Four methods for analyzing partial interval recording data, with application to single-case research

James E. Pustejovsky and Daniel M. Swan

The University of Texas at Austin

This is the pre-print version of an article forthcoming in *Multivariate Behavioral Research*

Author Note

James E. Pustejovsky, Department of Educational Psychology, University of Texas at Austin. Daniel M. Swan, Department of Educational Psychology, University of Texas at Austin.

An earlier version of this paper was presented at the annual convention of the American Educational Research Association, April 4, 2014 in Philadelphia, PA.

Address correspondence to James E. Pustejovsky, Department of Educational Psychology, University of Texas at Austin, 1 University Station D5800, Austin, TX 78712. Email: pusto@austin.utexas.edu.

Abstract

Partial interval recording is a procedure for collecting measurements during direct observation of behavior. It is used in several areas of educational and psychological research, particularly in connection with single-case research. Measurements collected using partial interval recording suffer from construct invalidity because they are not readily interpretable in terms of the underlying characteristics of the behavior. Using an alternating renewal process model for the behavior under observation, we demonstrate that ignoring the construct invalidity of PIR data can produce misleading inferences, such as inferring that an intervention reduces the prevalence of an undesirable behavior when in fact it has the opposite effect. We then propose four different methods for analyzing PIR summary measurements, each of which can be used to draw inferences about interpretable behavioral parameters. We demonstrate the methods by applying them to data from two single-case studies of problem behavior.

Keywords: behavioral observation; partial interval recording; single-case research; alternating renewal process

Four methods for analyzing partial interval recording data, with application to single-case research

Direct observation of behavior is used in many areas of educational and psychological research, for a variety of purposes. For example, several existing instruments for diagnosis and assessment of behavioral disorders incorporate direct observation procedures (Volpe, DiPerna, Hintze, & Shapiro, 2005). In between-subjects randomized trials, behavioral observations are used both to define entrance criteria and as outcome measures (e.g., Durand, Hieneman, Clarke, Wang, & Rinaldi, 2012; Landa, Holman, O'Neill, & Stuart, 2011). Direct observation of behavior is considered a hallmark of single-case research, because intervention effects on behavioral outcomes often have immediate and recognizable social implications for the individual participants involved (Hartmann & Wood, 1990; Horner et al., 2005). The measurements produced by direct observation are considered low inference insofar as they correspond with easily understood, readily interpretable characteristics of a behavior. In comparison, other measurement procedures (such as rating scales) require an observer or respondent to make more abstract, global assessments of a subject's behavior and can thus be more difficult to ground in interpretable constructs.

Given the advantages and wide use of behavioral observation data, corresponding methods for statistical analysis are required. In between-subjects settings, the goal of such analysis might be to assess the correlates of a behavioral characteristic, such as the extent to which a parent's response on an assessment scale predicts some aspect of her child's behavior, or to test for differences between groups in the behavioral characteristic. In within-subjects settings, there is growing interest in methods for statistical analysis and meta-analysis of single-case research (e.g., Horner, Swaminathan, Sugai, & Smolkowski, 2012; Shadish, Rindskopf, & Hedges, 2008). Single-case designs involve measuring the behavior of an individual case repeatedly over time, including both before and after the controlled introduction of an intervention. Effects are identified by assessing changes in the pattern of behavior that correspond to the introduction or removal of the intervention.

With behavioral observation data from single-case designs, the goal of statistical analysis might be to test the hypothesis that the intervention had no effect on a target behavior, to quantify the magnitude of a behavior change for an individual case using an effect size metric, or to synthesize effect sizes across multiple cases.

Direct observation procedures require that the behavior of interest be precisely defined, so that a properly trained observer can judge whether it is present or absent at a given point in time. For a given, operationally defined behavior, an investigator may be interested in any of several distinct characteristics, including its prevalence, incidence, average duration, or average interim time. A behavior's prevalence is the proportion of time that it occurs. Incidence is the rate at which new instances of the behavior occur (per unit time). Mean duration is the average length of each unique episode of the behavior. Mean interim time is the average length of time in between episodes of the behavior.[1] In order to measure these characteristics, an observer will monitor the behavior of a subject for a length of time while recording data using one of several different procedures. Data from the observation session are then usually summarized into a single measurement.

Several different procedures are used to record data during direct observation, including continuous recording, frequency counting, and partial interval recording (Ayres & Gast, 2010). Continuous recording entails noting the times at which each behavioral episode begins and ends, thus capturing the full sequence of events during a session. Such data are typically summarized by calculating the proportion of session time during which the behavior was observed, which is a measure of prevalence. Frequency counting involves tallying the number of occurrences of the target behavior over the course of the session; standardizing the number of occurrences per unit of time produces a measure of incidence. Partial interval recording (PIR) involves dividing an observation session into short time

[1]The literature on direct observation of behavior often uses different, somewhat less concise terms for these quantities. Prevalence is sometimes referred to as "percent duration," incidence is sometimes called simply "rate", mean duration may be termed "duration per occurrence", and mean interim time is sometimes called "inter-response time" (Ayres & Gast, 2010).

intervals and scoring each interval according to whether or not the behavior occurs for any part of that interval. Typically, the raw scores from an observation session are then summarized by the proportion of intervals during which the behavior occurred. For instance, a 20 minute session may be divided into 80 intervals, each 15 seconds in length. Each interval is scored as a one if the behavior occurs at any point during the interval, and otherwise receives a score of zero. A summary score is calculated as the proportion of intervals receiving scores of one (equivalently, the mean interval score).

In contrast to continuous recording and frequency counting data, summarized PIR data measure neither the prevalence nor the incidence of a behavior (J. Altmann, 1974; Mann, Ten Have, Plunkett, & Meisels, 1991). Rather, PIR measurements depend on a combination of both prevalence and incidence, and the form of this relationship itself depends on the chosen length of each interval (Kraemer, 1979). Despite this ambiguity about the construct being measured, interval recording remains in wide use, particularly within single-case research (Mudford, Taylor, & Martin, 2009; Rapp et al., 2007). Some methods textbooks even recommend its use; for example, Kazdin (2011) advises: "Whenever there is doubt as to what assessment strategy should be adopted, an interval approach is almost always applicable" (p. 79).

Only a few available methods of analyzing PIR measurements account for how it confounds multiple constructs. S. A. Altmann and Wagner (1970) noted that if the behavior being observed follows a Poisson process, then applying a complementary log transform to the summary measurements yields a measure of incidence. However, the Poisson process model is only appropriate for behaviors where individual behavioral events have negligible duration (which implies that the prevalence of the behavior is essentially zero). Furthermore, the complementary log transformation is only suitable when the time between instances of behavior follows an exponential distribution, and the interpretation of the transformed measurements is quite sensitive to deviations from that parametric model (Fienberg, 1972).

Ary and Suen (1983) and Suen and Ary (1986) proposed methods for estimating both incidence and duration from PIR data by counting the frequency of certain patterns in the sequence of interval scores during an observation session. However, the authors added the qualification that the proposed methods apply only if the duration of individual behavioral events is always larger than the chosen interval length and the time in between instances of behavior is always longer than twice the chosen interval length.[2] Even if these conditions are met, the methods still require access to the sequence of raw scores from an observation session. If the analyst has access only to one summary measurement from each observation session–as would be the case with secondary data analysis or meta-analysis–then the required calculations cannot be carried out.

Lacking alternatives that are both valid and feasible, an analyst faced with PIR data may be inclined to simply ignore the construct validity problems with the measurements. In this paper, we demonstrate that doing so can produce misleading inferences. We then propose several methods for analyzing PIR data that are framed in terms of readily interpretable behavioral characteristics and that take into account the unusual characteristics of the measurements. We demonstrate the methods in examples drawn from single-case designs.

The proposed methods are all motivated by a particular model for the sequence of behaviors actually observed during a given session, or what we will call the behavior stream. The model, known as an alternating renewal process (ARP), treats the lengths of individual behavioral events and the interim times between behavioral events as mutually independent random quantities, each following some probability distribution. The ARP model provides a basis for expressing the properties of PIR summary measurements as functions of prevalence and incidence. Rogosa and Ghandour (1991) used the ARP model to study the psychometric properties of various behavioral observation procedures,

[2]Even under the stated conditions, it is unclear whether the methods are useful; Rogosa and Ghandour (1991) reported simulations in which the methods produce highly biased estimates of incidence and duration.

including PIR. Pustejovsky (2014b) used the model to define effect size metrics that are comparable across a variety of different procedures for collecting behavioral observation data.

The four analytic methods presented in the following sections rely on distinct set of further assumptions and yield different information about the behavior. The four methods are as follows:

1. Assume that the mean event duration is greater than some known value, which leads to upper and lower bounds on the prevalence of the behavior.

2. Assume that the mean event duration is less than some known value and that the probability of a new behavioral event occurring within a certain length of time from the end of the previous event is less than some known value. These assumptions lead to upper and lower bounds on the incidence of the behavior.

3. Assume that the mean event duration is equal across two samples of behavior and that the interim times in each sample follow exponential distributions. These assumptions lead to upper and lower bounds for the ratio of average interim times in the two samples.

4. Assume that both the event durations and interim times follow exponential distributions with different means. Moment estimators for prevalence and incidence can then be derived from analytic expressions for the mean and variance of PIR data.

All four methods rely on fairly strong assumptions about the behavior under observation. Their application will therefore require careful justification on the basis of prior knowledge. Furthermore, different sets of assumptions will be appropriate under very different empirical circumstances, and are not meant to be equal alternatives or competing approaches to analyzing the same data. Instead, this collection of different methods helps to delineate the various circumstances under which PIR measurements are actually informative about characteristics of the behavior stream.

The remainder of this paper is organized as follows. The next section describes the modeling assumptions (including the ARP model) that are common to all four methods,

then defines several effect size parameters that are potential targets of estimation. The following section uses the model to demonstrate how ignoring the construct invalidity of PIR data can produce misleading inferences. Each of the following four sections then discusses a method for analyzing PIR data, describing underlying assumptions and presenting a brief application to a single-case study.[3] A final section discusses limitations and extensions.

