Journal of Applied Behavior Analysis
JOURNAL OF APPLIED BEHAVIOR ANALYSIS 2014, 47, 83–100 NUMBER 1 (SPRING)

INTERVAL SAMPLING METHODS AND MEASUREMENT ERROR: A COMPUTER SIMULATION

OLIVER WIRTH, JAMES SLAVEN, AND MATTHEW A. TAYLOR
NATIONAL INSTITUTE FOR OCCUPATIONAL SAFETY AND HEALTH

A simulation study was conducted to provide a more thorough account of measurement error associated with interval sampling methods. A computer program simulated the application of momentary time sampling, partial-interval recording, and whole-interval recording methods on target events randomly distributed across an observation period. The simulation yielded measures of error for multiple combinations of observation period, interval duration, event duration, and cumulative event duration. The simulations were conducted up to 100 times to yield measures of error variability. Although the present simulation confirmed some previously reported characteristics of interval sampling methods, it also revealed many new findings that pertain to each method's inherent strengths and weaknesses. The analysis and resulting error tables can help guide the selection of the most appropriate sampling method for observation-based behavioral assessments.

Key words: interval sampling, measurement error, momentary time sampling, observation, observational data, partial-interval recording, simulation, whole-interval recording

Observers often use interval-based sampling methods to obtain duration estimates of behavior or other target events when continuous recording methods are not suitable. Interval-based sampling methods commonly found in applied behavior analysis studies include *momentary time sampling* (MTS), *partial-interval recording* (PIR), and *whole-interval recording* (WIR) (Cooper, Heron, & Heward, 2007). All three methods involve dividing an observation period into many brief intervals in which an observer determines whether a target event occurs (Barlow, Nock, & Hersen, 2009). The total number of intervals in which the target event occurs is then counted to derive an estimate of cumulative event duration in the entire observation period. The rule for determining whether an interval is counted towards the estimate varies by method (Mayer, Sulzer-Azaroff, & Wallace, 2012). With MTS, also known as *instantaneous sampling* or *point sampling*, an interval is counted when the target event is observed at the exact moment the interval ends, regardless of whether the event occurs during any other portion of the interval. With PIR, also known as *one-zero sampling*, an interval is counted if the target event, regardless of its duration, is observed in any portion of the interval. With WIR, an interval is counted only if a target event is observed throughout the entire interval. Because interval sampling methods do not record frequency and duration of every event, these methods are inherently prone to measurement error (Gardenier, MacDonald, & Green, 2004). The major aim of the present study is to assess the magnitude and pattern of measurement error associated with these methods and describe the influence of procedural variations and external factors.

Measurement error, in the context of observation-based interval sampling, is inaccuracy in the recorded data caused by the sampling strategy or other features of the observational method (G. Murphy & Goodall, 1980). Many studies have revealed consistently that the magnitude and

James Slaven is now at Indiana University–Purdue University Indianapolis.

We thank Matthew Weaver at Mercyhurst University for his review of an early draft of this manuscript. The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the National Institute for Occupational Safety and Health.

Address correspondence to Oliver Wirth, CDC/NIOSH, 1095 Willowdale Road MS 2027, Morgantown, West Virginia 26505 (e-mail: oaw5@cdc.gov).

doi: 10.1002/jaba.93

direction of measurement error are influenced by several procedural variables, such as the sampling method (Alvero, Struss, & Rappaport, 2007; Gardenier et al., 2004; Devine, Rapp, Testa, Henrickson, & Schnerch, 2011; Green, McCoy, Burns, & Smith, 1982; G. Murphy & Goodall, 1980; Powell, Martindale, Kulp, Martindale, & Bauman, 1977; Rapp, Colby-Dirksen, Michalski, Carroll, & Lindenberg, 2008; Simpson & Simpson, 1977), the number or duration of observation intervals (Alvero, Rappaport, & Taylor, 2011; Alvero et al., 2007; Brittle & Repp, 1984; Devine et al., 2011; Dunbar, 1976; Leger, 1977; Powell et al., 1977; Rapp et al., 2008), and the duration of the observation period (Devine et al., 2011; Mansell, 1985; Mudford, Beale, & Singh, 1990) and some event-related factors, such as frequency (Alvero et al., 2007, 2011; Gardenier et al., 2004; Green et al., 1982; McDowell, 1973; G. Murphy & Goodall, 1980; Powell et al., 1977) and duration (Green et al., 1982; G. Murphy & Goodall, 1980; Sanson-Fisher, Poole, & Dunn, 1980).

Many studies of interval sampling methods have consisted of computer-based simulations (Ary & Suen, 1983; Engel, 1996; Harrop & Daniels, 1986; Jacobsen & Wiggins, 1982; Kearns, Edwards, & Tingstrom, 1990; Milar & Hawkins, 1976; Powell, 1984; Powell & Rockinson, 1978; Repp, Roberts, Slack, Repp, & Berkler, 1976; Rhine & Ender, 1983; Rojahn & Kanoy, 1985; Tyler, 1979; Wilson, Jansen, & Krausman, 2008). Simulations offer researchers several advantages over analyses of observational records obtained from real-life events in the field. For instance, simulations give researchers control over event-related parameters throughout the observation period, such as frequency, duration, rate, and distribution. In a simulation, the occurrence of events can be precisely programmed to allow parametric and factorial analyses of their effects on measurement error. Furthermore, an evaluation of sampling methods with a simulation removes error attributable to the human observer. *Observer error* is inaccuracy in the data that is introduced by the observer and can be indirectly attributed to a range of variables, such as age, training, history, vigilance, reaction time, motivation, and stimulus discriminability (Green et al., 1982; Mudford et al., 1990; M. J. Murphy & Harrop, 1994; Repp et al., 1976; Saudargas & Zanolli, 1990; Taylor, Skourides, & Alvero, 2012; Tyler, 1979). Measurement error is studied more efficiently when it is disentangled from the confounding effects of observer error.

Simulation studies of interval sampling methods have provided several consistent findings that are well known by applied behavior analysts. For instance, it is widely believed that MTS yields small error magnitudes and no consistent or systematic bias towards either overestimation or underestimation of event occurrences, whereas PIR yields greater error and a consistent bias towards overestimation of event occurrences (Harrop & Daniels, 1986; Tyler, 1979). Simulation studies have also revealed with either MTS or PIR that measurement error increases when the duration of sampling intervals increases (Kearns et al., 1990; Rhine & Ender, 1983; Tyler, 1979), whereas error decreases when the duration of target events increase (Harrop & Daniels, 1986; Rhine & Ender, 1983). Although no simulation studies have directly assessed measurement error associated with WIR, it is reported that WIR yields a consistent bias towards underestimation of event occurrence that increases with interval duration (Alvero et al., 2007; Powell et al., 1977).

Simulation studies of interval sampling methods also have produced several inconsistent or discrepant findings. For example, PIR can be biased towards underestimation of event occurrences by approximately 35% (Repp et al., 1976), whereas MTS can be biased towards overestimation (G. Murphy & Goodall, 1980). Some findings suggest that the direction of MTS error depends on *cumulative event duration*, which is usually quantified as a percentage of the observation period. For example, Tyler (1979) reported that MTS error is biased towards

underestimations when cumulative event duration is low and overestimations when cumulative event duration is high. On the other hand, Kearns et al. (1990) reported that MTS error does not vary systematically across a wide range of cumulative event durations, whereas Wilson et al. (2008) reported that MTS error decreases as cumulative event duration increases. MTS also was shown to produce low levels of unbiased error no matter the interval or event duration in one study (Ary & Suen, 1983) and decreased error with shorter event durations in another study (Rhine & Ender, 1983).

Procedural variations also hinder comparisons across studies, further clouding our understanding of measurement error. For example, it is difficult to compare and contrast results of studies that used atypical sampling procedures, such as the insertion of intermittent recording intervals with PIR (Harrop & Daniels, 1986; Repp et al., 1976) or different ranges of sampling intervals, such as 5 s to 60 s in one study (Tyler, 1979) and 30 s to 1,200 s in another (Kearns et al., 1990). In other studies, cumulative event durations, event durations, and event rates were either confounded or not reported (Kearns et al., 1990; Rhine & Ender, 1983; Tyler, 1979; Wilson et al., 2008). Finally, measurement error was quantified differently across several studies, with some reporting absolute error values (i.e., difference between estimated and actual event durations) and some reporting relative error values (i.e., difference expressed as a percentage of actual event durations).

