of stereotypy and the estimates produced by MTS 10 s (striped bars), MTS 20 s (white bars), MTS 30 s (checked bars) and PIR 10 s (black bars). The *x*-axis depicts the participant and sample. The *y*-axis depicts the relative duration of stereotypy.

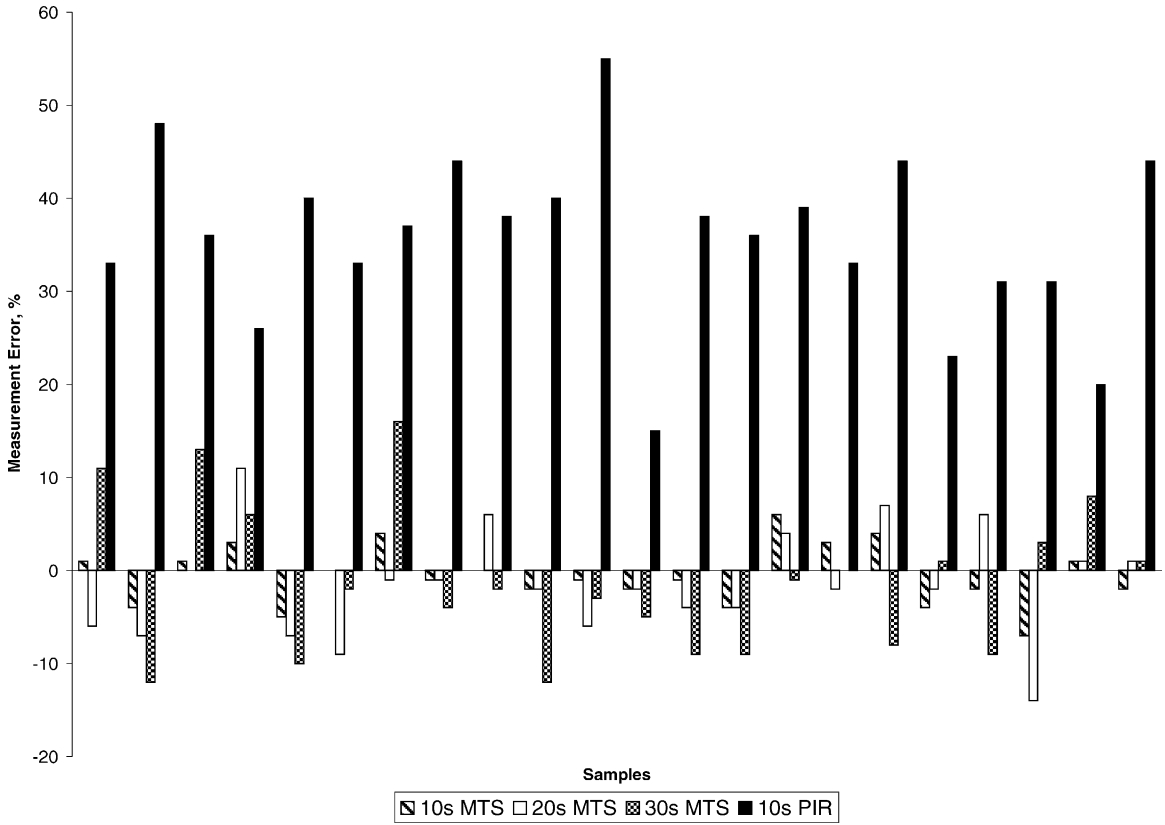



Fig. 3. Measurement error for MTS 10 s (striped bars), MTS 20 s (white bars), MTS 30 s (checked bars) and PIR 10 s (black bars). The *x*-axis depicts the participant and sample. The *y*-axis depicts the degree of measurement error.

In addition, the percent difference relative to the actual duration of stereotypy was calculated for each estimate method. For example, when the actual duration of stereotypy was 9% and 10 s MTS estimated 10%, the difference (1) was divided by the duration measure and multiplied by 100% ($[1/9] \times 100\%$). The result, 11%, is the relative percent difference between the actual duration measure and the 10 s MTS estimate. Across all samples, PIR overestimated the actual duration of the behavior by an average of 164% (range: 60–367%). MTS 10, 20 and 30 s over- and under-estimated the duration of the behavior by an average of 12% (range: 0–44%), 25% (range: 0–122%), and 28% (range: 0–122%), respectively.