Common modeling assumptions

This section describes the assumptions of the alternating renewal process model, how PIR summary measurements are generated under that model, and effect size parameters of interest. Suppose that the analyst has two samples of PIR measurements, of size n_0 and n_1 respectively, and that interest is in comparing the samples. Let Y_{si} denote the i^{th} PIR measurement from sample s, for $s = 0, 1$.

All of the methods described below are premised on the assumption that each sample consists of statistically independent summary measurements, drawn from a common data generating process. For example, an analyst might have several measurements of a behavior on a single individual during a baseline phase and several further measurements after a treatment is introduced; here, interest is in comparing the characteristics of the individual's behavior during baseline to those during treatment. Given that repeated measurements are made on a single individual, the assumption that the data points within each phase are statistically independent and identically distributed is admittedly quite strong. In particular, the independence assumption would not apply if the prevalence or incidence of the behavior is increasing during the baseline phase (i.e., a baseline time trend), nor if the measurements are serially correlated. However, analysis of PIR data is challenging even in this simplistic case, as will be seen in later sections. We comment on alternative assumptions in the final section.

[3]Software implementing the four proposed methods is available in the ARPobservation package (Pustejovsky, 2014a) for the R statistical computing environment (R Core Team, 2014).

The equilibrium alternating renewal process model

The equilibrium alternating renewal process is a model for the stream of behavior during a single observation session, all of which goes into the generation of a single summary measurement Y_{si}. The model is operationalized in terms of the length of behavioral events and the interim times between events, both of which are treated as random quantities.[4] Specifically, the event durations during observation session i in sample s are assumed to be identically distributed with mean μ_s and cumulative distribution function $F_s(x; \mu_s)$. The interim times during the same observation session are assumed to be identically distributed with mean λ_s and cumulative distribution function $G_s(x; \lambda_s)$. All interim times and all event durations are assumed to be mutually independent, so that the length of the next event or interim time does not depend on the sequence of events leading up to it. Finally, the entire process is assumed to be aperiodic and in equilibrium, so that there is a constant probability that an event is occurring at any given point in time during the observation session.

Under the ARP model, the prevalence and incidence of the behavior are functionally related to the the mean event duration μ_s and the mean interim time λ_s. Prevalence, which we will denote as ϕ_s, is equal to the ratio of μ_s to the sum of μ_s and λ_s; incidence, which we will denote as ζ_s, is equal to the inverse of the sum of μ_s and λ_s. If the behavior was measured using continuous recording, then all of these parameters could be estimated directly (i.e., by taking the mean duration of the events and the mean of the interim times observed during the session). In contrast, the use of partial interval recording leads to a more complicated relationship between the recorded data and the parameters, making it comparatively difficult to learn about any of the characteristics of the behavior stream.

Partial interval recording

To use the PIR procedure, the observer divides a given observation session into K intervals of equal length. For each interval, the first c time units are used to observe the

[4] Kulkarni (2010, Chp. 8) provides an introduction to the mathematics of the ARP.

behavior, with the remainder of the interval used to record notes; we call c the *active interval length*. The observer counts a behavior as present if it occurs at any point during the active interval. Let $U_{sik} = 1$ if the behavior occurs at any point during the k^{th} interval, $U_{sik} = 0$ otherwise, for $k = 1, ..., K$, $i = 1, ..., n_s$, and $s = 0, 1$. The summary measurement Y_{si} is calculated as the proportion of intervals during which the behavior is observed at any point:

$$Y_{si} = \sum_{k=1}^{K} U_{sik}/K. \tag{1}$$

Under the assumptions of the ARP, it can be shown that the expected values of $U_{si1}, ..., U_{siK}$ and of Y_{si} are all equal, with

$$\mathrm{E}\left(Y_{si}\right) = \phi_s + \zeta_s \int_0^c \left[1 - G_s(x; \lambda_s)\right] dx, \tag{2}$$

where $\int_0^c$ denotes the definite integral taken over the interval $0 \leq x < c$. A proof of (2) is given in Pustejovsky (2014b).

Equation (2) makes apparent the crux of the construct validity problem with PIR data: the expectation of a PIR summary measurement depends on both the prevalence and incidence of the behavior, as well as on the exact distribution of interim times and the length of the active interval used for observation. If one interprets PIR data as measuring prevalence, then it is an upwardly biased measure. If one interprets PIR data as measuring the incidence of a behavior, as might be done when individual event durations are all very short and prevalence is near zero, then its bias is multiplicative and depends on the probability that an interim time is less than the active interval length. The analytic methods described in later sections employ different sets of further assumptions in order to learn about prevalence, incidence, or other behavior stream parameters based on samples of PIR data.

Target parameters

Under the assumptions of the ARP model, the prevalence, incidence, mean event duration, and mean interim time of the behavior are the primary parameters of interest.

The analyst's goal will often be to estimate or construct confidence intervals for effect sizes that compare these parameters across the two samples of measurements. In what follows, we focus on comparisons in the form of ratios or log-ratios. For instance, to compare the behavioral prevalence in the samples, we will consider the ratio of prevalence in the second sample to prevalence in the first sample, ϕ_1/ϕ_0, or its natural logarithm, $\ln \phi_1 - \ln \phi_0$. There are three reasons for focusing on ratio comparisons, as discussed in greater detail elsewhere (Pustejovsky, 2014b). First, proportionate changes are relatively simple to interpret. Second, proportionate changes are useful for comparing quantities that are both strictly positive, in a way that avoids range restrictions. For example, an 80% decrease in mean event duration is sensible regardless of the initial value of event duration, whereas a decrease of 5 s does not make sense if the initial event duration is only 3 s. Third, in certain circumstances, a proportionate change in one parameter can be meaningfully compared to a proportionate change in another parameter. For instance, if mean event duration is constant across samples ($\mu_0 = \mu_1$), then the proportionate change in incidence is equal to the proportionate change in prevalence: $\zeta_1/\zeta_0 = \phi_1/\phi_0$. Such equivalence can be useful when meta-analyzing results across studies that use different measurement procedures (Pustejovsky, 2014b). Finally, taking the natural logarithm of ratios is often a useful transformation for constructing confidence intervals because it can make the range of the comparison metric less restricted.

Ignoring construct invalidity

One method for analyzing PIR data is to simply ignore the construct invalidity issue and treat them just like any other measurements. In the context of a single-case design where the goal is to evaluate the effect of some intervention on the behavior of a case, it might be argued that the use of PIR is acceptable so long as the procedure is applied consistently across measurement occasions (i.e., holding the active interval length and session length constant for the duration of the study), so that the internal validity of the study is preserved. This line of argument is incorrect. In this section, we develop two

hypothetical examples demonstrating how using partial interval recording can produce misleading inferences about whether an intervention has the intended effect. The first example deals with a state behavior, in which individual instances can last for non-negligible lengths of time and the target of measurement is the behavior's prevalence. The second example deals with a discrete behavior, in which each instance of the behavior has negligible duration and primary interest is in the behavior's incidence.

Partial interval recording for measuring prevalence of a state behavior

Consider a study evaluating the effect of a particular teaching technique thought to prevent disruptive behavior. A particular child displays disruptive behavior that can last for non-trivial lengths of time, and so the main dimension of interest is prevalence. Prior to intervention, the child displays disruptive behaviors that last an average of $\mu_0 = 6$ seconds, with durations that follow a gamma distribution where $F_0(x) = \Gamma(x|2, 3)$.[5] The interim time between instances of disruptive behavior also follows a gamma distribution with $G_0(x) = \Gamma(x|3, 4)$, so that the average interim time is $\lambda_0 = 12$ seconds. It follows that the prevalence of disruptive behavior is $\mu_0/(\mu_0 + \lambda_0) = 0.33$. Further suppose that the teaching technique causes an increase in both the average duration of disruptive events and the average interim time. Specifically, when the intervention is applied, $F_1(x) = \Gamma(x|2, 10)$, $\mu_1 = 20$, $G_1(x) = \Gamma(x|3, 10)$, and $\lambda_1 = 30$. The prevalence of the behavior thus increases by 20%, from 0.33 to 0.40, meaning that the intervention does not produce the desired reduction in behavior–instead, it is actually harmful.

Suppose that the investigator uses an ABAB design, in which an initial control phase is followed by a treatment phase, a return to the control phase, and a final phase where the treatment is re-introduced. There are eight observation sessions per phase. During each session, she measures disruptive behavior using partial interval recording with an active interval length of $c = 15$ s, 5 s of rest time for recording, and a total session length of 20

[5]Here and following, we will write $\Gamma(x|k, \theta)$ to denote the cumulative distribution function of a gamma random variable with shape k, scale θ, and mean $k\theta$.

min. Figure 1 plots an example of how the results of this study might appear; we created it by simulating behavior stream data, applying the partial interval recording procedure, and calculating summary measurements for each session. In this simulated example, the average proportion of partial intervals is 0.89 during the A (baseline) phases and 0.68 during the B (intervention) phases; the proportion actually *decreases* slightly, even though the true prevalence of the behavior has increased. From Equation (2), one can verify that the decrease is not just a fluke of the particular sample, but rather will be observed generally. The mistaken inference arises because PIR data is upwardly biased as a measure of prevalence and the magnitude of bias depends on both the distribution of interim times and the average event duration. In this example, the increase in interim times leads to a change in the magnitude of the bias that masks the increase in prevalence.

