Estimating the prevalence and incidence of a state behavior: Models for interval recording data and a novel observation system

James E. Pustejovsky and Daniel M. Swan

The University of Texas at Austin

Austin H. Johnson

University of Connecticut

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. Austin H. Johnson, Department of Educational Psychology, University of Connecticut.

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

Data based on direct observation of behavior are used widely in many areas of educational and psychological research, particularly in applied research areas such as the treatment of behavioral disorders. A number of different methods are used to record data during direct observation, including continuous recording, momentary time sampling (MTS), partial interval recording (PIR), and whole interval recording (WIR). Among these methods, PIR and WIR have long been recognized as problematic because, as typically reported, the mean of such data measures neither the prevalence nor the incidence of the observed behavior. Though the problems with these methods have long been recognized, little research has examined methods of analyzing interval recording data other than simply taking the mean. This paper proposes a Alternating Poisson Process model for interval recording data that permits estimation of both prevalence and incidence via maximum likelihood or penalized maximum likelihood methods. The paper also describes a novel observation recording method that involve combinations of MTS, PIR, and WIR and that provides considerably more efficient estimators of prevalence and incidence.

Keywords: behavioral observation; interval recording; alternating Poisson process; Markov chain

Estimating the prevalence and incidence of a state behavior: Models for interval recording data and a novel observation system

Measurements derived from systematic, direct observation of human behavior are used in many areas of psychological and educational research. For example, direct observation of student classroom behavior is a primary component of several existing instruments for screening and diagnosis of emotional and behavioral problems (Volpe, DiPerna, Hintze, & Shapiro, 2005); direct observation of childrens' challenging behavior in home settings has been employed to collect pre- and post-test measures in randomized trials of behavioral interventions (e.g., Durand, Hieneman, Clarke, Wang, & Rinaldi, 2012); and direct observation of infant-parent interaction patterns is employed in studies of child development (Mann, Ten Have, Plunkett, & Meisels, 1991) and cross-cultural differences (Bornstein, 2002). Direct observation also plays a prominent role in single-case research, where it is used to assess individual responses to intervention by measuring changes in behavior over time (Kazdin, 2011).

Systematic direct observation procedures require that the behavior of interest have a clear operational definition, so that its occurrence or absence can be judged at a given point in time. In forming such an operationally definition, is useful to distinguish between behaviors that are events, where each occurrence is of negligible duration, versus behaviors that are states, where individual episodes of behavior have positive duration (J. Altmann, 1974). The primary characteristic of an event behavior is its incidence, or frequency of occurrence per time unit. In contrast, a state behavior has two primary characteristics: incidence, which is the frequency (per unit time) with which new episodes of the behavior begin and prevalence, which is the proportion of time that the behavior occurs. Given an operationally defined behavior, measurements of its characteristics are obtained by recording data while observing the behavior (either in person, or by video-recording) for a specified length of time.

There are several different procedures for recording data during direct observation,

varying in ease of implementation, the level of detail in the resulting data, and the aspect of behavior to which the resulting measurement corresponds (for surveys of major recording procedures, see J. Altmann, 1974; Ayres & Gast, 2010; Hartmann & Wood, 1990; Primavera, Allison, & Alfonso, 1996). The most intensive procedure is continuous recording (sometimes called duration recording or real-time recording), in which the observer records the time at which each behavioral episode begins and ends. Data from continuous recording is very rich, in that it permits direct estimation of prevalence, incidence, and further aspects of the pattern of behavior; continuous recording data can also be subjected to more sophisticated forms of modeling (e.g., Bakeman & Quera, 2011; Haccou & Meelis, 1992). However, less effort-intensive data collection methods are often required, particularly for use in clinical and applied research settings.

Other commonly used systems for collecting behavioral data do not capture a complete record of the behavior during an observation session, but rather involve making observations only intermittently. Among intermittent recording systems, the three main procedures are momentary time sampling, partial interval recording, and whole interval recording. In all three methods, an observation session is divided into a fixed number of equally spaced intervals, of perhaps 10 or 15 s in length, and a binary data-point is recorded for each interval. The systems differ only in the rule for scoring each interval. Using momentary time sampling (MTS), an interval is scored as a one if a behavioral event is happening during the final moment of the interval (and is otherwise scored as a zero). Using partial interval recording (PIR, also known as one-zero sampling, modified frequency sampling, or Hansen sampling), an interval is scored as a one if the behavior occurs at any point during the interval. Using whole interval recording (WIR), an interval is scored as a one only if the behavior occurs for the entire duration of the interval. In some PIR and WIR systems, a small length of time is left between each interval so that the observer does not have to maintain continuous attention.