## 3. Experiment 2

The purpose of Experiment 2 was to further investigate the accuracy of measurement methods with samples of stereotypy occurring at low, moderate, and high levels.

### *3.1. Method*

#### *3.1.1. Samples and participants*

Twenty-seven videotaped samples of stereotypy were scored for this experiment. The samples were collected from the first 10 min of standardized assessment sessions (as described in Experiment 1) conducted with 16 children diagnosed with autism or PDD-NOS. The children ranged in age from 2 years, 5 months to 5 years, 11 months. All were enrolled in an intensive behavior analytic preschool. They are described in Table 2.

#### *3.1.2. Data collection*

All samples were first scored using the continuous duration recording method described for Experiment 1. The range of relative durations of stereotypy for the entire pool of samples was examined, and samples were grouped arbitrarily into categories of low, moderate, and high levels of stereotypy. Low-level stereotypy was defined as behavior that occurred for a relative duration of less than 19% (10 samples). Moderate-level stereotypy occurred for a relative duration of 20–39% (11 samples). High-level stereotypy was defined as behavior that consumed 40% or more of the observation period (6 samples). MTS 10, 20, 30 s and PIR 10 s estimates of relative duration were then derived from the continuously measured duration data for all samples, as in Experiment 1.

#### *3.1.3. Interobserver agreement and procedural integrity*

The first 3 min of each sample were scored by an independent observer and the experimenter and IOA calculated as described in Experiment 1. Across all samples, IOA ranged from 77 to 100%, with a mean of 94%. Cohen's kappa was also calculated to assess IOA, yielding a mean of .96 ($p < .001$) and a range

Table 2
Characteristics of participants in Experiment 2

| Participant | Gender | Chronological age | Diagnosis |
|---|---|---|---|
| CAZ | Male | 2 years, 10 months | Autism |
| CCO[a] | Male | 2 years, 5 months | Autism |
| CRN | Male | 3 years, 3 months | Autism |
| CRR | Male | 5 years, 11 months | Autism |
| DUL | Male | 3 years, 2 months | Autism |
| LBY | Male | 4 years, 4 months | Autism |
| NBE | Female | 3 years, 9 months | Autism |
| PCL | Female | 4 years, 4 months | Autism |
| RPO | Male | 4 years, 4 months | Autism |
| SRN | Male | 3 years, 1 month | Autism |
| BCD | Male | 2 years, 8 months | PDD-NOS |
| CAG | Male | 5 years, 9 months | PDD-NOS |
| JGN | Male | 4 years, 2 months | PDD-NOS |
| JLY | Male | 3 years, 1 month | PDD-NOS |
| PES | Male | 5 years, 1 month | PDD-NOS |
| WGN | Male | 3 years, 4 months | PDD-NOS |

*Note*. Age reflects participant's age at time of first sample.
[a] Participant not included in Experiment 1.

of .94–.98. Procedural integrity (i.e., the accuracy with which the experimenter calculated the data produced by each measurement method) was evaluated as described in Experiment 1. An independent observer calculated the actual and estimated relative durations from the raw, continuously recorded duration data for 7 of the 27 samples (26%). Those calculations were compared with the experimenter's calculations. Agreement ranged from 89 to 100%, with a mean of 99%.

### 3.2. Results

When samples of stereotypy were categorized into low, moderate and high levels, MTS methods yielded more accurate estimates of the relative duration of the behavior than PIR methods. Fig. 4 shows the percent occurrence of stereotypy as measured by continuous duration recording as well as the estimates produced by the MTS and PIR methods. MTS methods both over- and under-estimated the actual duration of stereotypy across the low, moderate and high samples, while PIR overestimated the durations at all levels.

Fig. 5 depicts the degree to which the MTS and PIR estimates deviated from the actual duration recording, expressed as mean measurement error across low, moderate, and high samples. Measurement errors were calculated in the same manner as described in Experiment 1. For example, across all low samples, PIR 10 s estimates deviated from duration recording by an average of 0.8%. As Fig. 5 shows, MTS yielded more accurate estimates of the duration of stereotypy than did PIR at all levels of behavior. For low levels of stereotypy, MTS estimates deviated very little from actual duration measures with intervals of 10 s (0.8% measurement error) and 20 s (0.2% measurement error), while the deviation