Partial interval recording for measuring incidence of a discrete behavior

Similarly deceptive results are also possible when using partial interval recording to measure the incidence of discrete behaviors. Consider a study evaluating the effect of an intervention for reducing the self-injurious behavior of a child with autism; the child displays self-injurious behaviors that have very short duration, so that incidence is the primary dimension of interest. Suppose that, prior to intervention, the behaviors follow an alternating renewal process with all event durations equal to 0 (so $\mu_0 = 0$) and interim times that follow a gamma distribution with $G_0(x) = \Gamma(x|5, 9)$, so that the average interim time between self-injurious behaviors is $\lambda_0 = 45$ s. Further suppose that the intervention causes a change in the distribution of interim times, so that the behaviors are more frequent on average but also more sporadic, with a larger variance in interim times. Specifically, assume that after introducing the treatment, the behaviors follow an alternating renewal process with $\mu_1 = 0$ and interim time distribution $G_1(x) = \Gamma(x|0.5, 60)$, so that the average interim time between self-injurious behaviors is now $\lambda_1 = 30$ s. See Figure 2a for a plot of the interim time densities before and after intervention. The behaviors are substantially more frequent (going from 1.3 per minute to 2.0 per minute), meaning that the intervention

does not produce the desired reduction in behavior, and is instead actually harmful.

Suppose that the investigator again uses PIR to measure this behavior. As illustrated in Figure 2b, her conclusions will depend substantially on what active interval length she uses. If $c = 20$ s, she will observe a decrease of 10% on average, from 0.44 intervals in the absence of intervention to 0.40 intervals during the intervention, even though the true incidence has increased by 50%. For longer interval lengths, she would observe an even larger decrease. Only for active interval lengths of less than 13 s would the change in the expected proportion of intervals have the same sign as the true proportionate change in incidence.

These two examples illustrate that naïve analysis of PIR data can lead to mistaken inferences under various circumstances, including both when it is treated as a measure of prevalence and when it is treated as a measure of incidence. Of course, the mere possibility of these scenarios does not imply that construct invalidity will always be a plausible threat for studies that employ PIR. Unfortunately, we currently know little about how often such severe problems occur in practice nor about the extent to which the findings of extant research are distorted by the use of PIR. In this situation, analytic methods that account for the properties of PIR measurements–including that they are in fact sensitive to both the prevalence and incidence of the behavior–present one way to address potential construct validity threats. The following sections describe several different methods for analyzing PIR data that do just this, providing estimates that are directly interpretable in terms of behavioral characteristics.

Method 1: A bound for prevalence

The bias in PIR data as a measure of prevalence arises because entire intervals are counted even when the behavior occurs for only a fraction of the interval duration. As a result, the proportion of intervals is always rounded up relative to the actual proportion of session time that the behavior occurs. Intuitively, one might expect that the upward bias would be relatively minor if most instances of behavior are longer than the active interval

length.[6] Formalizing this intuition leads to a bound for the prevalence of a behavior in relation to the mean of PIR measurements.

Because PIR always involves rounding up, $\mathrm{E}(Y_s)$ is an upper bound for prevalence in sample s. Now suppose that a lower limit on the average event duration can be established based on an investigator's experience or knowledge about the behavior being observed, so that $0 < \mu_L^* \leq \mu_s$ for a known value μ_L^*. It follows that the prevalence of the behavior will be no less than a certain fraction of the expected value of the PIR measurements:

$$\frac{\mu_L^*}{\mu_L^* + c}\mathrm{E}(Y_s) \leq \phi_s \leq \mathrm{E}(Y_s). \tag{3}$$

A proof is given in Appendix A. Consistent with intuition, a larger value for μ_L^* will lead to a narrower bound for prevalence. However, the value of μ_L^* must be large relative to the active interval length in order for the bound to be narrow. For example, $\mu_L^* = 2c$ implies that ϕ_s will be between 67% and 100% of $\mathrm{E}(Y_s)$, a relatively wide range.

The bound given in (3) can be used to establish bounds on the log of the prevalence ratio in two samples. Assume that the lower bound on the mean event duration holds for both samples: so that $0 < \mu_L^* \leq \mu_0$ and $\mu_L^* \leq \mu_1$. Let $h^\phi = \ln(\mu_L^* + c) - \ln(\mu_L^*)$. It follows from (3) that bounds on the log of the prevalence ratio are given by

$$\ln\left[\frac{\mathrm{E}(Y_1)}{\mathrm{E}(Y_0)}\right] - h^\phi \leq \ln\left(\frac{\phi_1}{\phi_0}\right) \leq \ln\left[\frac{\mathrm{E}(Y_1)}{\mathrm{E}(Y_0)}\right] + h^\phi. \tag{4}$$

The bounds for the log of the prevalence ratio can be estimated by replacing $\mathrm{E}(Y_0)$ and $\mathrm{E}(Y_1)$ with the sample means. Let $\bar{y}_s$ be the mean and s_s^2 be the variance of the measurements in sample $s = 0, 1$. Some of the estimators presented in this and following sections cannot be calculated if the sample mean is at ceiling or floor levels. To account for this possibility, define the truncated sample means

$$\tilde{y}_s = \begin{cases} 1/(n_s K) & \text{if} \quad \bar{y}_s = 0 \\ \bar{y}_s & \text{if} \quad 0 < \bar{y}_s < 1 \\ 1 - 1/(n_s K) & \text{if} \quad \bar{y}_s = 1 \end{cases}$$

[6] Ary and Suen (1983) make a similar heuristic argument.

for $s = 0, 1$. Now, let

$$R = \ln \tilde{y}_1 - \ln \tilde{y}_0 \tag{5}$$

denote the log-response ratio of the truncated sample means. An estimate of the bounds in (4) is then given by $R \pm h^{\phi}$. Because h^{ϕ} is a known constant, the approximate variance of the estimated bounds is equivalent to the variance of R, which can be estimated as

$$V_R = \frac{s_0^2}{n_0 \tilde{y}_0^2} + \frac{s_1^2}{n_1 \tilde{y}_1^2}. \tag{6}$$

Furthermore, an approximate $(1 - \alpha)$ confidence interval (CI) for the log of the prevalence ratio is given by

$$R \pm \left(h^{\phi} + z_{\alpha/2} \sqrt{V_R} \right), \tag{7}$$

where $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of a standard normal distribution. Finally, an approximate $(1 - \alpha)$ CI for the prevalence ratio is given by exponentiating the upper and lower endpoints of (7).

Example 1

Moes (1998) used a single-case design to evaluate the effect of providing choice-making opportunities in the context of homework tutoring sessions on the disruptive behavior of four children with autism. The investigators measured disruptive behavior using PIR with $c = 10$ s active intervals, 5 s for recording, and $K = 80$ intervals per observation session. Each case was measured for a total of $n_0 = 10$ sessions in the no-choice condition and $n_1 = 10$ sessions in the choice condition; each condition was introduced across two phases, using a randomized ABAB/BABA design. Based on a visual analysis of graphed outcome data, Moes (1998) concluded that the children engaged in less disruptive behavior during choice conditions than during the no-choice conditions, and that the difference between conditions was larger for two cases (Charles and James).

Suppose that, based on experience with the types of disruptive behaviors exhibited by the study participants, the average length of disruptive behaviors can be established as greater than $\mu_L^* = 10$ s. Based on this assumption, Table 1 reports estimated bounds on

the log of the prevalence ratio for each of the four cases, along with approximate 95% CIs for each case.[7] The bounds estimates are largely consistent with the conclusions of a visual analysis, in that they are more extreme for Charles and James than for the other two cases.

An advantage of our statistical analysis is that it provides effect size estimates that characterize not just the presence but also the magnitude of changes in behavior, along with measures of the uncertainty in the estimates. These effect size estimates can be used to synthesize findings across the cases. The final row of the table reports fixed-effects meta-analyses of the end-points of the bounds.[8] The average log-prevalence ratio across the four cases is estimated to be between -2.58 and -1.20, equivalent to a 70-92% reduction in disruptive behavior. Based on the 95% confidence interval of [-3.27,-0.51] for the average-log prevalence ratio, the treatment leads to a reduction in disruptive behavior of 40-96%. While it is apparent that the treatment is beneficial, considerable uncertainty remains about the magnitude of the average effect.

Method 2: A bound for incidence

PIR is sometimes also used in context of discrete behaviors, where event durations are very short and incidence is the primary characteristic of interest. In this circumstance, if the interim times between events tend to be longer than the active interval length, then few intervals will contain multiple behavioral events and the number of intervals scored as a one will closely approximate the total number of events. Formalizing this argument leads to a bound for the incidence of a behavior in relation to the mean of PIR measurements.

Suppose that average event durations in sample s are shorter than some known value established based on prior experience, $\mu_U^* \geq \mu_s$. Also suppose that the interim time between behavioral events is rarely less than the active interval length, so that $G_s(c; \lambda_s) \leq p < 1$ for known value p. It follows that the incidence of the behavior is

[7] The estimates can be calculated directly from the summary statistics reported in Table 1.

[8] We used a fixed-effects meta-analysis both for simplicity of presentation and because generalization to a large population of cases (as implied by a random-effects meta-analysis) entails considerations that are beyond the scope of this article.

bounded by known factors of the expected value:

$$\frac{\mathrm{E}\left(Y_s\right)}{\mu_U^* + c} \leq \zeta_s \leq \frac{\mathrm{E}\left(Y_s\right)}{(1-p)c}. \tag{8}$$

A proof is given in Appendix B. Here, a smaller value for μ_U^* (corresponding to shorter event durations) will lead to a narrower bound for incidence. A smaller value of p, corresponding to a lower probability of interim times shorter than c, will also lead to a narrower bound for incidence. In the limit, if $\mu_U^* = 0$ and $p = 0$ then incidence is a constant fraction of the expected value of the PIR measurements.

The bound given in (8) can be used to establish bounds on the log of the incidence ratio in two samples. Assume that the upper limits on the mean event duration and probability of interim times less than the active interval length hold for both samples, so that $\mu_0 \leq \mu_U^*$, $\mu_1 \leq \mu_U^*$, $G_0(c; \lambda_0) \leq p$, and $G_1(c; \lambda_1) \leq p$. Letting $h^\zeta = \ln(\mu_U^* + c) - \ln(1 - p^*) - \ln(c)$, it follows that

$$\ln\left[\frac{\mathrm{E}\left(Y_1\right)}{\mathrm{E}\left(Y_0\right)}\right] - h^\zeta \leq \ln\left(\frac{\zeta_1}{\zeta_0}\right) \leq \ln\left[\frac{\mathrm{E}\left(Y_1\right)}{\mathrm{E}\left(Y_0\right)}\right] + h^\zeta. \tag{9}$$

These bounds can be estimated from partial interval recording data using $R \pm h^\zeta$, where R is the log-response ratio given in Equation (5). Because h^ζ is a fixed quantity, the variance of the bounds estimators can again be estimated as V_R from Equation (6). An approximate $(1-\alpha)$ CI for the log of the incidence ratio can be calculated as $R \pm \left(h^\zeta + z_{\alpha/2}\sqrt{V_R}\right)$; exponentiating the end-points of the CI provides an equivalent CI for the incidence ratio.