In many applications, the interval-by-interval data generated by these recording

systems is summarized by the proportion of intervals scored as a one. Often, only this summary proportion is used for later analysis, where it is interpreted as a measure of prevalence. However, whether it is reasonable to reduce the data to a summary proportion depends on which recording system is used to collect the data.

Under quite general modeling assumptions, the proportion of MTS intervals is an unbiased estimate of prevalence (Rogosa & Ghandour, 1991). Thus, reducing the data to the summary proportion is entirely reasonable if the investigator's interest is solely in the prevalence of the behavior. The drawback of doing so is that simple summaries of MTS data do not provide any measure of the behavior's incidence. Even if not of substantive interest, an estimates of incidence is needed in order to assess the magnitude of measurement error in the prevalence estimate.

Brown, Solomon, and Stephens (1977, see also Griffin & Adams, 1983) described a method for estimating both prevalence and incidence from MTS data. Their approach was to first posit a stochastic process for the underlying pattern of behavior as perceived by the observer, or what is often termed the behavior stream. The particular model they considered was an Alternating Poisson Process, which is a simple, two-state continuous time Markov chain where transitions between states follow exponential distributions. Brown and colleagues showed that if the behavior stream is generated by such a process, then interval-by-interval MTS scores follow a discrete-time Markov chain, from which closed-form expressions for the maximum likelihood estimators of prevalence and incidence can be derived.

Unlike MTS, PIR and WIR systems do not produce clearly interpretable summary measurements. Rather, the PIR summary proportion systematically over-estimates prevalence and the WIR summary proportion systematically under-estimates prevalence; in both cases, the extent of the bias depends on characteristics of the behavior as well as operational features of the recording system (Kraemer, 1979; Rogosa & Ghandour, 1991), making the construct interpretation of such data quite difficult. Consequently,

methodologists have long argued against the use of PIR and WIR systems (J. Altmann, 1974; Lane & Ledford, 2014; Mann et al., 1991, cf.). Despite such objections, the systems remain in common use, particularly as part of behavioral time series designs and single-case research (Lane & Ledford, 2014; Mudford, Taylor, & Martin, 2009; Rapp et al., 2007).

Little previous research has considered methods of analyzing PIR and WIR data beyond using the summary proportion. For PIR data, S. A. Altmann and Wagner (1970) proposed a transformation of the summary proportion as an estimate of incidence, motivated by a model in which behavioral episodes follow a Poisson process. While this model applies well to event behaviors, it is not a suitable description of state behaviors, where individual episodes have non-negligible duration. Suen and Ary (1986, 1989) have proposed a method for obtaining estimates of prevalence and incidence from PIR data, provided that the behavior stream conforms to certain conditions. However, their proposed procedure is not motivated by any explicit data-generating process, and later simulation studies reported that the method produces badly biased estimates (Rogosa & Ghandour, 1991, sec. 5.2). Pustejovsky and Swan (2014) proposed several methods for bounding the bias of the PIR summary proportion as an estimate of prevalence, based on various prior assumptions about the behavior stream. These methods are useful for analysis of summarized PIR data, as would be available from a published single-case study, but are not full models of the data-generating process. Other methods of analysis, involving fully specified data-generating models for the interval-by-interval scores, are therefore of interest.

This paper examines models for PIR and WIR data, from which principled estimates of prevalence and incidence can be obtained. Following Brown et al. (1977), we use an Alternating Poisson Process for the underlying behavior stream to derive a model for the interval-by-interval scores. Under this model, maximum likelihood estimates for prevalence and incidence can be obtained using conventional numerical

techniques (although they do not have closed-form expression). To remedy some problems with the maximum likelihood estimators, we introduce penalized likelihood estimators that have better operating characteristics and that can be tailored to express prior information about the behavioral parameters. The final section of the paper describes a novel procedure for intermittent recording of a behavior that entails combining MTS and interval recording methods and that can be used to obtain more efficient estimators of prevalence and incidence.

We use a parametric bootstrap procedure for obtaining interval estimates.

Alternating Poisson Process models

The Alternating Poisson Process is a stochastic model that can be used to describe a stream of behavior, as it is perceived in time. The model applies to state behaviors, where the behavior is either occurring or not occurring at any given point in time and where each episode of behavior has non-negligible duration. The stream of a state behavior can be described in terms of two components: sequentially ordered, non-overlapping episodes of behavior, which we will call event durations, and spans of time in between episodes, which we will call interim times. Let $\{Z(t), 0 \leq t\}$ denote the state of the behavior stream over the course of an observation session, where $Z(t) = 1$ indicates that an event is occurring at time t and $Z(t) = 0$ otherwise.