A more thorough review of the relevant research literature would reveal additional discrepant findings that pertain to interval sampling methods and measurement error, but a consistent conclusion will emerge nevertheless: No study has sufficiently taken into account the multifactorial nature of measurement error. Although results of a few studies hint at the magnitude and direction of error being dependent on the levels of other factors, in most studies measurement error was evaluated across a narrow range of parameters or without a sufficient account of possible interaction effects. As a result, limitations remain in our understanding of the error produced by interval sampling methods.

The purpose of the present study was to conduct a more thorough computer-based simulation study of measurement error associated with MTS, PIR, and WIR. A computer program was written to simulate the application of these different sampling methods under a wider range of procedural and event-related parameters, including interval duration, duration of the observation period, event duration, and cumulative event duration. The reliability of the results was also assessed across up to 100 iterations of simulation. We hypothesized that the general magnitude and direction of measurement error would be a direct function of the sampling method, consistent with previous studies. We also hypothesized that the specific patterns of measurement error across the wide range of procedural and event-related parameters would reveal the effects of these parameters to be highly interdependent.

METHOD

Variables and Parameters

Table 1 shows the tested procedural and event-related parameters. All combinations of the variables and parameters were tested with three different sampling methods: MTS, PIR, and WIR. MTS sampled events only during the last second of each interval, and PIR and WIR sampled events continuously throughout each interval without additional recording intervals. Other procedural parameters included a wide range of interval durations (15 s to 450 s) and observation periods (1 hr to 8 hr). Cumulative event duration was chosen as a variable, rather than event frequency, to be consistent with previous simulation studies. Cumulative event durations from 1% to 100% of the observation period were tested, except when it was not possible. For example, with 256-s events and a

Table 1
Procedural and Event-Related Parameters That Were Tested in the Simulation

Parameters	Levels tested
Sampling method	MTS, PIR, WIR
Interval duration (s)	15, 30, 60, 120, 240, and 450
Observation period (hr)	1, 4, and 8
Cumulative event duration (%)	1 to 100 (in 1% increments)
Event duration (s)	1, 2, 4, 8, 16, 32, 64, 128, and 256

3,600-s observation period, cumulative event durations below 8% were not tested because the lowest possible percentage, rounded to the nearest whole number, is 8% (3,600 s ÷ 256 s × 100 = 7.11%). Event durations ranged from 1 s to 256 s.

Procedure

To simulate the application of the different interval sampling methods under different combinations of procedural and event-related parameters, a computer program was written using Microsoft Visual Basic. The basic structure of the program consisted of nested for next loops in which every combination of observation period, event duration, and cumulative event duration was passed to a subroutine to generate an array of randomly distributed target events. Elements of each *event array* consisted of *ones*, depicting the occurrence of a target event, and *zeros*, depicting the absence of the target event. Each element of the array represented 1 s of an observation period. Thus, for example, an event array for a 1-hr observation period consisted of 3,600 elements.

The location and distribution of target events were determined by using the random number generator in Visual Basic. For each level of cumulative event duration from 1% to 100%, the appropriate number of events was added to the array. For example, if given an observation period of 3,600 s, an event duration of 4 s each, and a cumulative event duration of 30% of the observation period, then it would be necessary to add at least 270 events to the event array (3,600 s × 30% ÷ 4 s = 270). Each event was assigned to the array at a randomly determined location. When an event was assigned, the zeros in that portion of the array, starting with the first element of the location, were replaced with ones; however, an event was assigned to a location only if the starting element did not already contain an existing event. If the selected location already contained an event, then another location was randomly selected. This iterative procedure was repeated as many times as necessary to assign all events and yield the desired cumulative event duration as a percentage of the observation period. With this procedure, it was possible that two or more events could partially overlap leaving a succession of ones in the array that exceeded the desired event duration. As expected, overlapping events occurred more often as the actual cumulative event duration increased; however, events were never less than the desired event duration, except in two conditions: (a) when the desired cumulative event duration was achieved after only a portion of the final event had been assigned, or (b) when events were assigned near the end of the array. In the latter condition, for example, if a 16-s event was assigned to start at the 3,595th element of a 3,600-element array, then only 6 s of the event would be assigned to the array. This strategy of event assignment, instead of a more random strategy, resulted in better control over event duration by maximizing the discreteness of each event.

After the event array was generated, it was passed to a subroutine that simulated the application of the MTS, PIR, and WIR sampling methods. The same event array was used for every level of interval duration to ensure that the duration estimates and error measures obtained for every combination of sampling method and interval duration were based on the same distribution of events; this requirement is an improvement over past simulation studies in which comparisons between sampling methods

were made across different event distributions (e.g., Powell, Martindale, & Kulp, 1975). This procedure was repeated once for every level of observation period from 1 hr to 8 hr. Finally, the entire simulation was conducted 100 times to yield measures of error variability across multiple iterations.

Dependent Measures

The simulation program yielded four measures for every combination of procedural and event-related parameters. The *actual cumulative event duration* was expressed as a percentage of the total observation period. The *estimated cumulative event duration* was obtained following each application of the sampling method and was expressed as a percentage of the total observation period. *Absolute error* was calculated by subtracting the actual cumulative event duration from the estimated cumulative event duration. *Relative error* was calculated by dividing the absolute error by the actual cumulative event duration, and then converting the result to a percentage. Positive error values represented overestimates of event duration, and negative error values represented underestimates of event duration.

RESULTS

Comparison of Actual and Estimated Cumulative Event Durations

Figures 1 and 2 show the duration estimates and relative error obtained from the simulation of MTS within a 1-hr observation period. Estimated cumulative event duration is plotted as a function of the actual cumulative event duration for each interval duration and event duration; results are presented from one iteration of the simulation. In general, Figures 1 and 2 show that MTS yielded both underestimates and overestimates of actual cumulative event duration. The difference between estimated cumulative event duration and actual cumulative event duration (i.e., absolute error) tended to increase with increasing interval durations across all levels of actual cumulative event duration. Absolute error appeared to be small or negligible when interval duration was nearly equal to or less than the event duration. The actual cumulative event durations had a much greater effect on the relative errors of estimates. As shown on the right axes of Figures 1 and 2, the magnitude of relative error tended to decrease with increasing actual cumulative event durations; however, the magnitude of this decrease appeared to be dependent on event duration and interval duration. For instance, relative error appeared to be either small or negligible when interval duration was similar to or less than the event duration. Conversely, relative error appeared to be large when interval duration exceeded event duration, especially at lower levels of actual cumulative event duration.

Figure 3 shows the duration estimates and relative error obtained from the simulation of PIR within a 1-hr observation period. In general, the results reveal the characteristic tendency of PIR to overestimate cumulative event durations systematically. In addition, the magnitude of overestimation appears to be a direct function of interval duration and event duration. For instance, the magnitude of overestimation increased when interval duration increased and event duration decreased. Across actual cumulative event durations of 1% to 100%, the patterns of duration estimates tended to be curvilinear. In general, as interval duration increased, the curvilinear pattern appeared to be more pronounced. Furthermore, increases in event duration tended to shift the patterns of duration estimates from curvilinear to more linear (or toward actual cumulative event durations).