Example 2

Dunlap et al. (1994) used a single-case design to evaluate the effect of providing choice between academic activities on the disruptive behavior of three elementary school students with emotional and behavioral disorders. The investigators used PIR to measure disruptive behavior; for two cases (Sven and Ahmad), measurements were based on an active interval length of $c = 10$ s and 5 s for recording, while for the third case (Wendell), measurements were based on an active interval length of $c = 15$ s with no time for

recording. Observation sessions lasted 15 min, and each summary measurement was based on $K = 60$ intervals. Using visual analysis, Dunlap et al. (1994) concluded that providing choice-making opportunities reduced disruptive behavior for each of the three cases.

Based on descriptions of the types of disruptive behaviors exhibited by the three students, we assume that the average duration of behavioral events was fairly short. Primary interest is therefore in the incidence of disruptive behavior. To bound the log-incidence ratio, we assume that $\mu_0, \mu_1 \leq \mu_U^* = 10$ s. Since the active interval length varies across cases, we make separate assumptions about p for each case: for Sven and Ahmad we assume that the probability of interim times less than $c = 10$ s is at most $p = 0.15$, implying that $h^\zeta = 0.86$; for Wendell, we assume that the probability of interim times less than $c = 15$ s is at most $p = 0.25$, implying that $h^\zeta = 0.80$. The original investigators (or a meta-analyst with relevant clinical experience) would likely be able to further refine these assumptions based on contextual information.

Table 2 provides summary statistics by treatment condition for each case in the study, along with estimated incidence ratio bounds and approximate confidence intervals. The point-estimates of the bounds are below zero (corresponding to no effect of the treatment) for all three cases. However, the 95% CIs include zero for Sven and Wendell; thus, in contrast to the original investigators' visual assessment, accounting for both sampling uncertainty and identification-related uncertainty suggests that one cannot rule out the possibility that choice-making had no effect on the behavior of these two cases. The final row of Table 2 reports a fixed-effects meta-analysis of the bounds, which yields estimates of the bounds on the mean of the log-incidence ratio across the cases in the study. Given that the 95% CI for the average log-incidence ratio is [-3.69,-1.23], it is possible to conclude that, on average, the treatment reduced the incidence of disruptive behavior by more than 71% (i.e., $100\% \times [1 - \exp(-1.23)]$), and possibly as much as 97% (i.e., $100\% \times [1 - \exp(-3.69)]$). Based on conservative assumptions regarding the average behavioral event duration and probability of short interim times, it is reasonable to

conclude that this treatment is effective *on average* for these three cases, even though we cannot make as precise inferences about the presence of effects for individual cases.

Method 3: A bound for changes in mean interim time

The first two methods make assumptions only about the mean event duration and the probability of short interim times, but not about the full distribution of event durations or interim times. Entertaining a stronger set of distributional assumptions about the behavior stream will yield narrower bounds for parameters of interest. For instance, S. A. Altmann and Wagner (1970) proposed analyzing partial interval recording data under the assumptions that events have zero duration and that interim times follow an exponential distribution. Loosening the former assumption to allow for behaviors with longer event durations leads to a method that can be used to evaluate changes in interim time. The assumptions described below imply a bound on the ratio of interim times, which we define as λ_0/λ_1 so that its sign is consistent with the interpretation of ratios of the other parameters.

Assume that the mean event durations in each sample are equal, so that $\mu_0 = \mu_1$, but that this quantity is unknown. In the context of a single-case study, this assumption means that the average length of each behavioral event does not change between phases, although the distribution of interim times could still change. Further assume that the interim times in each sample follow exponential distributions, so that $G_s\left(x; \lambda_s\right) = 1 - \exp\left(-x/\lambda_s\right)$ for $s = 0, 1$. Denote the logistic transformation as $\text{logit}(x) = \ln(x) - \ln(1-x)$ and the complementary-log-log transformation as $\text{cll}(x) = \ln(-\ln(1-x))$. If follows that

$$f_L\left[\mathrm{E}\left(Y_0\right), \mathrm{E}\left(Y_1\right)\right] < \ln\left(\frac{\lambda_0}{\lambda_1}\right) < f_U\left[\mathrm{E}\left(Y_0\right), \mathrm{E}\left(Y_1\right)\right], \tag{10}$$

where the bounds are defined by the functions

$$f_L(x,y) = \begin{cases} \text{logit}(y) - \text{logit}(x) & \text{if} \quad x > y \\ \text{cll}(y) - \text{cll}(x) & \text{if} \quad x \leq y \end{cases}$$
$$f_U(x,y) = \begin{cases} \text{cll}(y) - \text{cll}(x) & \text{if} \quad x > y \\ \text{logit}(y) - \text{logit}(x) & \text{if} \quad x \leq y. \end{cases} \tag{11}$$

A proof is given in Appendix C. The bound involving the complementary-log-log corresponds to $\mu_0 = \mu_1 = 0$ and is equivalent to the estimator proposed by S. A. Altmann and Wagner (1970). As the mean event duration increases, the implied value of the log of the interim ratio approaches the log-odds ratio of the expected values in each sample. The bounds will be narrow if $\mathrm{E}(Y_0)$ and $\mathrm{E}(Y_1)$ are both near zero or are close to equal.

Point-estimates for the bounds in (10) can be formed by substituting the sample means for expectations, taking $(\hat{f}_L, \hat{f}_U) = (f_L(\tilde{y}_0, \tilde{y}_1), f_U(\tilde{y}_0, \tilde{y}_1))$. The variance of $(\hat{f}_L, \hat{f}_U)$ is complicated by the fact that f_L and f_U are not smooth functions. Define the large-sample variance of the log-odds ratio as

$$V_{LOR} = \frac{s_0^2}{n_0 \tilde{y}_0^2 (1-\tilde{y}_0)^2} + \frac{s_1^2}{n_1 \tilde{y}_1^2 (1-\tilde{y}_1)^2} \tag{12}$$

and the large-sample variance of the complementary-log-log ratio as

$$V_{CLR} = \frac{s_0^2}{n_0 (1-\tilde{y}_0)^2 [\ln(1-\tilde{y}_0)]^2} + \frac{s_1^2}{n_1 (1-\tilde{y}_1)^2 [\ln(1-\tilde{y}_1)]^2}. \tag{13}$$

The variance of $\left(\hat{f}_L, \hat{f}_U\right)$ can then be estimated as

$$V_{fL} = \begin{cases} V_{LOR} & \text{if} \quad \text{cll}(\tilde{y}_1) - \text{cll}(\tilde{y}_0) \leq z_{\alpha/2}\sqrt{V_{LOR}} \\ V_{CLR} & \text{if} \quad \text{cll}(\tilde{y}_1) - \text{cll}(\tilde{y}_0) > z_{\alpha/2}\sqrt{V_{LOR}} \end{cases}$$
$$V_{fU} = \begin{cases} V_{CLR} & \text{if} \quad \text{cll}(\tilde{y}_1) - \text{cll}(\tilde{y}_0) < -z_{\alpha/2}\sqrt{V_{LOR}} \\ V_{LOR} & \text{if} \quad \text{cll}(\tilde{y}_1) - \text{cll}(\tilde{y}_0) \geq -z_{\alpha/2}\sqrt{V_{LOR}} \end{cases} \tag{14}$$

A CI that covers the interval (f_L, f_U) (and thus the true log of the interim ratio) with probability of approximately $(1-\alpha)$ is given by $\left[\hat{f}_L - z_{\alpha/2}\sqrt{V_{fL}}, \hat{f}_U + z_{\alpha/2}\sqrt{V_{fU}}\right]$. Exponentiating the end-points of the CI provides an equivalent CI for the ratio of interim times.

Example 1, continued

Returning to the data from the study by Moes (1998), suppose that the investigators are confident that the choice-making intervention did not alter the average length of the participants' disruptive behaviors. If it is further assumed that the interim times between episodes of disruptive behavior are exponentially distributed, then the change in average interim times can be quantified by using Method 3. Table 3 reports estimated bounds on the log of the interim time ratio for each of the four cases, along with approximate 95% CIs. The final row of the table reports fixed-effects meta-analyses based on the end-points of the bounds. The average log-interim ratio across the four cases is estimated to be between -2.26 and -2.08, equivalent to a 87-90% reduction in disruptive behavior. Based on the 95% confidence interval of [-3.01,-1.36] for the average log-interim ratio, which accounts for sampling uncertainty in the bounds estimates, the treatment leads to a reduction of 74-95%. The narrow confidence interval suggests that one may be confident that the treatment is very efficacious, on average, for these four cases. However, this is due in large part to the strength of the parametric assumptions upon which Method 3 is premised.

Method 4: Moment estimators for prevalence and incidence

Method 3 introduced the assumption that interim times are exponentially distributed, but made no assumption about the parametric form of event durations. A final method for analyzing PIR data introduces a further parametric assumption for the event time distribution, in order to obtain point estimates (rather than interval bounds) for the parameters of the behavior stream. Specifically, it is assumed that the behavior stream follows an Alternating Poisson Process, which is a special case of the ARP model where both event durations and interim times follow exponential distributions. Under these

assumptions, expressions for both the mean and variance of PIR data can be obtained in terms of the underlying parameters of the behavior stream. These expressions can be used to form moment estimators for both prevalence and incidence (or equivalently, for the mean event duration and the mean interim time).