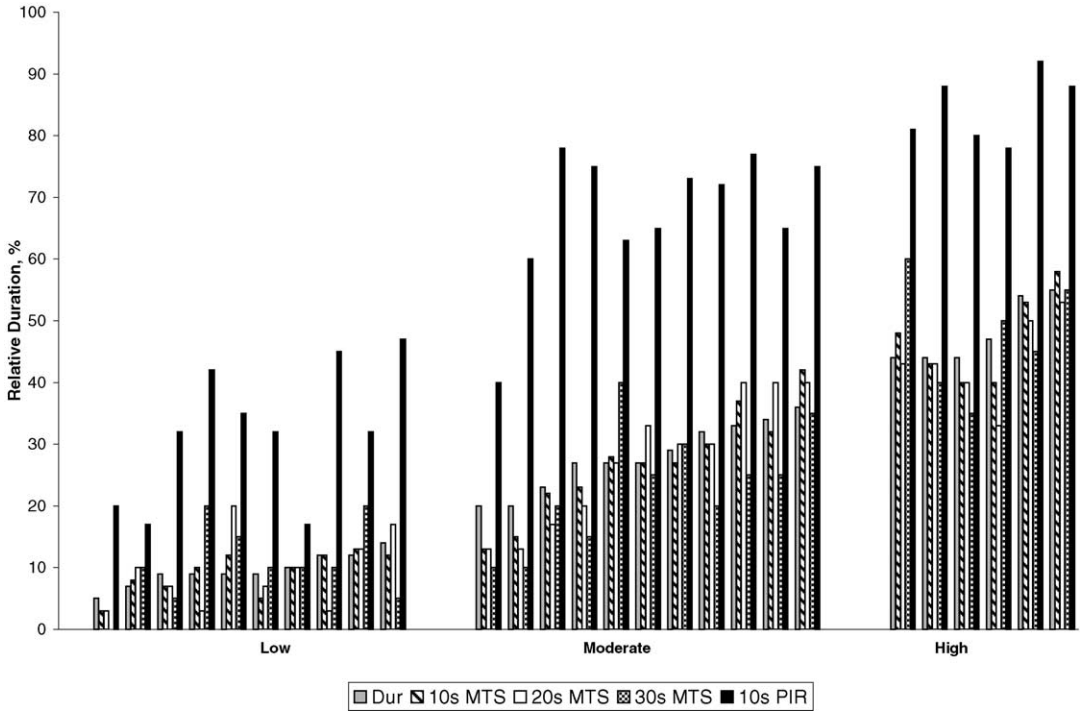



Fig. 4. Comparison of measurements of low-level stereotypy (relative duration of less than 19%), moderate-level stereotypy (relative duration between 20 and 39%), and high-level stereotypy (relative duration of more than 40%) produced by MTS 10 s (striped bars), MTS 20 s (white bars), MTS 30 s (checked bars) and PIR 10 s (black bars). The *x*-axis depicts the participant and sample. The *y*-axis depicts the relative duration of stereotypy.