Relative error of duration estimates with the PIR method showed a systematic pattern across actual cumulative event durations, event durations, and interval durations. In general, relative error decreased with increasing cumulative event durations and event durations, and it increased with increasing interval duration. In some cases, relative error was large (up to 10,000% of the observation period), especially when event

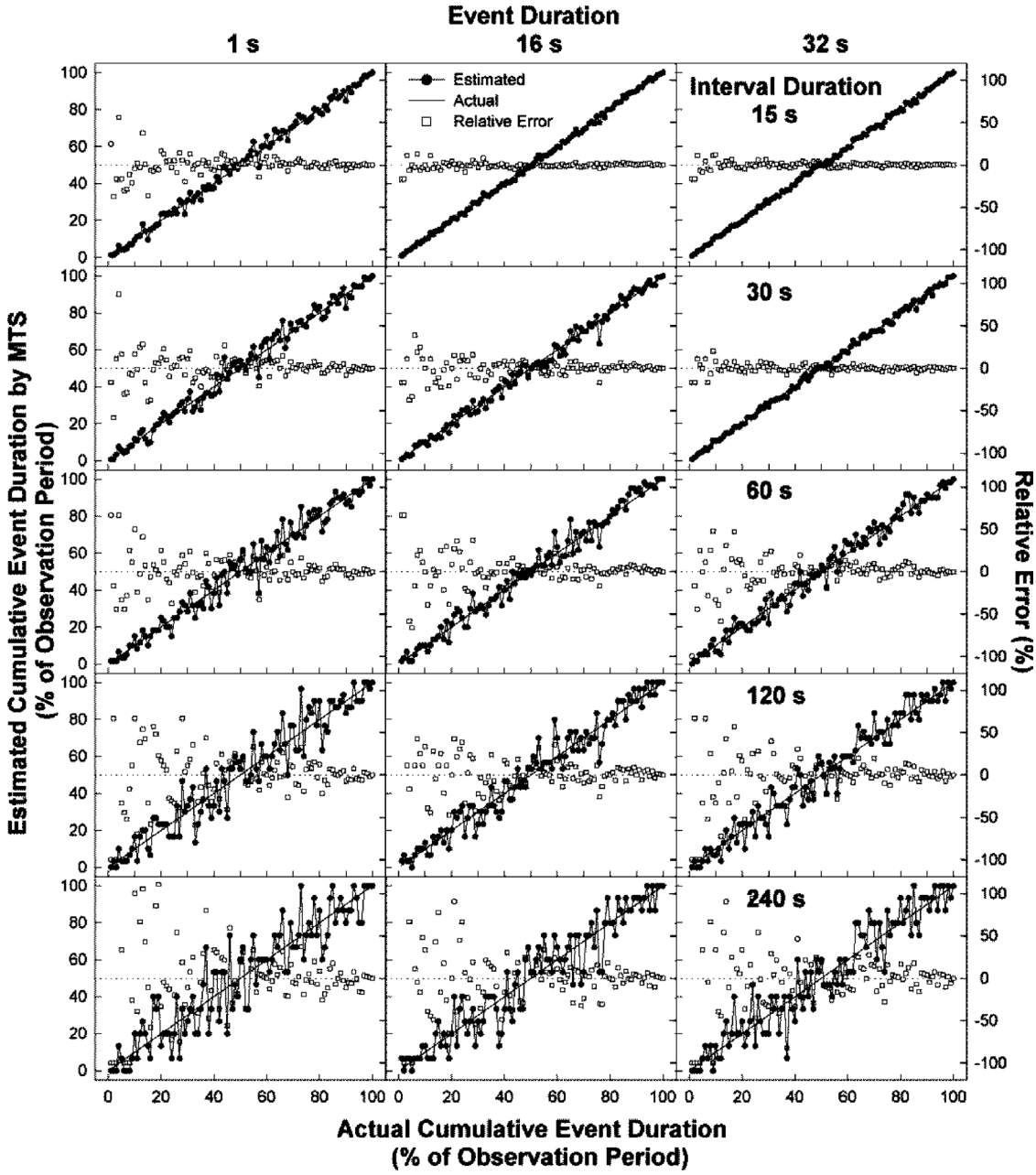

Figure 1. Estimated and actual cumulative event durations (*y* axis, left) and relative error (*y* axis, right) obtained from one iteration of the MTS method plotted across different actual cumulative event durations (*x* axis), event durations of 1 s, 16 s, and 32 s (left, middle, and right), and interval durations (top to bottom). Relative error was the difference between estimated and actual cumulative event durations divided by actual cumulative event durations. Cumulative event durations and relative error are expressed as the percentage of a 1-hr observation period. Positive error values represent overestimates, and negative values represent underestimates.

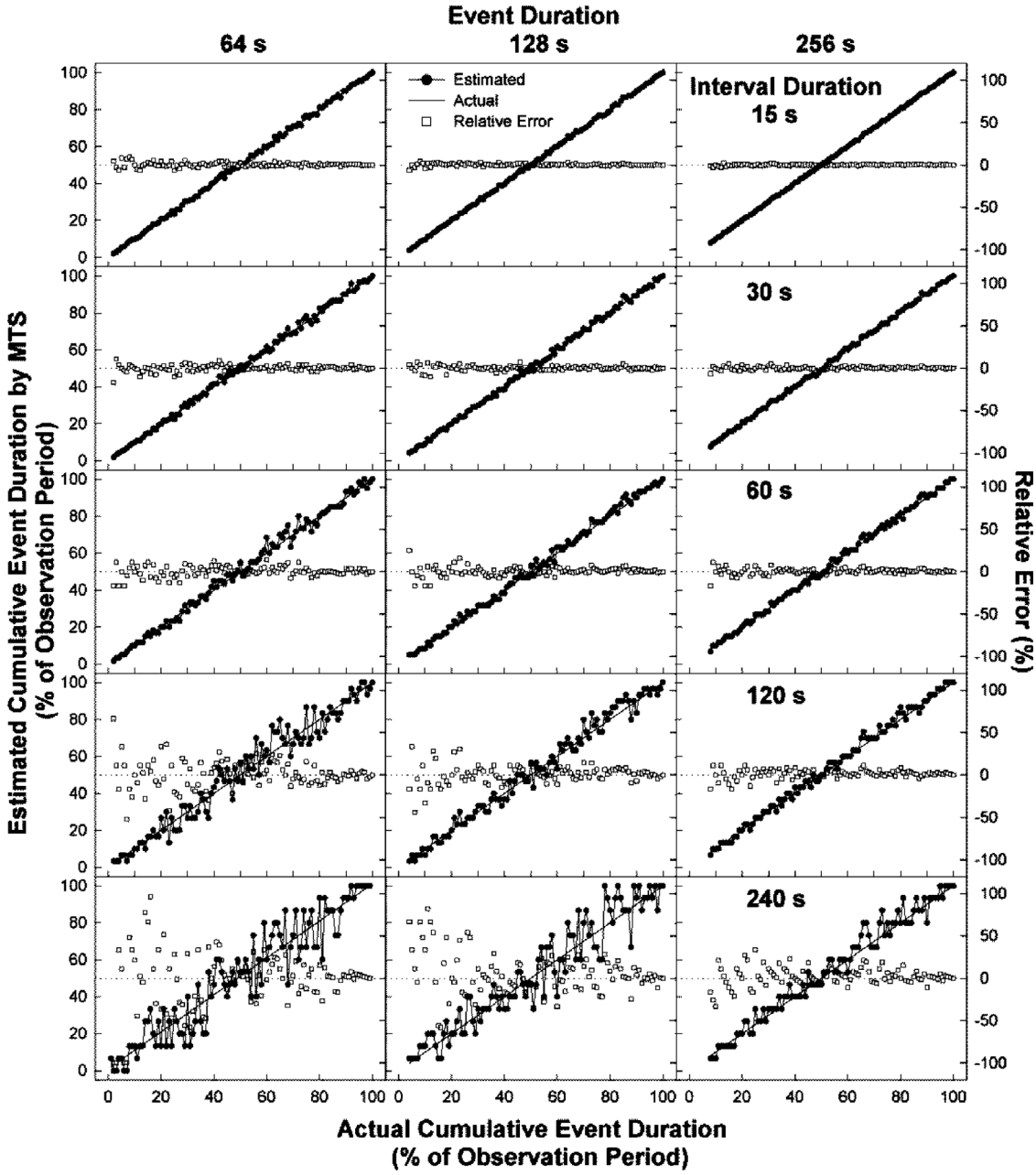

Figure 2. Continued from Figure 1. Estimated and actual cumulative event durations (*y* axis, left) and relative error (*y* axis, right) obtained from one iteration of the MTS method plotted across different actual cumulative event durations (*x* axis), event durations of 64 s, 128 s, and 256 s (left, middle, and right), and interval durations (top to bottom). Other details are the same as in Figure 1.