Assume that the event durations and the interim times both follow exponential distributions, so that $F_s(x;\mu_s) = 1 - \exp(-x/\mu_s)$ and $G_s(x;\lambda_s) = 1 - \exp(-x/\lambda_s)$. It follows from (2) that

$$\mathrm{E}(Y_s) = 1 - (1-\phi_s)\exp\left(\frac{-\zeta_s c}{(1-\phi_s)}\right). \tag{15}$$

It follows further that the variance of the PIR measurement is given by

$$\mathrm{Var}(Y_s) = \frac{\mathrm{E}(Y_s)\left[1-\mathrm{E}(Y_s)\right]}{K}\left[1 + \frac{2\phi_s}{K\mathrm{E}(Y_s)}\sum_{k=1}^{K-1}(K-k)\exp\left(\frac{\zeta_s c}{\phi_s} - \frac{\zeta_s k L}{\phi_s(1-\phi_s)K}\right)\right]. \tag{16}$$

Appendix D provides a derivation of (16).

Moment estimators for prevalence and incidence are obtained for each $s = 0, 1$ by replacing $\mathrm{E}(Y_s)$ with the sample mean in (15) and (16), replacing $\mathrm{Var}(Y_s)$ with the sample variance in (16), and solving both expressions for ϕ_s and ζ_s. Solutions exist when the sample moments satisfy $0 < \bar{y}_s < 1$ and $\bar{y}_s(1-\bar{y}_s)/K \leq s_s^2 < \bar{y}_s(1-\bar{y}_s)$. However, it is possible to obtain values of the sample variance in the range $0 \leq s_s^2 \leq \bar{y}_s(1-\bar{y}_s)n_s/(n_s-1)$. To ensure that the estimators are well-defined over the entire space of sample moments, we use the truncated estimator

$$\tilde{s}_s^2 = \min\left[\max\left(s_s^2, \frac{\tilde{y}(1-\tilde{y})}{K} + \frac{1}{n_s K^2}\right), \tilde{y}(1-\tilde{y}) - \frac{1}{n_s K^2}\right]$$

in place of s_s^2.

The solution to the moment equations can be simplified by profiling ζ_s. Let $\hat{\phi}_s$ and $\hat{\zeta}_s$ denote the moment estimators. For a given value of $\tilde{y}_s$ and $\hat{\phi}_s$,

$$\hat{\zeta}_s = \frac{-(1-\hat{\phi}_s)}{c}\ln\left(\frac{1-\tilde{y}_s}{1-\hat{\phi}_s}\right). \tag{17}$$

Substituting (17) into (16), $\hat{\phi}_s$ is then the solution to

$$\tilde{s}_s^2 = \frac{\tilde{y}_s(1-\tilde{y}_s)}{K}\left[1 + \frac{2\phi_s}{K\tilde{y}_s}\sum_{k=1}^{K-1}(K-k)\exp\left[\left(\frac{kL}{\phi_s cK} - \frac{1-\phi_s}{\phi_s}\right)\ln\left(\frac{1-\tilde{y}_s}{1-\phi_s}\right)\right]\right]. \tag{18}$$

Moment estimators of the mean event duration and mean interim time are given by $\hat{\mu}_s = \hat{\phi}_s/\hat{\zeta}_s$ and $\hat{\lambda}_s = \left(1 - \hat{\phi}_s\right)/\hat{\zeta}_s$, respectively.

It is quite difficult to obtain analytic expressions for the sampling variance of the moment estimators. Instead, we consider a parametric bootstrap approach (Efron & Tibshirani, 1998) for obtaining standard errors and constructing confidence intervals for parameters of interest. Appendix E provides details regarding computation of the parametric bootstrap.

Simulation results

We conducted a simulation study to assess the performance of Method 4 over a wide range of behavioral parameters (ϕ_s, ζ_s), intervals per observation session (K), and sample sizes (n_s).[9] The simulations examined the bias of the moment estimators in both absolute and relative terms, the efficiency of the moment estimator for prevalence relative to the sample mean of the PIR measurements, and the coverage rates of bootstrap CIs for prevalence and incidence.

The moment estimators have reasonably small biases in absolute terms, except when prevalence or incidence is large. Furthermore, at all sample sizes and over most of the parameter space, the moment estimator of prevalence has smaller root mean-squared error than the "naïve" estimator given by the sample mean of the PIR measurements. For instance, when $n \geq 12$, the moment estimator is nearly always more efficient than the naïve estimator when prevalence is less than one half; over a large part of this space, it is at least 80% more efficient.

However, the moment estimators perform less well if assessed in terms of relative (proportionate) bias. Even using a rather liberal criteria for relative bias and restricting attention to only part of the parameter space, rather large sample sizes are required in order to obtain close to unbiased estimates of prevalence or incidence. Specifically, the

[9]The online supplementary materials include a much more detailed description of the simulation design and results.

simulation results suggest that sample sizes of $n_s \geq 20$ and $K \geq 40$ are required. When based on samples of this size, the moment estimator for prevalence has bias of less than 5% when prevalence is moderate ($.25 \leq \phi_s \leq .75$) and incidence is relatively low when measured in terms of the active interval length ($\zeta_s \leq 0.10/c$); the moment estimator for incidence has bias of less than 5% when prevalence is low ($\phi_s \leq .10$) and incidence is also low ($\zeta_s \leq 0.10/c$). Similarly large sample sizes are required to obtain CIs with acceptable coverage rates. When $n \geq 20$ and $K \geq 40$, nominally 95% parametric bootstrap CIs have coverage of between 92.5% and 97.5% over a portion of the parameter space that mirrors the conditions where the moment estimators have small relative bias.

On the whole, the simulation results indicate that the moment estimators and associated CIs may be used for tentative, exploratory purposes, particularly when the alternative would be to rely on the naïve estimator of prevalence. However, the moment estimators may produce somewhat biased estimates of log-ratio effect size estimates when based on sample sizes that are typical for single-case designs. Furthermore, the conditions under which the prevalence estimator performs well overlap very little with the conditions where the incidence estimator performs well. Consequently, use of this method should be restricted to contexts where only one of the parameters is of primary interest, just as with the methods discussed in previous sections.

Example 1, continued

Returning yet again to the data from the study by Moes (1998), we now assume that the behavior streams followed an Alternating Poisson Process, in which both event durations and interim times are exponentially distributed. We use Method 4 to estimate change in the prevalence of disruptive behavior, as quantified by the log of the prevalence ratio. Table 4 reports the results.

The estimated log-prevalence ratios for Charles, Chuck, and James correspond to reductions in prevalence of 97%, 66%, and 94%, respectively. For Carl's data in the no-choice condition, the sample variances are less than $\tilde{y}_0(1 - \tilde{y}_0)/K$; the estimates are

therefore based on the truncated value of $\tilde{s}_0^2$. This leads to a large, positive estimate of 0.662 for the log-prevalence ratio, which may be implausible. Given the small sample sizes in each condition ($n_0 = n_1 = 10$), our simulation results suggest that the estimates may be biased in all four cases. In particular, the prevalence estimates in the choice condition are well outside of the range where we might expect to obtain accurate results. This example illustrates a crucial drawback of using Method 4 to estimate changes across conditions: even if prevalence (or incidence) in one condition is within the range where approximately unbiased estimates can be obtained, it may not remain so in the other condition.

Discussion

Behavioral data collected using partial interval recording suffer from construct invalidity because such measurements are not readily interpretable in terms of the underlying characteristics of the behavior; in short, PIR measures neither prevalence nor incidence (J. Altmann, 1974; Mann et al., 1991). Using an alternating renewal process model for the behavior stream, we have demonstrated that ignoring the construct invalidity of PIR data can produce misleading inferences, such as inferring that an intervention reduces the prevalence of an undesirable behavior when in fact it has the opposite effect. We then proposed four different methods for analyzing PIR summary measurements, all of which are based on the ARP model for the behavior stream, and all of which produce estimates of interpretable behavioral parameters, such as the prevalence ratio or incidence ratio.

All four methods rely on strong sets of assumptions about the behavior being measured. The first method involves assuming that mean duration is greater than a known value, which implies bounds for the behavior's prevalence. For the resulting bound to be narrow and informative, the assumed value for minimum duration must be large relative to the active interval length. The second method involves assuming that the mean duration is less than a known value and that interim times less than the active interval length occur only rarely; together, these assumptions imply bounds for the behavior's incidence. The

shorter the mean event duration, and the smaller the maximum interim time probability, the narrower will be the resulting bound.

Though Methods 1 and 2 both involve assumptions about certain aspects of the behavior stream, neither invokes assumptions regarding the parametric form of the event duration or interim time distributions. In contrast, Methods 3 and 4 both involve assuming that interim times follow an exponential distribution. Method 3 adds the further assumption that the mean event duration is constant across samples, which implies a relatively narrow bound for the ratio of two mean interim times. Method 4 instead adds the assumption that event durations also follow an exponential distribution, which leads to moment estimators for prevalence and incidence. While Methods 3 and 4 will tend to produce more informative estimates than Methods 1 and 2, this advantage is due to the use of more restrictive assumptions that may be difficult to verify without detailed data on individual behavior streams. Furthermore, Method 4 requires relatively large sample sizes to generate estimates with small bias and CIs with acceptable coverage. Consequently, we recommend that Methods 1 and 2 should be privileged except when prior evidence can be used to establish that the stronger modeling assumptions of Method 3 or 4 are reasonable.

Given the strength of the required assumptions, it should be emphasized that valid application of any of the proposed methods will require credible prior information about the behavior being observed. We have attempted to state the assumptions, particularly those entailed in Methods 1 and 2, in terms of quantities that a knowledgeable observer could interpret, assess, and perhaps roughly estimate. Apart from contextual knowledge, all of the assumptions are susceptible to empirical investigation. Systematic studies of the characteristics of different behavior profiles (such as the disruptive behaviors of students in classroom settings, self-injurious behavior, etc.) would be extremely useful for determining the settings and circumstances in which any of the proposed methods for analyzing PIR data could reasonably be applied. Of course, in order for such studies to provide informative data, observation procedures other than PIR would need to be employed, and

careful attention would need to be given to variation in the behavioral characteristics across individuals and across settings.