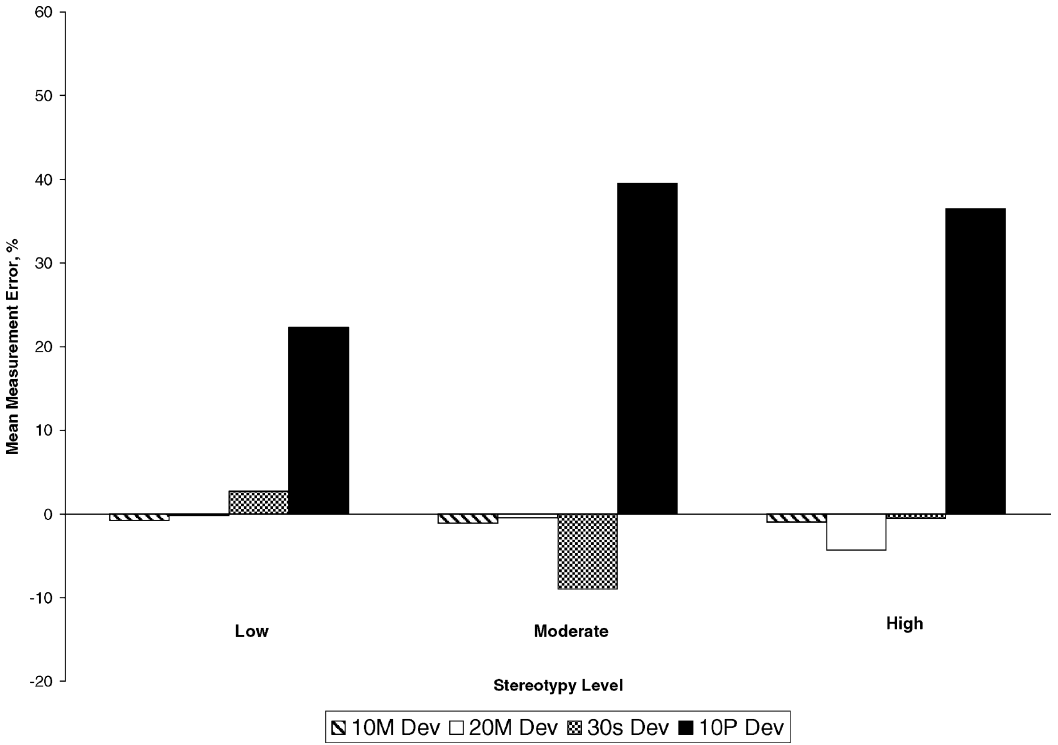



Fig. 5. Mean measurement error across low, moderate and high levels of stereotypy produced by MTS 10 s (striped bars), MTS 20 s (white bars), MTS 30 s (checked bars) and PIR 10 s (black bars). The *x*-axis depicts the level of stereotypy. The *y*-axis depicts the degree of measurement error.

Table 3
Relative percent difference from duration recording

| | Level of stereotypy | | |
|---|---|---|---|
| | Low (%) | Moderate (%) | High (%) |
| 10 s MTS | 8 | 4 | 2 |
| 20 s MTS | 2 | 2 | 9 |
| 30 s MTS | 28 | 32 | 1 |
| 10 s PIR | 229 | 141 | 76 |

*Note*. 10 s MTS: 10 s momentary time-sampling; 20 s MTS: 20 s momentary time-sampling; 30 s MTS: 30 s momentary time-sampling; 10 s PIR: 10 s partial-interval recording; low: relative duration of less than 19%; moderate: relative duration 20–39%; high: relative duration of 40% or more.

found with MTS 30 s was slightly higher (2.7%). For moderate levels of stereotypy, MTS 30 s was less accurate (9% measurement error) than MTS 10 s (1.09% measurement error) and 20 s (0.45% measurement error). For high levels of stereotypy, MTS 20 s (4.3% measurement error) was the least accurate of the three (MTS 10 s—1% measurement error; MTS 20 s—0.5% measurement error). PIR consistently overestimated the actual duration of stereotypy at all levels. The mean measurement error for low-level samples was 22%. It was 39.5% for moderate-level samples and 36.5% for high-level samples.

The relative percent difference from the continuous duration recording method for each interval method was calculated as described for Experiment 1. These values are summarized in Table 3.

## 4. Discussion

Stereotypic behavior has at least three dimensions: the frequency or rate of occurrence of episodes, the duration of episodes, and the relative duration of stereotypic behavior in any given observation period. Any or all of those dimensions may be of interest to the practitioner concerned with reducing the behavior, or the researcher asking basic questions about stereotypy. Because it is multidimensional, stereotypy can be difficult to measure accurately, but accurate measurement is the cornerstone of sound treatment as well as research. It is therefore surprising that few studies to date have examined the accuracy of direct observational methods for measuring stereotypy. Our Experiment 1 results replicated the findings of the one previous study that compared discontinuous, interval-based estimates with continuously recorded duration measures of stereotypy (Murphy & Goodall, 1980): MTS methods yielded more accurate estimates than PIR methods. Although MTS both under- and overestimated the actual level of stereotypy, the error margins were generally small, while PIR consistently overestimated by large margins. These findings are also consistent with the results of previous experiments that compared methods of measuring the duration of programmed human behavior as well as simulated behavior (Green et al., 1982; Harrop & Daniels, 1986; Powell et al., 1975, 1977).

Experiment 2 both replicated and extended previous research. When methods of measuring stereotypy occurring at low and high levels were compared, MTS 10 s was found to produce the most accurate estimates overall. For samples of moderate-level stereotypy, MTS 20 s was found to produce the most accurate estimates. For samples of high-level stereotypy, MTS 10 and 30 s produced the smallest errors (1 and 0.5%), while MTS 20 s produced a 4% error. Overall, PIR produced the largest errors with samples of moderate-level stereotypy. It is important to note that these data were collapsed across all samples. Examination of the estimates by category (low, moderate, and high levels) did not reveal any systematic error pattern for the estimates produced by MTS methods. This experiment included only 27 samples of stereotypic behavior, however; replications with additional samples would be informative.