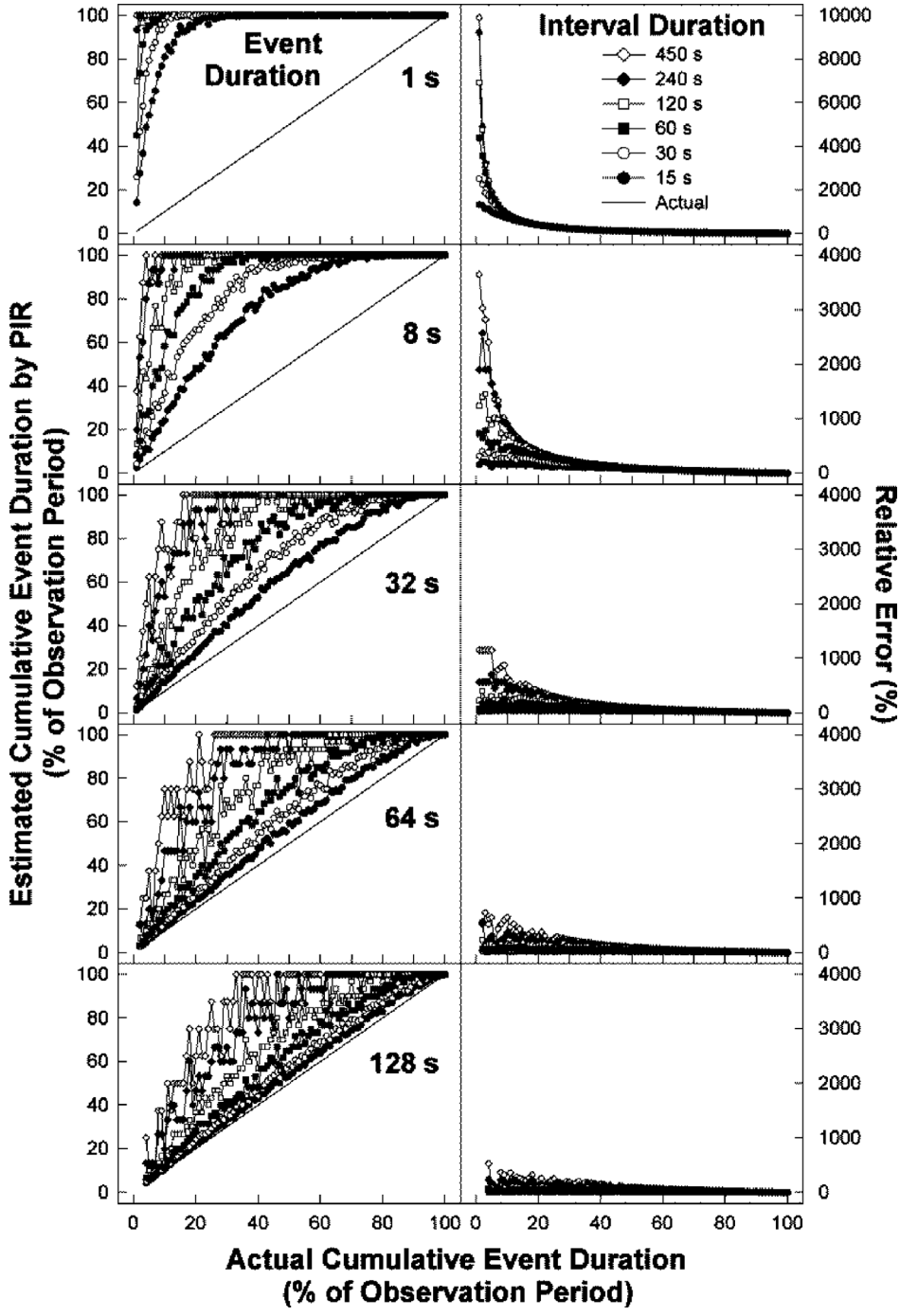

Figure 3. Estimated and actual cumulative event durations (*y* axis, left) and relative error (*y* axis, right) obtained from one iteration of the PIR method plotted across different actual cumulative event durations (*x* axis) and different event durations (top to bottom). Multiple plots in each panel show the effects of different interval durations and the actual cumulative event durations. Other details are the same as in Figure 1.

durations were short and interval durations were long.

Figure 4 shows the duration estimates and relative error obtained from the simulation of WIR within a 1-hr observation period. The results show a systematic pattern of error similar to PIR, except that event durations were underestimated and the magnitude of error was slightly greater. Similar to PIR error patterns, the magnitude of underestimation with WIR appears to depend on interval duration, actual cumulative event duration, and event duration. As interval duration increased and event duration decreased, the magnitude of underestimation increased. The curvilinear patterns of underestimation between 1% and 100% of the actual cumulative event durations were consistent across the interval durations and event durations, except when event duration far exceeded interval duration. The overall patterns of relative error associated with the WIR method were similar to those with PIR; however, unlike the PIR method, the WIR method imposes a floor effect on the magnitude of relative error. Because the maximum magnitude of absolute error cannot be greater than the actual event duration, relative error varied from 100% to 0%. (Note the different *y*-axis scale on the right side of Figure 3.) In general, relative error decreased with increasing cumulative event durations and event durations and increased with increasing interval duration.

Duration of Observation Period

Figure 5 shows how different observation periods affected estimates of cumulative event durations when event duration was held constant at 32 s. Increasing the duration of the observation period from 1 hr to 8 hr systematically reduced the variability of error with all three methods. This reduction in error variability was also associated with a reduction in the overall magnitude of error with MTS but not with PIR and WIR. Increasing the observation period had similar effects with other event durations not shown.

Repeated Sampling across Multiple Iterations

Figure 6 shows the effects of averaging the estimates obtained across 100 different iterations of the simulation in a 1-hr observation period. Results for only 32-s event durations are presented, but they are representative of the effects seen with other event durations. In general, the results demonstrate the reliability of the duration estimates across multiple repetitions of the simulation, each with a different random distribution of events. The small standard deviations associated with the duration estimates after 100 iterations further support the general conclusion that the duration estimates derived from the MTS, PIR, and WIR methods are highly repeatable, particularly with shorter interval durations.

Absolute Error

Because the duration estimates derived from MTS, PIR, and WIR methods were systematic, tables were generated to show the range of absolute errors obtained in the present simulation across the many combinations of procedural and event-related parameters. These tables, one for each method, can be obtained from the first author or by accessing the Supporting Information for this article on the Wiley Online Library. These tables could serve as a reference for researchers and practitioners who rely on interval sampling methods. They show the median, minimum, and maximum absolute error obtained from the present simulation, given different levels of estimated or actual cumulative event durations. Given the expected *actual* cumulative event duration, one set of tables can be consulted in advance of designing or implementing an interval-based observation study to better understand and account for variables that might affect error in cumulative duration estimates. Accordingly, researchers will be able to determine a priori the possible