We see several avenues for further development of the four proposed analytic methods. We have studied the methods from a frequentist perspective, but taking a Bayesian approach could be fruitful. For Methods 1 and 2, which require the analyst to specify thresholds (minimum or maximum values) of certain parameters, a Bayesian approach would allow the analyst to incorporate prior uncertainty regarding the thresholds, rather than treating them as known. In Method 4, incorporation of prior information regarding prevalence or incidence may improve the operating characteristics of the moment estimators in small samples. However, the textbook Bayesian approach requires an expression for the full likelihood of the PIR summary measurements, which is quite difficult to calculate. Approximate Bayesian methods (e.g. Turner & Van Zandt, 2012) might present a solution here and should be investigated further.

All of the methods that we have described are based on the assumption that the PIR measurements from a given sample are independently and identically distributed. In between-subjects contexts, the assumption that the measurements are identically distributed may not be reasonable if there is person-specific heterogeneity in the behavioral parameters (i.e., variation in prevalence or incidence across individuals). In repeated-measures contexts such as arise in single case designs, the independence assumption is often considered unreasonable. Instead, recent discussions of statistical methods for single-case research have emphasized the need to allow for simple forms of auto-correlation (Horner et al., 2012). We would expect that Methods 1, 2, and 3 are somewhat robust to auto-correlation in the measurements because they depend only on the expected value of the measurements. So long as the auto-correlation does not enter in such a way as to invalidate Expression (2), the bounds that we have derived will remain valid. Furthermore, even when the data are auto-correlated, the proposed methods will provide reasonable estimates of these bounds, which could be synthesized across cases in a

single-case design or used in a larger meta-analysis of multiple studies. However, the variance estimation methods we have proposed will tend to under-state uncertainty if the data are auto-correlated; methods that better account for auto-correlation remain a topic for further investigation. We also suspect that Method 4 will be relatively sensitive to auto-correlation because it assumes a relationship between the variance and the behavioral parameters (Expression 16) that will be inaccurate in the presence of serial dependence.

It would be useful to extend the proposed methods to handle regressions of PIR measurements on predictor variables, such as time trends. However, such extensions appear to be quite difficult. While Methods 1 and 2 could in principle be extended to handle regression models, interval-bound regressions require specialized, complex estimation techniques (Manski & Tamer, 2002). In the more highly parameterized Methods 3 and 4, the behavioral parameters cannot be easily separated to formulate a generalized linear model. If the analyst has access only to session-level PIR summary measurements, then the very simple model that we have considered may stand near the limit of what is feasible.[10]

Despite the strength of the assumptions on which they are based, the methods that we have described in this paper represent an improvement over extant methods of analysis that simply ignore the construct invalidity of PIR data. The relative simplicity of the required calculations for the first three methods also means that they can readily be used for re-analysis of extant studies. Even if the authors of a study do not report the results of these proposed analyses, readers could still examine the extent to which the findings could be distorted due to the use of PIR procedures. Furthermore, the availability of the proposed methods may permit research synthesists to include studies that use PIR procedures in meta-analyses, as well as to draw comparisons with studies that use other, more readily interpretable measurement procedures (cf. Pustejovsky, 2014b).

[10]If the analyst has access to the individual interval-level data from a session, or what we have denoted above as $U_{si1}, ..., U_{siK}$, then more flexible modeling strategies may be possible. We are currently exploring models for such interval-level data (Pustejovsky, 2013).

Given the ready availability of electronic devices and programs for recording behavioral observations, the use of PIR for research purposes requires strong justification. An advantage of the analytic methods that we have described is that they can be used to formulate and assess such justifications, because the assumptions that motivate the analytic methods delineate circumstances under which informative parameter estimates can be obtained. Investigators can therefore use the methods as a guide to make more principled prospective decisions about the use of PIR. For example, if an investigator has a rough estimate of the mean duration of a certain behavior before measuring it using PIR, she could select an active interval length so that the bound from Method 1 is narrow. Similarly, an investigator who wishes to obtain accurate point estimates of prevalence or incidence could choose a total sample size large enough for Method 4 to perform well. We would suggest that, if the ARP is a reasonable model for the behavior under study, then the use of PIR for direct observation is only warranted if the investigator can justify the assumptions of one of the four analytic methods that we have proposed. An investigator who uses PIR under other circumstances risks being misled.

References

Altmann, J. (1974). Observational study of behavior: Sampling methods. *Behaviour*, *49*(3/4), 227–267.

Altmann, S. A., & Wagner, S. S. (1970). Estimating rates of behavior from Hansen frequencies. *Primates*, *11*(2), 181–183. doi: 10.1007/BF01731143

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*(2), 143–150.

Ayres, K., & Gast, D. L. (2010). Dependent measures and measurement procedures. In D. L. Gast (Ed.), *Single subject research methodology in behavioral sciences* (pp. 129–165). New York, NY: Routledge.

Dunlap, G., DePerczel, M., Clarke, S., Wilson, D., Wright, S., White, R., & Gomez, A. (1994). Choice making to promote adaptive behavior for students with emotional and behavioral challenges. *Journal of Applied Behavior Analysis*, *27*(3), 505–518.

Durand, V. M., Hieneman, M., Clarke, S., Wang, M., & Rinaldi, M. L. (2012). Positive Family Intervention for Severe Challenging Behavior I: A Multisite Randomized Clinical Trial. *Journal of Positive Behavior Interventions*, *15*(3), 133–143. doi: 10.1177/1098300712458324

Efron, B., & Tibshirani, R. J. (1998). *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.

Fienberg, S. E. (1972). On the use of Hansen frequencies for estimating rates of behavior. *Primates*, *13*(3), 323–325.

Hartmann, D. P., & Wood, D. D. (1990). Observational methods. In A. S. Bellack, M. Hersen, & A. E. Kazdin (Eds.), *International handbook of behavior modification and therapy* (2nd ed., pp. 107–138). New York, NY: Plenum Press.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S. L., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education.

Exceptional Children, *71*(2), 165–179.

Horner, R. H., Swaminathan, H., Sugai, G., & Smolkowski, K. (2012). Considerations for the systematic analysis and use of single-case research. *Education and Treatment of Children*, *35*(2), 269–290. doi: 10.1353/etc.2012.0011

Kazdin, A. E. (2011). *Single-Case Research Designs: Methods for Clinical and Applied Settings*. New York, NY: Oxford University Press.

Kraemer, H. C. (1979). One-zero sampling in the study of primate behavior. *Primates*, *20*(2), 237–244.

Kulkarni, V. G. (2010). *Modeling and Analysis of Stochastic Systems*. Boca Raton, FL: Chapman & Hall/CRC.

Landa, R. J., Holman, K. C., O'Neill, A. H., & Stuart, E. A. (2011). Intervention targeting development of socially synchronous engagement in toddlers with autism spectrum disorder: A randomized controlled trial. *Journal of child psychology and psychiatry, and allied disciplines*, *52*(1), 13–21. doi: 10.1111/j.1469-7610.2010.02288.x

Mann, J., Ten Have, T. R., Plunkett, J. W., & Meisels, S. J. (1991). Time sampling: A methodological critique. *Child Development*, *62*(2), 227–241.

Manski, C. F., & Tamer, E. (2002). Inference on regressions with interval data on a regressor or outcome. *Econometrica*, *70*(2), 519–546.

Moes, D. R. (1998). Integrating choice-making opportunities within teacher-assigned academic tasks to facilitate the performance of children with autism. *Research and Practice for Persons with Severe Disabilities*, *23*(4), 319–328.

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*(1), 165–169. doi: 10.1901/jaba.2009.42-165

Pustejovsky, J. E. (2013). *Observation procedures and Markov chain models for estimating the prevalence and incidence of a behavior.*

Pustejovsky, J. E. (2014a). *ARPobservation: Simulating recording procedures for direct observation of behavior.* Retrieved from `http://cran.r-project.org/web/packages/ARPobservation`

Pustejovsky, J. E. (2014b). Measurement-comparable effect sizes for single-case studies of free operant behavior. *Psychological Methods*, (In press). doi: 10.1037/met0000019

R Core Team. (2014). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from `http://www.r-project.org/`

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

Rogosa, D., & Ghandour, G. (1991). Statistical models for behavioral observations. *Journal of Educational Statistics*, *16*(3), 157–252.

Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention*, *2*(3), 188–196. doi: 10.1080/17489530802581603

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*(1), 31–38. doi: 10.1007/BF00960870

Turner, B. M., & Van Zandt, T. (2012). A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology*, *56*(2), 69–85. doi: 10.1016/j.jmp.2012.02.005

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

Table 1

Estimated bounds for the log of the prevalence ratio for cases from Moes (1998), based on Method 1

Case	$\bar{y}_0$	s_0	$\bar{y}_1$	s_1	Estimate	95% CI
Carl	0.14	0.03	0.04	0.11	(-1.92,-0.53)	[-3.69,1.24]
Charles	0.35	0.20	0.01	0.03	(-4.12,-2.74)	[-5.56,-1.30]
Chuck	0.27	0.15	0.09	0.13	(-1.81,-0.42)	[-2.80,0.57]
James	0.50	0.17	0.03	0.09	(-3.50,-2.12)	[-5.39,-0.24]
FE meta-analysis					(-2.58,-1.20)	[-3.27,-0.51]

Table 2

Estimated bounds for the log of the incidence ratio for cases from Dunlap et al. (1994), based on Method 2

Case	n_0	$\bar{y}_0$	s_0	n_1	$\bar{y}_1$	s_1	Estimate	95% CI
Ahmad	8	0.70	0.18	8	0.02	0.02	(-4.19,-2.48)	[-4.72,-1.96]
Sven	15	0.36	0.19	6	0.09	0.10	(-2.20,-0.49)	[-3.07,0.39]
Wendall	10	0.26	0.12	11	0.06	0.07	(-2.24,-0.64)	[-3.01,0.12]
FE meta-analysis							(-3.30,-1.62)	[-3.69,-1.23]