The outcomes of the experiments reported here highlight the importance of evaluating measurement methods with behaviors that have various characteristics. The characteristics of most stereotypic behaviors rule out simple frequency or rate measures. That leaves duration, which cannot always be recorded precisely in applied settings, or interval-based methods. The latter have certain advantages; for example, they can be useful for recording multiple behaviors within an observation period. They also have distinct disadvantages, the principal one being that because they are not continuous measurement methods, they produce estimates of the true level of behavior. This makes careful calibration against continuously recorded measures critically important (cf. Johnston & Pennypacker, 1993), as the experiments reported here illustrate.

Another set of factors that must be considered in selecting direct observational measurement methods is the demands placed on those who are charged with observing and recording. PIR, by definition, requires a dedicated observer to watch the target individual throughout each of a series of short intervals, and record something at the end of each interval. As noted earlier, this requirement makes PIR impractical for many applied settings, where data are often recorded by interventionists who must carry out other tasks (such as delivering instruction) concurrently. MTS using very short intervals (e.g., 10 s) also requires a dedicated observer, because it is very difficult to engage in anything else when one must observe and record behavior every few seconds. It may be feasible, however, to record stereotypy using MTS with 30 s intervals while engaging in other activities, such as providing instruction. We have tested this possibility in our intensive preschool program for children with autism spectrum disorders, where the student to teacher ratio is 1:1. Teachers used MTS 30 s procedures to record data on stereotypy in one or two 5 min sampling periods each day, during which the teachers implemented prescribed instructional procedures at the same time. Vibrating or beeping timers signaled the onset and offset of the sampling periods, as well as the end of each interval. Analyses of a sampling of these data indicated that MTS 30 s estimates of the levels of stereotypy were quite accurate, and the teachers were able to integrate the data recording with ongoing instructional activities without difficulty. Together with the results of the experiments reported here, this suggests that MTS 30 s recording can be implemented by practitioners

without severely compromising the accuracy of data on which programming decisions are based.

Further research on the accuracy of measurement methods with various topographies of stereotypy, as well as other categories of behavior that have characteristics similar to stereotypy (e.g., some self-injurious behavior) is needed. Although this study and others clearly revealed the inadequacies of PIR for estimating the level of a variety of behaviors, and the relative superiority of MTS, the limitations of the latter method are not yet completely understood. For example, analyses of estimates of response frequencies and durations produced by MTS procedures using longer interval lengths than we used, as well as investigations of the accuracy and reliability of MTS data recorded by practitioners while they implement instructional procedures, would be enlightening. Careful analysis of the measurement errors produced by MTS with intervals of various lengths and the circumstances under which measurement error can be minimized would help practitioners and researchers select the most accurate measurement method for the behavior(s) of interest. Of course, continuous recording throughout periodic observation periods remains the preferred method for accurately measuring the frequency and duration of stereotypy and other behaviors with similar characteristics (cf. Johnston & Pennypacker, 1993).

## Appendix A. Operational definition

*Stereotypy*: Responses that have no apparent function and are not teacher-directed.

Examples include:

- rocking or swaying of torso, head, or body (full motion down and up or left and right);
- vocalizations that are not recognizable words (in normal conversational tone and volume) and are not in direct response (within 5 s) to teacher request for vocal response;
- hand flapping or other non-functional hand movements;
- non-functional rotation of hand (more than 90°) with or without materials;
- positioning hands in front of face or over ears;
- finger flicking;
- spinning objects;
- addition of objects to a line (2 or more) objects;
- licking, mouthing, or smelling objects, people or surfaces;
- manipulation of objects in a manner not appropriate to materials, not including throwing;
- non-functional closing or squinting of eyes;
- non-contextual laughing or giggling (not in response to interaction with materials or interaction with another person);
- non-functional movement of any or all body parts or objects, including jumping when paired with screaming;

- pressing or rubbing fingers or whole hand against surface or body parts;
- tapping objects.