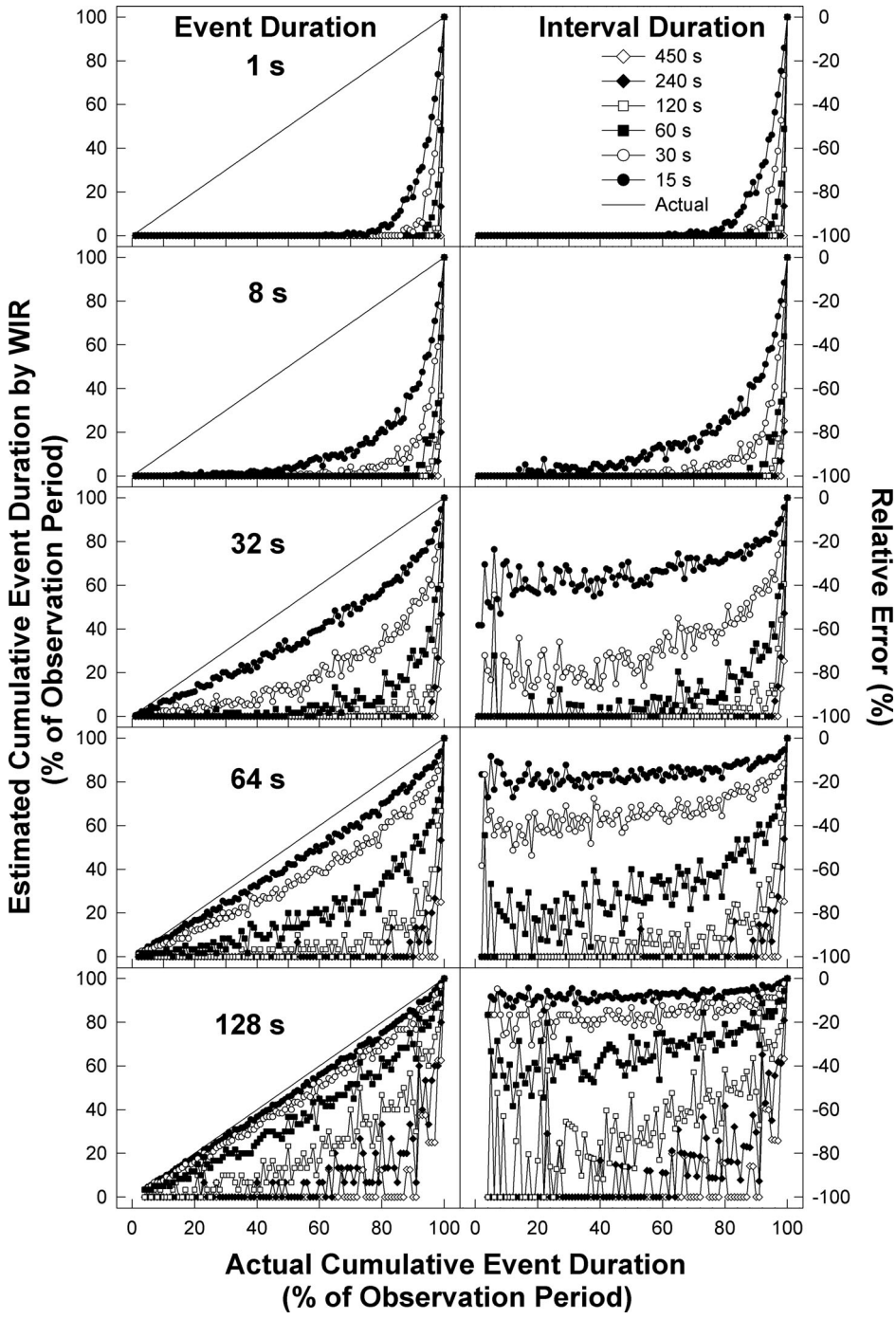

Figure 4. Estimated and actual cumulative event durations (*y* axis, left) and relative error (*y* axis, right) obtained from one iteration of the WIR method. Other details are the same as in Figure 3.

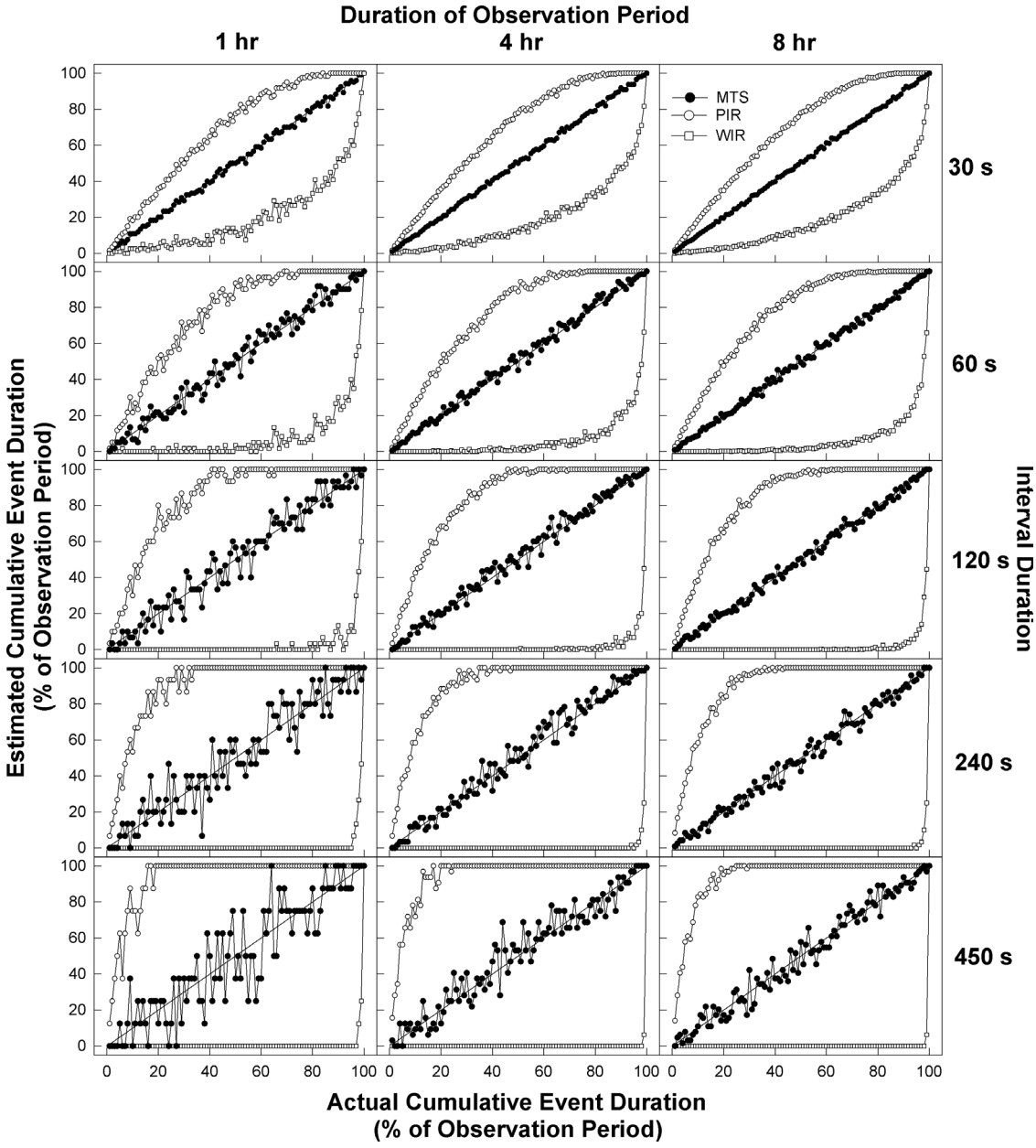

Figure 5. Estimated and actual cumulative event durations obtained from one iteration of the MTS, PIR, and WIR methods plotted in each panel. The effects of 1-hr, 4-hr, and 8-hr observation periods are shown from left to right. The effects of 30-s to 450-s interval durations are shown from top to bottom. Event duration was 32 s.

magnitude and direction of error that can be expected for a given set of parameters that most closely approximates their own application. Because error is highly dependent on event-related parameters, the selection of an interval method should also take into account any anticipated changes in the occurrence of events after an intervention. The sensitivity of a method