Table 3

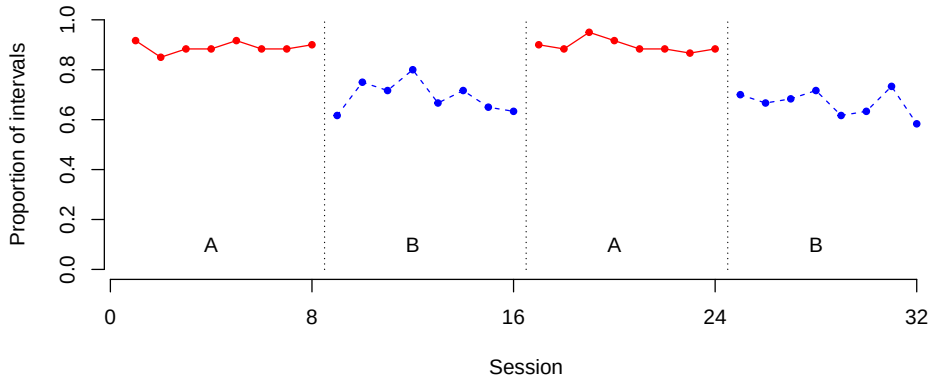
Estimated bounds for the log of the interim ratio for cases from Moes (1998), based on Method 3

Case	Estimate	95% CI
Carl	(-1.33,-1.28)	[-3.18,0.57]
Charles	(-3.85,-3.63)	[-5.36,-2.16]
Chuck	(-1.34,-1.23)	[-2.47,-0.17]
James	(-3.47,-3.12)	[-5.44,-1.20]
FE meta-analysis	(-2.26,-2.08)	[-3.01,-1.36]

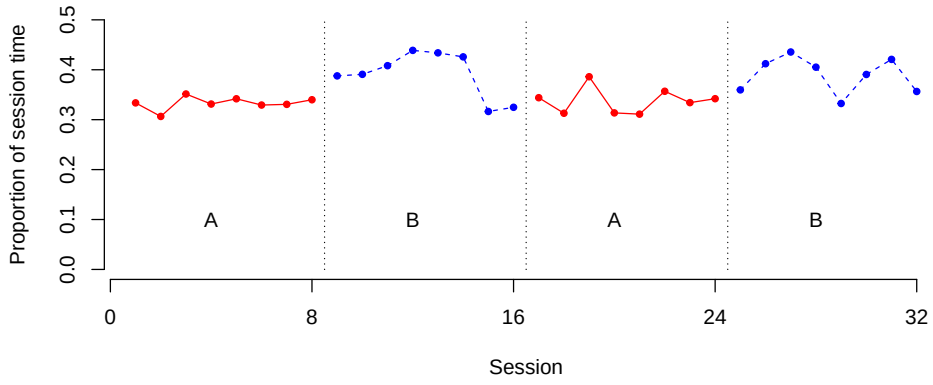
Table 4

Estimates for cases from Moes (1998), based on Method 4

	Prevalence				Incidence (per interval)			
Case	No Choice	Choice	log-ratio	95% CI	No Choice	Choice	log-ratio	95% CI
Carl	0.020	0.039	0.662	[-4.45,1.77]	0.124	0.001	-4.453	[-5.48,-2.77]
Charles	0.329	0.009	-3.630	[-6.28,-2.53]	0.019	0.003	-2.021	[-3.92,-0.57]
Chuck	0.247	0.084	-1.085	[-2.64,-0.22]	0.024	0.005	-1.540	[-3.01,-0.22]
James	0.465	0.029	-2.787	[-6.62,-1.51]	0.035	0.001	-3.241	[-4.88,-1.64]

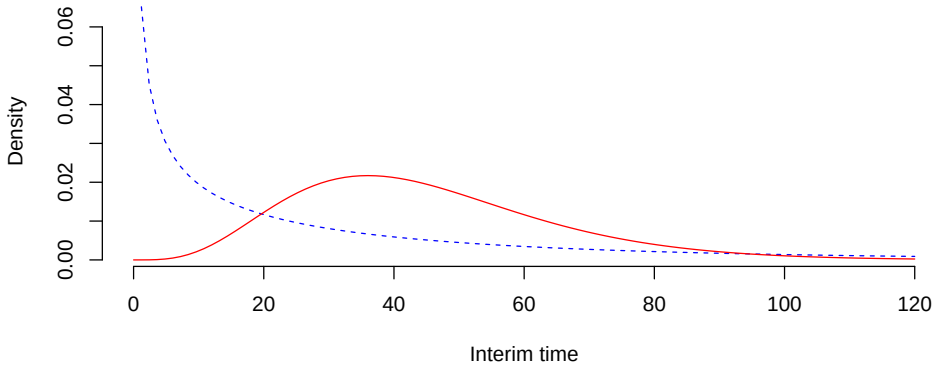


(a) Simulated single-case graph using 15 s partial interval recording.

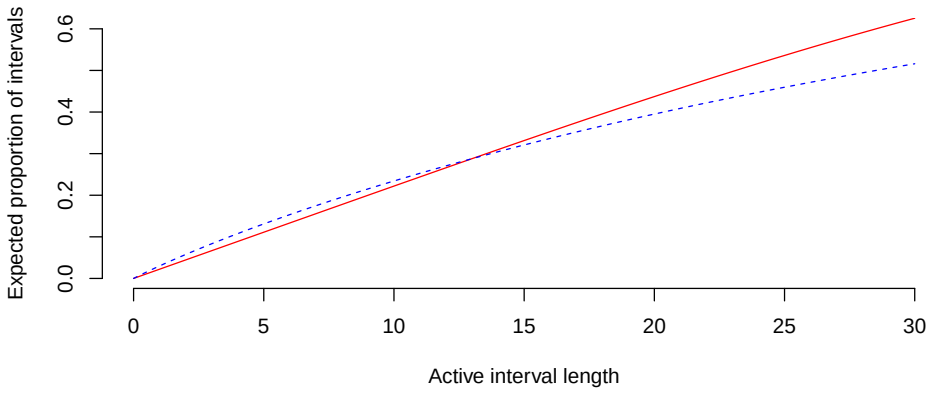


(b) Simulated single-case graph using continuous recording.

Figure 1. Example of partial interval recording with a state behavior.



(a) Density of interim time distributions.



(b) Expected proportion of intervals as a function of active interval length.

Figure 2. Example of partial interval recording with a discrete behavior. Solid lines correspond to the distribution prior to intervention. Dashed lines correspond to the distribution after intervention.

Appendix A

A bound for prevalence

Denote the expected value of the partial interval measurements in each sample as $\mathrm{E}(Y_s) = \pi_s$. The interim time distribution $G_s(x;\lambda_s)$ is bounded between 0 and 1, by which it follows that

$$0 \leq \int_0^c \left[1 - G_s(x;\lambda_s)\right] dx \leq c. \tag{19}$$

Substituting (19) into (2) leads to

$$\phi_s \leq \pi_s \leq \phi_s + \zeta_s c. \tag{20}$$

The right inequality in (3) follows. Assume that $0 < \mu_L^* < \mu_s$ for a known value μ^*. Multiplying (20) by $\mu_L^*/(\mu_L^* + c)$ and noting that $\phi_s = \mu_s\zeta_s$,

$$\frac{\mu_L^*\pi_s}{\mu_L^* + c} \leq \phi_s - \frac{c\phi}{\mu_L^* + c} + \frac{\mu_L^*\zeta c}{\mu_L^* + c} = \phi_s - \frac{c\zeta_s\left(\mu_s - \mu_L^*\right)}{\mu_L^* + c} \leq \phi_s, \tag{21}$$

which demonstrates the left inequality in (3).

Appendix B

A bound for incidence

Assume that $\mu_s \leq \mu_U^*$ for known μ_U^*. From the right side of (20), $\pi_s \leq \phi_s + \zeta_s c$. Dividing by $\mu_U^* + c$ and noting that $\phi_s = \mu_s \zeta_s$,

$$\frac{\pi_s}{\mu_U^* + c} \leq \frac{\mu_s + c}{\mu_U^* + c} \zeta_s \leq \zeta_s, \tag{22}$$

which demonstrates the left inequality in (8). Now assume that $G(x; \lambda_s) < p$ for known p. It follows that

$$(1 - p)c \leq \int_0^c [1 - G_s(x; \lambda_s)]\, dx. \tag{23}$$

Substituting (23) into (2) leads to

$$\pi_s \geq \phi_s + \zeta_s (1 - p)c \geq \zeta_s (1 - p)c. \tag{24}$$

The right inequality in (8) follows after dividing by $(1 - p)c$.