Non-examples include:

- ''walking'' toys (e.g., cars, stuffed animals, dolls);
- whining—high pitched prolonged vocalization;
- crying;
- screaming, vocalizations above normal conversation level;
- laughing in response to tickling or joke;
- student rocking in one direction and teacher redirecting back;
- movements generated from an unobservable body part, i.e., legs wiggling but view on tape is from waist up;
- smiling that does not produce an audible sound;
- wiping face or mouth;
- incorrect responses to teacher direction (note that this is specific to the direction, e.g., only incorrect motor responses to cues meant to set the occasion for a motor response and incorrect vocal responses to cues meant to set the occasion for a vocal response are considered non-examples);
- approximations of word or request;
- rubbing eyes;
- leaning on forearm or fist;
- tapping anywhere on teachers body to get attention;
- immediate echolalia: words identical to those spoken by another person.

## Acknowledgments

The research reported here was supported by the New England Center for Children. The authors thank Susan Silvestri, Victoria Bousquet, Tracy Ide, Kate Fiske and Michael Gardenier for their assistance with data collection. The authors would like to especially thank Dr. William Ahearn for his help with data analysis.

## References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Barton, E. J., & Ascione, F. R. (1984). Direct observation. In T. H. Ollendick & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures*. Pearson Allyn & Bacon, Upper Saddle River, NJ (pp. 166–194).

Cooper, J. O. (1987). Measuring and recording behavior. In J. O. Cooper, T. E. Heron, & W. L. Heward (Eds.), *Applied behavior analysis* (pp. 59–80). Columbus: Merrill.

Epstein, L. H., Doke, L. A., Sajwaj, T. E., Sorrell, S., & Rimmer, B. (1974). Generality and side effects of overcorrection. *Journal of Applied Behavior Analysis*, *7*, 385–390.

Green, S. B., McCoy, J. F., Burns, K. P., & Smith, A. C. (1982). Accuracy of observational data with whole interval, partial interval and momentary time-sampling recording techniques. *Journal of Behavioral Assessment*, *4*, 103–118.

Guess, D., & Carr, E. (1991). Emergence and maintenance of stereotypy and self-injury. *American Journal of Mental Retardation*, *96*, 299–329.

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*, 73–76.

Johnston, J. M. (1996). Distinguishing between applied research and practice. *The Behavior Analyst*, *19*, 35–47.

Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research* (2nd ed.). Hillsdale, NJ: Erlbaum.

Jones, R. S. P., Wint, D., & Ellis, N. C. (1990). The social effects of stereotyped behavior. *Journal of Mental Deficiency Research*, *34*, 261–268.

Koegel, R. L., & Covert, A. (1972). The relationship of self-stimulation to learning in autistic children. *Journal of Applied Behavior Analysis*, *5*, 381–387.

LaGrow, S. J., & Repp, A. C. (1984). Stereotypic responding: A review of intervention research. *American Journal of Mental Deficiency*, *88*, 595–609.

Lewis, M. H., & Bodfish, J. W. (1998). Repetitive behavior disorders in autism. *Mental Retardation and Developmental Disabilities Research Reviews*, *4*, 80–89.

McEntee, J. E., & Saunders, R. R. (1997). A response-restriction analysis of stereotypy in adolescents with mental retardation: Implications for applied behavior analysis. *Journal of Applied Behavior Analysis*, *30*, 485–506.

Miltenberger, R. G., Rapp, J. T., & Long, E. S. (1999). A low-tech method for conducting real-time recording. *Journal of Applied Behavior Analysis*, *32*, 119–120.

Morrison, K., & Rosales-Ruiz, J. (1997). The effect of object preferences on task performance and stereotypy in a child with autism. *Research in Developmental Disabilities*, *18*(2), 127–137.

Murphy, G., & Goodall, E. (1980). Measurement error in direct observations: A comparison of common recording methods. *Behaviour Research and Therapy*, *18*(2), 147–150.

Powell, J., Martindale, A., & Kulp, S. (1975). An evaluation of time-sample measures of behavior. *Journal of Applied Behavior Analysis*, *8*, 463–469.

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

Pyles, D. A. M., Riordan, M. M., & Bailey, J. S. (1997). The stereotypy analysis: An instrument for examining environmental variables associated with differential rates of stereotypic behavior. *Research in Developmental Disabilities*, *18*(1), 11–38.

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.

Rojahn, J., Matlock, S. T., & Tasse, M. J. (2000). The stereotyped behavior scale: Psychometric properties and norms. *Research in Developmental Disabilities*, *21*, 437–454.

Schroeder, S. R., Rojahn, J., Mulick, J. A., & Schroeder, C. S. (1990). Self-injurious behavior. In J. L. Matson (Ed.), *Handbook of behavior modification with the mentally retarded* (2nd ed., pp. 141–180). New York: Plenum.