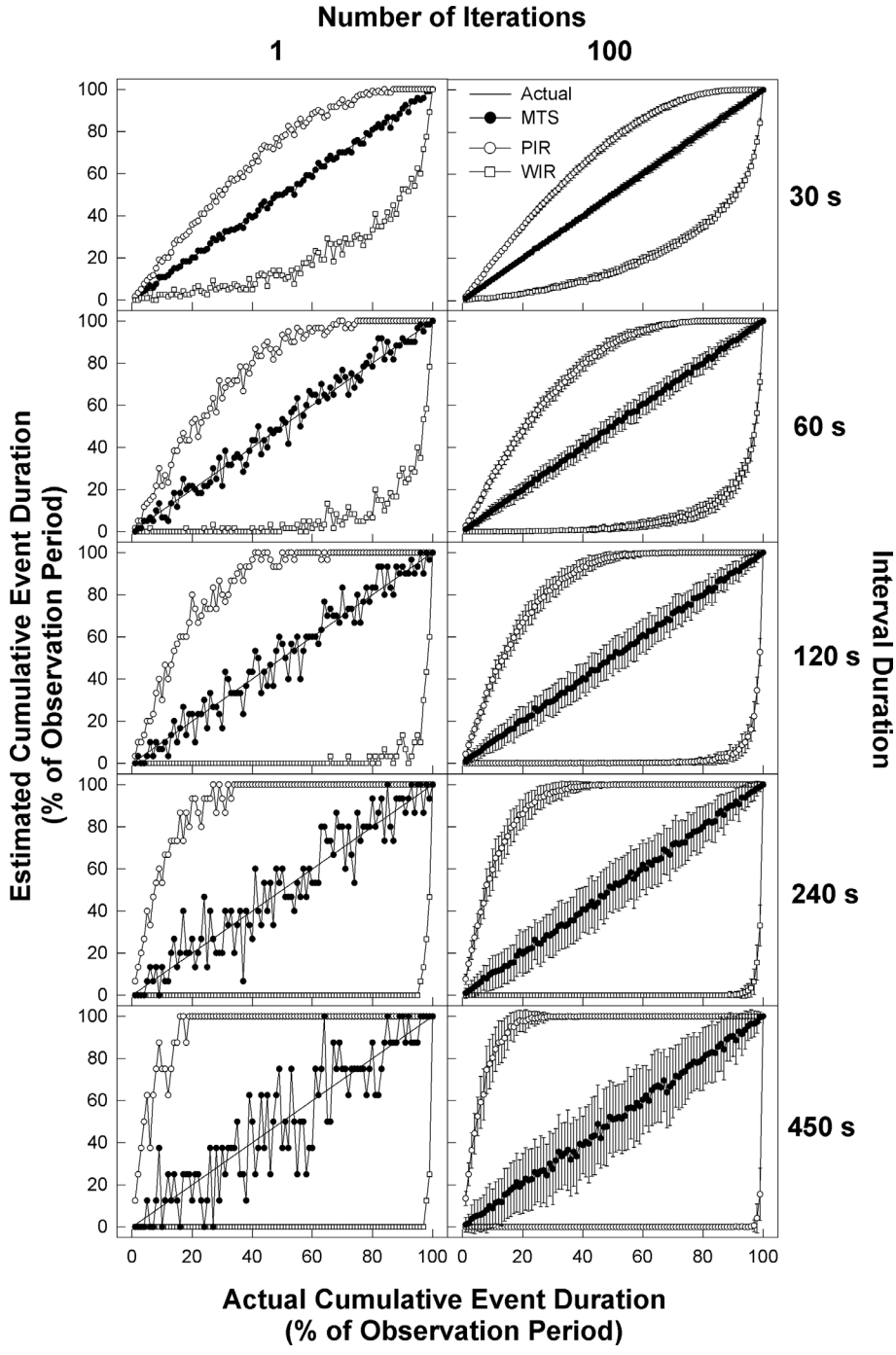

Figure 6. Estimated and actual cumulative event durations obtained from 1 and 100 iterations of the MTS, PIR, and WIR methods plotted in each panel. The effects of increasing the number of iterations from 1 to 100 are shown from left to right. The effects of 30-s to 450-s interval durations are shown from top to bottom. Event duration was 32 s. Estimates in the right panels are means and standard deviations (error bars) across the 100 iterations.

to detect a change would be substantiated if the range of possible duration estimates derived from the method across the study phases neither exceed nor overlap the intervention's anticipated effect size. Another set of tables can be consulted for post hoc analysis of *estimated* cumulative event durations after obtaining observational data or to evaluate empirical results of an already published study.

Our analysis revealed an unexpected finding with MTS. Examination of the median error values across different categories of actual cumulative event durations showed that MTS is somewhat biased towards underestimation (negative medians) with low cumulative event durations and overestimation (positive medians) with high cumulative event durations. This bias is most evident with longer interval durations and shorter event durations.

DISCUSSION

Consistent with many previous investigations of interval sampling methods, the present simulation study showed that measurement error associated with MTS, PIR, and WIR was highly systematic and dependent on interactions among interval duration, event duration, and cumulative event duration. The key findings can be summarized as follows: (a) PIR consistently overestimates and WIR consistently underestimates cumulative event duration; (b) MTS sometimes overestimates and sometimes underestimates cumulative event duration (error is not biased in either direction); (c) absolute error magnitude tends to be smaller with MTS and greater with PIR and WIR; (d) absolute error variability tends to be greater with MTS and smaller with PIR and WIR; (e) in general, absolute and relative error decrease when interval duration decreases; (f) absolute and relative error are minimal when interval duration is less than or equal to event duration; (g) absolute error decreases when observation period increases; and (h) assuming a random distribution of target events, patterns of error in a given measurement system are highly systematic, which makes it possible to estimate expected error a priori or obtained error post hoc.

The general hypothesis, that the magnitude and direction of measurement error are dependent on the sampling method used, was well supported. PIR and WIR produced, in general, greater magnitudes of error than MTS, consisted with previous studies (Alvero et al., 2007; Harrop & Daniels, 1986; Tyler, 1979). Also consistent was the finding that MTS produced mostly unbiased estimates of cumulative duration, whereas PIR tended to overestimate cumulative event duration and WIR tended to underestimate cumulative event duration. Similar effects with interval duration and event duration were also revealed. For example, regardless of the method used, error decreased when interval duration was decreased (Alvero et al., 2007; Kearns et al., 1990; Rhine & Ender, 1983; Tyler, 1979) and event duration was increased (Harrop & Daniels, 1986; Rhine & Ender, 1983). Error was reduced especially when event duration approached or exceeded interval duration, as reported by Ary and Suen (1983).

Some effects of sampling method, interval duration, event duration, and cumulative event duration differ from previous findings. However, instead of finding discrepant results, the present simulation study provides a more complete account of each variable's influence on error by assessing all factorial combinations of the variables. Furthermore, the present findings expand our current understanding of the relative strengths and weaknesses of MTS, PIR, and WIR.

Perhaps the most widely reported conclusion from previous studies of interval sampling methods is that MTS is preferred over either PIR or WIR because it produces unbiased or unsystematic patterns of error (G. Murphy & Goodall, 1980; Powell et al., 1977; Suen, Ary, & Covalt, 1991). The present study provided evidence that MTS is biased towards underestimation when cumulative event durations are low and overestimation when cumulative event

durations are high. The magnitude of underestimation or overestimation with MTS sometimes exceeded 50% of the observation period.

Another reported conclusion from previous studies is that PIR and WIR produce excessive and inconsistent error (Powell, 1984; Powell et al., 1977). Indeed, Powell (1984) once described PIR as "equivalent to using a yardstick that aperiodically expands and contracts, and does so without the user's awareness" (p. 218). For similar reasons, many experts do not recommend PIR or WIR at all (Alvero et al., 2007; G. Murphy & Goodall, 1980; Saudargas & Zanolli, 1990). Contrary to these sentiments, the present findings reveal some respectable characteristics of PIR and WIR. For example, unlike previous characterizations of PIR, the error associated with both PIR and WIR was found to be systematic and, in many instances, predictable across different interval durations, event durations, and cumulative event durations. For instance, the present study revealed that when PIR (or WIR) estimates of cumulative event duration were values other than 0% and 100%, then there was little variability in the estimates, even though the magnitude of overestimation (or underestimation) tended to be great. Thus, with some combinations of factors, clinicians and researchers can expect to obtain reliable estimates with PIR and WIR, and, if the obtained estimates are adjusted to take into account the expected magnitude of error, the results can be accurate. Such adjustments are not possible with MTS because of the notable amount of variability in error. In summary, using Powell's (1984) analogy, our results suggest that it is MTS that is equivalent to using a yardstick that aperiodically expands and contracts because it can yield overestimations and underestimations, whereas PIR is equivalent to a yardstick that is consistently longer than standard because it almost always yields overestimations. Conversely, WIR is equivalent to a yardstick that is consistently shorter than standard because it consistently yields underestimations.