Appendix C

A bound for the log-interim ratio

Assume that $\mu_0 = \mu_1 = \mu$ and that $G_s(x; \lambda_s) = 1 - \exp(-x/\lambda_s)$ for $s = 0, 1$. It follows from (2) that $\pi_s = 1 - \lambda_s \exp(-c/\lambda_s) / (\mu + \lambda_s)$. For fixed π_s, write

$$\mu = f(\lambda_s, \pi_s) = \frac{\lambda_s \exp(-c/\lambda_s)}{1 - \pi_s} - 1$$

with inverse $\lambda_s = f^{-1}(\mu, \pi_s)$. Observe that

$$\begin{aligned} \lim_{\mu \to 0} f^{-1}(\mu, \pi_s) &= \frac{c}{-\ln(1 - \pi_s)} \\ \lim_{\mu \to \infty} \frac{f^{-1}(\mu, \pi_s)}{\mu} &= \lim_{\mu \to \infty} \frac{1 - \pi_s}{\exp[-c/f^{-1}(\mu, \pi_s)] - 1 + \pi_s} = \frac{1 - \pi_s}{\pi_s}, \end{aligned} \tag{25}$$

Denote the log-interim ratio as $\omega^\lambda = \ln(\lambda_0/\lambda_1)$ and note that it can be written as

$$\omega^\lambda(\mu, \pi_0, \pi_1) = \ln f^{-1}(\mu, \pi_0) - \ln f^{-1}(\mu, \pi_1).$$

It follows from (25) that

$$\begin{aligned} \lim_{\mu \to 0} \omega^\lambda(\mu, \pi_0, \pi_1) &= \mathrm{cll}(\pi_1) - \mathrm{cll}(\pi_0) \\ \lim_{\mu \to \infty} \omega^\lambda(\mu, \pi_0, \pi_1) &= \mathrm{logit}(\pi_1) - \mathrm{logit}(\pi_0). \end{aligned} \tag{26}$$

Next, note that the derivative of ω^λ with respect to μ is

$$\begin{aligned} \frac{\partial \omega^\lambda}{\partial \mu} &= \frac{1}{\lambda_0 \partial f(\lambda_0, \pi_0)/\partial \lambda_0} - \frac{1}{\lambda_1 \partial f(\lambda_1, \pi_1)/\partial \lambda_1} \\ &= \frac{\lambda_0}{c\mu + (\mu + c)\lambda_0} - \frac{\lambda_1}{c\mu + (\mu + c)\lambda_1}. \end{aligned} \tag{27}$$

Also note that

$$\frac{\partial f^{-1}(\mu, \pi_s)}{\partial \pi_s} = \frac{-\lambda_s(\mu + \lambda_s)^2}{[c\mu + (\mu + c)\lambda_s] \exp(-c/\lambda_s)} < 0. \tag{28}$$

Now suppose that $\pi_0 > \pi_1$. It follows from (28) that $\lambda_0 < \lambda_1$, from (27) that ω^λ is strictly decreasing in μ, and from (26) that

$$\mathrm{logit}(\pi_1) - \mathrm{logit}(\pi_0) < \omega^\lambda < \mathrm{cll}(\pi_1) - \mathrm{cll}(\pi_0). \tag{29}$$

Similarly, $\pi_0 \leq \pi_1$ implies that

$$\text{cll}(\pi_1) - \text{cll}(\pi_0) < \omega^{\lambda} < \text{logit}(\pi_1) - \text{logit}(\pi_0). \tag{30}$$

The limits given in (10) follow.

Appendix D

Variance of PIR data

This appendix provides a derivation of the variance of a PIR measurement given in (16), under the assumption that both the event durations and the interim times are exponentially distributed (i.e., an alternating Poisson process model). Let $Z(t) = 1$ indicate that an event is occurring at time t and $Z(t) = 0$ indicate that no event is occurring at time t. Formally,

$$Z(t) = \sum_{j=1}^{J} I\left[0 \leq t - \sum_{i=1}^{j-1}(A_i + B_i) < A_j\right].$$

Under the alternating Poisson process, $Z(t)$ is a continuous time Markov chain, having the property that

$$\begin{aligned} Pr(Z(s+t) = 1 | Z(s) = a, Z(r) : 0 \leq r < s) &= Pr(Z(s+t) = 1 | Z(s) = a) \\ &= Pr(Z(t) = 1 | Z(0) = a) \end{aligned} \tag{31}$$

for $a \in \{0, 1\}$ and $s, t \geq 0$ (Kulkarni, 2010, Thm. 6.1). Denote the transition probabilities of this continuous time Markov chain by

$$p_0(t) = Pr(Z(t) = 1 | Z(0) = 0) = \phi\left(1 - e^{-\rho t}\right) \tag{32}$$

$$p_1(t) = Pr(Z(t) = 1 | Z(0) = 1) = (1 - \phi)e^{-\rho t} + \phi, \tag{33}$$

(*ibid.*, Equation 6.17, p. 207) where the alternate parameterization $\rho = \frac{1}{\mu} + \frac{1}{\lambda} = \frac{\zeta}{\phi(1-\phi)}$ is employed for ease of notation. Under the assumption that the process is in equilibrium, $\Pr(Z(t) = 1) = \phi$ for any fixed t.

The variance of the Y is derived by evaluating $\mathrm{Cov}(U_h, U_k)$ for $1 \leq h < k \leq K$. Let $t_0 = (h-1)L/K$ denote the beginning of the h^{th} interval, $t_1 = t_0 + c$ denote the end of the active portion of the h^{th} interval, and $t_2 = (k-1)L/K$ denote the beginning of the k^{th}

interval. Observe that

$$\begin{aligned}
\Pr\left(Z(t_1)=1, U_h=1|Z(t_0)=1\right) &= \Pr\left(Z(t_1)=1|Z(t_0)=1\right) \\
&= p_1(c) = (1-\phi)e^{-\rho c} + \phi, \\
\Pr\left(Z(t_1)=1, U_h=1|Z(t_0)=0\right) &= \Pr(Z(t_1)=1 \cap Z(s)=1, t_0 \le s < t_1 | Z(t_0)=0) \\
&= \int_0^c p_1\left(c-x\right)\left[1-\exp\left(-\phi\rho x\right)\right] dx \\
&= \phi\left(1-e^{-\rho c}\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
\Pr\left(Z(t_1)=1 \cap U_h=1\right) &= \phi \Pr\left(Z(t_1)=1, U_h=1|Z(t_0)=1\right) \\
&\quad + (1-\phi)\Pr\left(Z(t_1)=1, U_h=1|Z(t_0)=0\right) = \phi \\
\Pr\left(Z(t_1)=1|U_h=1\right) &= \frac{\phi}{1-(1-\phi)e^{-\rho\phi c}}.
\end{aligned}$$

Conditioning on $Z(t_1)$,

$$\begin{aligned}
\Pr\left(Z(t_2)=1|U_h=1\right) &= \Pr\left(Z(t_1)=0|U_h=1\right) p_0(t_2-t_1) \\
&\quad + \Pr\left(Z(t_1)=1|U_h=1\right) p_1(t_2-t_1) \\
&= \phi + \frac{\phi(1-\phi)e^{-\rho(t_2-t_1)-\rho\phi c}}{1-(1-\phi)e^{-\rho\phi c}}.
\end{aligned}$$

Now conditioning on $Z(t_2)$,

$$\begin{aligned}
\Pr\left(U_k=1|U_h=1\right) &= \sum_{a=0}^{1} \Pr\left(Z(t_2)=a|U_h=1\right) Pr\left(U_k=1|Z(t_2)=a\right) \\
&= 1-e^{-\rho\phi c} + e^{-\rho\phi c}\Pr\left(Z(t_2)=1|U_h=1\right) \\
&= 1-(1-\phi_i)e^{-\rho\phi c} + \frac{\phi(1-\phi)e^{-\rho(t_2-t_1)-2\rho\phi c}}{1-(1-\phi)e^{-\rho\phi c}}.
\end{aligned}$$

It therefore follows that

$$\begin{aligned}
\mathrm{Cov}(U_h, U_k) &= \Pr(U_h=1)\left[\Pr\left(U_k=1|U_h=1\right) - \Pr(U_k=1)\right] \\
&= \phi(1-\phi)\exp\left[-\rho(k-h)L/K - (2\phi-1)\rho c\right].
\end{aligned}$$

Thus,

$$\begin{aligned}\mathrm{Var}\left(Y\right) &= \frac{1}{K}\mathrm{Var}(U_1) + \frac{2}{K^2}\sum_{k=1}^{K-1}(K-k)\mathrm{Cov}(U_1, U_{k+1}) \\ &= \frac{\mathrm{E}\left(Y\right)\left[1-\mathrm{E}\left(Y\right)\right]}{K} \times \left[1 + \frac{2\phi e^{(1-\phi)\rho c}}{K\mathrm{E}\left(Y\right)}\sum_{k=1}^{K-1}(K-k)\exp\left(\frac{-\rho k L}{K}\right)\right].\end{aligned}$$

Expression (16) follows by substituting $\frac{\zeta}{\phi(1-\phi)}$ for ρ and rearranging terms. Note that the derivation of this expression depends strongly on the independence of increments in the alternating Poisson process; this inhibits generalizations to alternating renewal processes based on event duration and interim time distributions other than the exponential.

Appendix E

Parametric bootstrap procedure for the Alternating Poisson Process

This appendix describes a parametric bootstrapping procedures for obtaining standard errors and confidence intervals for the moment estimators in Method 4. The parameter of interest θ is assumed to be some function of the behavior stream parameters, $\theta = f(\phi_0, \zeta_0, \phi_1, \zeta_1)$, so that the moment estimator $\hat{\theta}$ is calculated by replacing the parameters with corresponding moment estimators. For example, a moment estimator for the prevalence ratio is $\hat{\theta} = \hat{\phi}_1/\hat{\phi}_0$.

A parametric bootstrapping procedure with R replications involves the following:

1. For each $s = 0, 1$, simulate $R \times n_s$ behavior streams that follow an Alternating Poisson Process with mean event duration $\hat{\mu}_s$ and mean interim time $\hat{\lambda}_s$.

2. For each of the simulated behavior streams, apply the PIR procedure to generate summary measurements, $Y^r_{s1}, ..., Y^r_{sn_s}$, for $r = 1, ..., R$ and $s = 0, 1$. Calculate the sample mean and sample variance, $\tilde{y}^r_s$ and $(s^r_s)^2$, for $r = 1, ..., R$ and $s = 0, 1$.

3. Solve for the moment estimators based on each set of simulated sample moments, yielding estimates $\left(\hat{\phi}^r_s, \hat{\zeta}^r_s\right)$ for $r = 1, ..., R$ and $s = 0, 1$. Calculate an estimate for the parameter of interest $\hat{\theta}^r = f\left(\hat{\phi}^r_0, \hat{\zeta}^r_0, \hat{\phi}^r_1, \hat{\zeta}^r_1\right)$.

4. The sampling variance of $\hat{\theta}$ can be estimated by taking the variance over the bootstrap replicates $\hat{\theta}^1, ..., \hat{\theta}^R$. A $(1 - \alpha)$ CI for $\hat{\theta}$ can be formed by taking the $\alpha/2$ and $1 - \alpha/2$ sample quantiles of the bootstrap replicates.