Regarding the influence of interval duration, many researchers have offered general but often disparate recommendations. For instance, several researchers advised caution when using MTS with interval durations greater than 60 s (Brittle & Repp, 1984), 120 s (Rhine & Ender, 1983), and 240 s (Gunter, Venn, Patrick, Miller, & Kelly, 2003). Many researchers have acknowledged that, in practice, the selection of interval duration is often arbitrary (Rojahn & Kanoy, 1985; Sanson-Fisher et al., 1980), based on convenience of the observer (Mansell, 1985), or otherwise rarely based on achieving a specific level of accuracy (Mansell, 1985; Rojahn & Kanoy, 1985; Sanson-Fisher et al., 1980). The present simulation results seem to shed more light on these confusing recommendations by demonstrating robustly that error was dependent on the relation between interval duration and event duration. When interval duration was less than or closely approximated event duration, error magnitude and variability were minimal, even when interval duration was either short and frequent or long and infrequent. Conversely, when interval duration exceeded event duration, error magnitude and variability increased appreciably. This relation between interval and event duration was seen consistently across all three sampling methods. Therefore, it seems reasonable to conclude any one-dimensional rule of thumb for selecting interval duration will be misleading. This conclusion is not entirely new; many researchers have concluded that error is problematic when interval durations are long and event durations are short (Ary, 1984; Ary & Suen, 1986; Choi, Nam, & Lee, 2007; G. Murphy & Goodall, 1980; Rhine & Ender, 1983; Saudargas & Zanolli, 1990; Suen & Ary, 1986); however, results of the analyses reported herein and in the accompanying error tables reveal the nature of this interaction more precisely and comprehensively across a wide range of parameters.

Although there is a dearth of evidence regarding the effects of observation period on error (Mudford et al., 1990), the findings from

the present analysis with MTS are consistent with those of Devine et al. (2011) by showing that an increase of observation period reduced the magnitude and variability of error. The present analysis showed that the impact of observation period differed across sampling method. With PIR and WIR, increasing the observation period reduced the variability of error for any given set of parameters; however, it did not affect the magnitude of error. The overall pattern of error magnitude and variability can be expected to extend to observation periods that are less than 60 min. For example, observation periods of 10 to 15 min, commonly used in behavior-analytic research, would be expected to generate greater error variability than shown here with 60-min periods and all three methods, but error magnitude would be expected to be greater only with MTS.

The calculation and reporting of error statistics obtained from multiple iterations of the simulation were useful for demonstrating the reliability of the error patterns across the multiple combinations of the procedural and event-related parameters. The present findings show that statistical averaging of estimates obtained from multiple iterations of PIR and WIR did not change the magnitude or pattern of error but, instead, showed the error patterns to be highly reliable across the various parameters. To demonstrate a reliable pattern of error, the present study used 100 iterations, but fewer may have been adequate. Similar studies reported that iterations of 10, 15, and 20 generated data with marginal differences (Wilson et al., 2008) and no significant differences with iterations of 20, 50 and 100 (Rhine & Ender, 1983); therefore, iterations above 20 appear to have diminishing returns.

Despite several advantages, simulation studies have some disadvantages. First, generality of the results to real-world settings is limited. The controlled manipulation of variables removed variability that would be encountered in the field. Second, the evaluation of event duration was confounded with cumulative event duration; as cumulative event duration increased, average event durations also increased because many events tended to overlap or run in series. Third, the distribution of events in real-world settings can vary widely, but the present study evaluated only a random distribution. It is unclear how other distributions would affect the patterns of error.

Several important topics remain to be studied. For example, variable or random sampling strategies have not been thoroughly evaluated. These strategies can be more user-friendly and still provide levels of error similar to fixed-interval methods (Test & Heward, 1984). More research also is needed to understand how cyclical and irregular temporal patterns of events and non-events affect the accuracy of interval sampling methods, especially with shorter observation periods that are commonly used by applied behavior analysts. Furthermore, future research should take into account duration of both occurrences and nonoccurrences of events, because studies have shown that error can be exacerbated when interval duration exceeds the duration of nonoccurrences (Ary & Suen, 1986; Suen & Ary, 1986).

Also relevant are additional studies to determine the sensitivity of sampling methods for detecting changes in duration events (e.g., Bartlett, Rapp, & Henrickson, 2011; Carroll, Rapp, Colby-Dirksen, & Lindenberg, 2009; Devine et al., 2011; Meany-Daboul, Roscoe, Bourret, & Ahearn, 2007; Rapp et al., 2007, 2008). These studies reveal how various combinations of sampling and event-related parameters give rise to either false positives or false negatives in experimenter judgments of behavior change and intervention effectiveness. A synthesis of these approaches with more direct assessments of measurement error will further advance our understanding of the strengths and limitations of sampling methods.

Finally, practical tools are needed to help researchers and practitioners determine the magnitude and direction of error they can expect

to encounter either for a priori or post hoc considerations. As argued by others (e.g., Rojahn & Kanoy, 1985), we believe the systematic and reliable nature of error patterns supports the feasibility of developing these tools. Examples might include a mathematical equation (Wilson et al., 2008), process algorithm (Suen & Ary, 1986), decision tree or flow chart (Fiske & Delmolino, 2012), or error tables, such as those provided herein. For example, Wilson et al. (2008) derived an equation based on sampling theory to identify the number of observations needed to achieve a given level of error (see also Scheaffer, Mendenhall, Ott, & Gerow, 2012). Although the equation used by Wilson et al. does not take into account the interrelations among interval duration, event duration, or interevent duration, it is a good example of mathematical modeling that can be applied to this topic.

It was recently reported that 45% of the articles published between 1995 and 2005 in the *Journal of Applied Behavior Analysis* used discontinuous methods of recording (Mudford, Taylor, & Martin, 2009). The widespread reliance on these methods to collect behavioral data and measure intervention outcomes highlights the importance of efforts to investigate the accuracy, reliability, and sensitivity of these measures. Studies that further describe the error patterns of observational sampling methods, such as those reported herein, are important for the continued refinement of our data-collection methods.

REFERENCES

Alvero, A. M., Rappaport, E., & Taylor, M. A. (2011). A further assessement of momentary time-sampling across extended interval lengths. *Journal of Organizational Behavior Management*, *31*, 117–129. doi: 10.1080/01608061.2011.569203

Alvero, A. M., Struss, K., & Rappaport, E. (2007). Measuring safety performance: A comparison of whole, partial, and momentary time-sampling recording methods. *Journal of Organizational Behavior Management*, *27*, 1–28. doi: 10.1300/J075v27n04_01

Ary, D. (1984). Mathematical explanation of error in duration recording using partial interval, whole interval, and momentary time sampling. *Behavioral Assessment*, *6*, 221–228.

Ary, D., & Suen, H. K. (1983). The use of momentary time sampling to assess both frequency and duration of behavior. *Journal of Behavioral Assessment*, *5*, 143–150. doi: 10.1007/BF01321446

Ary, D., & Suen, H. K. (1986). Interval lengths required for unbiased frequency and duration estimates with partial, whole, and momentary time sampling. *Midwestern Educational Researcher*, *8*, 17–24.

Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior change*, (3rd ed.) Boston, MA: Pearson Education.

Bartlett, S. M., Rapp, J. T., & Henrickson, M. L. (2011). Detecting false positives in multielement designs: Implications and brief assessments. *Behavior Modification*, *35*, 531–552. doi: 10.1177/0145445511415396

Brittle, A. R., & Repp, A. C. (1984). An investigation of the accuracy of momentary time sampling procedures with time series data. *British Journal of Psychology*, *75*, 481–488. doi: 10.1111/j.2044-8295.1984.tb01917.x

Carroll, R. A., Rapp, J. T., Colby-Dirksen, A. M., & Lindenberg, A. M. (2009). Detecting changes in simulated events II: Using variations of momentary time-sampling to measure changes in duration events. *Behavioral Interventions*, *24*, 137–155. doi: 10.1002/bin.286

Choi, C.-Y., Nam, H.-Y., & Lee, W.-S. (2007). Measuring behaviors of wintering black-faced spoonbills (*Platalea minor*): Comparison of behavioral sampling techniques. *Waterbirds*, *30*, 310–316. doi: 10.1675/1524-4695(2007)30[310:MTBOWB]2.0.CO;2

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Upper Saddle River, NJ: Pearson Education.

Devine, S. L., Rapp, J. T., Testa, J. R., Henrickson, M. L., & Schnerch, G. (2011). Detecting changes in simulated events using partial-interval recording and momentary time sampling III: Evaluating sensitivity as a function of session length. *Behavioral Interventions*, *26*, 103–124. doi: 10.1002/bin.328

Dunbar, R. I. M. (1976). Some aspects of research design and their implications in the observational study of behaviour. *Behaviour*, *58*, 78–98.

Engel, J. (1996). Choosing an appropriate sample interval for instantaneous sampling. *Behavioural Processes*, *38*, 11–17. doi: 10.1016/0376-6357(96)00005-8

Fiske, K., & Delmolino, L. (2012). Use of discontinuous methods of data collection in behavioral intervention: Guidelines for practitioners. *Behavior Analysis in Practice*, *5*, 77–81.

Gardenier, N. C., MacDonald, R., & Green, G. (2004). Comparison of direct observational methods for measuring stereotypic behavior in children with autism spectrum disorders. *Research in Developmental Disabilities*, *25*, 99–118. doi: 10.1016/j.ridd.2003.05.004

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.

Gunter, P. L., Venn, M. L., Patrick, J., Miller, K. A., & Kelly, L. (2003). Efficacy of using momentary time samples to determine on-task behavior of students with emotional/behavioral disorders. *Education and Treatment of Children*, *26*, 400–412.

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. doi: 10.1901/jaba.1986.19-73

Jacobsen, N. K., & Wiggins, A. D. (1982). Temporal and procedural influences on activity estimated by time-sampling. *Journal of Wildlife Management*, *46*, 313–324.

Kearns, K., Edwards, R., & Tingstrom, D. H. (1990). Accuracy of long momentary time-sampling intervals: Implications for classroom data collection. *Journal of Psychoeducational Assessment*, *8*, 74–85. doi: 10.1177/073428299000800109

Leger, D. W. (1977). An empirical evaluation of instantaneous and one-zero sampling of chimpanzee behavior. *Primates*, *18*, 387–393. doi: 10.1007/BF02383116

Mansell, J. (1985). Time sampling and measurement error: The effect of interval length and sampling pattern. *Journal of Behavior Therapy and Experimental Psychiatry*, *16*, 245–251. doi: 10.1016/0005-7916(85)90070-9

Mayer, G. R., Sulzer-Azaroff, B., & Wallace, M. (2012). *Behavior analysis for lasting change* (2nd ed.). Cornwall-on-Hudson, NY: Sloan.

McDowell, E. E. (1973). Comparison of time-sampling and continuous recording techniques for observing developmental changes in caretaker and infant behaviors. *Journal of Genetic Psychology*, *123*, 99–105. doi: 10.1080/00221325.1973.10533192

Meany-Daboul, M. G., Roscoe, E. M., Bourret, J. C., & Ahearn, W. H. (2007). A comparison of momentary time sampling and partial-interval recording for evaluating functional relations. *Journal of Applied Behavior Analysis*, *40*, 501–514. doi: 10.1901/jaba.2007.40-501

Milar, C. R., & Hawkins, R. P. (1976). Distorted results from the use of interval recording procedures. In T. A. Brigham, R. Hawkins, J. Scott, & T. F. McLaughlin (Eds.), *Behavior analysis in education: Self-control and reading* (pp. 261–273). Dubuque, IA: Kendall/Hunt.

Mudford, O. C., Beale, I. L., & Singh, N. N. (1990). The representativeness of observational samples of different durations. *Journal of Applied Behavior Analysis*, *23*, 323–331. doi: 10.1901/jaba.1990.23-323

Mudford, O. C., Taylor, S. A., & Martin, N. T. (2009). Continuous recording and interobserver agreement algorithms reported in the *Journal of Applied Behavior Analysis* (1995–2005). *Journal of Applied Behavior Analysis*, *42*, 165–169. doi: 10.1901/jaba.2009.42-165

Murphy, G., & Goodall, E. (1980). Measurement error in direct observations: A comparison of common recording methods. *Behaviour Research & Therapy*, *18*, 147–150. doi: 10.1016/0005-7967(80)90109-6

Murphy, M. J., & Harrop, A. (1994). Observer error in the use of momentary time sampling and partial interval recording. *British Journal of Psychology*, *85*, 169–179. doi: 10.1111/j.2044-8295.1994.tb02517.x

Powell, J. (1984). On the misrepresentation of behavioral realities by a widely practiced direct observation procedure: Partial interval (one-zero) sampling. *Behavioral Assessment*, *6*, 209–219.

Powell, J., Martindale, A., & Kulp, S. (1975). An evaluation of time-sample measures of behavior. *Journal of Applied Behavior Analysis*, *8*, 463–469. doi: 10.1901/jaba.1975.8-463

Powell, J., Martindale, B., Kulp, S., Martindale, A., & Bauman, R. (1977). Taking a closer look: Time sampling and measurement error. *Journal of Applied Behavior Analysis*, *10*, 325–332. doi: 10.1901/jaba.1977.10-325

Powell, J., & Rockinson, R. (1978). On the inability of interval time sampling to reflect frequency of occurrence data. *Journal of Applied Behavior Analysis*, *11*, 531–532. doi: 10.1901/jaba.1978.11-531

Rapp, J. T., Colby, A. M., Vollmer, T. R., Roane, H. S., Lomas, J., & Britton, L. N. (2007). Interval recording for duration events: A re-evaluation. *Behavioral Interventions*, *22*, 319–345. doi: 10.1002/bin.239

Rapp, J. T., Colby-Dirksen, A. M., Michalski, D. N., Carroll, R. A., & Lindenberg, A. M. (2008). Detecting changes in simulated events using partial-interval recording and momentary time sampling. *Behavioral Interventions*, *23*, 237–269. doi: 10.1002/bin.269

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. doi: 10.1901/jaba.1976.9-501

Rhine, R. J., & Ender, P. B. (1983). Comparability of methods used in the sampling of primate behavior. *American Journal of Primatology*, *5*, 1–15. doi: 10.1002/ajp.1350050102

Rojahn, J., & Kanoy, R. C. (1985). Toward an empirically based parameter selection for time-sampling observation systems. *Journal of Psychopathology and Behavioral Assessment*, *7*, 99–120. doi: 10.1007/BF00961077

Sanson-Fisher, R. W., Poole, A. D., & Dunn, J. (1980). An empirical method for determining an appropriate interval length for recording behavior. *Journal of Applied Behavior Analysis*, *13*, 493–500. doi: 10.1901/jaba.1980.13-493

Saudargas, R. A., & Zanolli, K. (1990). Momentary time sampling as an estimate of percentage time: A field validation. *Journal of Applied Behavior Analysis*, *23*, 533–537. doi: 10.1901/jaba.1990.23-533

Scheaffer, R. L., Mendenhall, W., Ott, R. L., & Gerow, K. G. (2012). *Elementary survey sampling* (7th ed.). Boston, MA: Brooks/Cole.

Simpson, M. J. A., & Simpson, A. E. (1977). One-zero and scan methods for sampling behaviour. *Animal Behaviour*, *25*, 726–731. doi: 10.1016/0003-3472(77)90122-1

Suen, H. K., & Ary, D. (1986). A post hoc correction procedure for systematic errors in time-sampling duration estimates. *Journal of Psychopathology and Behavioral Assessment*, *8*, 31–38. doi: 10.1007/BF00960870

Suen, H. K., Ary, D., & Covalt, W. (1991). Reappraisal of momentary time sampling and partial-interval recording. *Journal of Applied Behavior Analysis*, *24*, 803–804. doi: 10.1901/jaba.1991.24-803

Taylor, M. A., Skourides, A., & Alvero, A. M. (2012). Observer error when measuring safety-related behavior: Momentary time sampling versus whole-interval recording. *Journal of Organizational Behavior Management*, *32*, 307–319. doi: 10.1080/01608061.2012.729389

Test, D. W., & Heward, W. L. (1984). Accuracy of momentary time sampling: A comparison of fixed- and variable-interval observation schedules. In W. L. Heward, T. E. Heron, D. Hill, & J. Trap-Porter (Eds.), *Focus on behavior analysis in education* (1st ed., pp. 177–194). Columbus, OH: Merrill.

Tyler, S. (1979). Time-sampling: A matter of convention. *Animal Behaviour*, *27*, 801–810. doi: 10.1016/0003-3472(79)90016-2

Wilson, R. R., Jansen, B. D., & Krausman, P. R. (2008). Planning and assessment of activity budget studies employing instantaneous sampling. *Ethology*, *114*, 999–1005. doi: 10.1111/j.1439-0310.2008.01544.x

Received February 8, 2013
Final acceptance June 28, 2013
Action Editor, Jeffrey Tiger

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.