The Alternating Poisson Process makes several further assumptions. Specifically, it is assumed that event durations and interim times are mutually independent, random quantities, that the event durations follow an exponential distribution with mean $\mu > 0$, and that the interim times follow an exponential distribution with mean $\lambda > 0$. Under the model, the prevalence of the behavior is equal to the ratio of μ to the sum of μ and λ and the incidence of the behavior is equal to the reciprocal of the sum of μ and λ. We will denote prevalence by ϕ, where $0 < \phi < 1$, and incidence by ζ, where $\zeta > 0$. Finally, it is assumed that the process is in equilibrium, with $\Pr(Y(0) = 1) = \phi$. This assumption implies that there is a constant marginal probability of observing an event at any given point in time.

The Alternating Poisson Process is a special case of a continuous time Markov chain, and thus has the Markov property that the future evolution of the behavior depends only on the current state, but not on the past history of the behavior. More precisely, the probability that a behavior will be occurring t seconds into the future is independent of the state of the behavior for $0 \leq r < s$:

$$\Pr\left[Z(s+t) = 1 \,|Z(s) = a, Z(r) : 0 \leq r < s\right] = \Pr\left[Z(s+t) = 1 \,|Z(s) = a\right] \tag{1}$$

for $a = 0, 1$ and $s, t \geq 0$ (Kulkarni, 2010, Thm. 6.1). The assumption that the process is in equilibrium further implies that the probability that a behavior will be occurring t seconds into the future does not depend on the current time, i.e.,

$$\Pr\left[Z(s+t) = 1 \,|Z(s) = a\right] = \Pr\left[Z(t) = 1 \,|Z(0) = a\right]. \tag{2}$$

Let $p_a(t)$ denote the conditional probability that an event will be occurring t seconds into the future, given that the behavior is currently in state a, for $a = 0, 1$. These conditional probabilities can be expressed as follows:

$$\begin{aligned} p_0(t) &= \Pr(Z(t) = 1 | Z(0) = 0) = \phi \left[1 - \exp\left(\frac{-t\zeta}{\phi(1-\phi)}\right)\right] \\ p_1(t) &= \Pr(Z(t) = 1 | Z(0) = 1) = \phi + (1-\phi)\exp\left(\frac{-t\zeta}{\phi(1-\phi)}\right) \end{aligned} \tag{3}$$

(Kulkarni, 2010, Eq. 6.17).

Momentary Time Sampling

Consider observing a behavior stream generated by the Alternating Poisson Process and recording observations using momentary time sampling with $K + 1$ recording times, equally spaced at intervals of length c. Denote the recorded data by the sequence of binary indicator variables $X_0, X_1, ..., X_K$. The MTS interval data are a record of the state of the behavior stream process at fixed moments in time: $X_k = Z(ck)$ for $k = 0, ..., K$.

Brown et al. (1977) demonstrated that MTS data follow a two-state, discrete-time Markov chain process with transition probabilities $Pr(X_k = 1 | X_{k-1} = a) = p_a(c)$ and

$Pr(X_k = 0|X_{k-1} = a) = 1 - p_a(c)$ for $a = 0, 1$. Therefore, sufficient statistics for the process are given by the table counting the number of transitions with $(X_{k-1} = a, X_k = b)$ for $a, b = 0, 1$ and $k = 1, ..., K$; let $n_{ab} = \sum_{k=1}^{K} I(X_{k-1} = a, X_k = b)$. Conditioning on X_0, the log-likelihood of MTS data is then given by

$$\begin{aligned} l_{MTS}(\phi, \zeta) = n_{01} \log \phi + n_{10} \log (1 - \phi) \\ + (n_{01} + n_{10}) \log \left[1 - \exp \left(\frac{-\zeta c}{\phi(1 - \phi)} \right) \right] \\ + n_{00} \log \left[1 - \phi + \phi \exp \left(\frac{-\zeta c}{\phi(1 - \phi)} \right) \right] \\ + n_{11} \log \left[\phi + (1 - \phi) \exp \left(\frac{-\zeta c}{\phi(1 - \phi)} \right) \right]. \end{aligned} \tag{4}$$

Brown et al. (1977) provided closed-form expressions for the maximum likelihood estimates (MLEs) for ϕ and ζ based on this model. Let $\hat{p}_0 = n_{01} / (n_{00} + n_{01})$ and $\hat{p}_1 = n_{11} / (n_{10} + n_{11})$. The MLE for ζ exists only when $\hat{p}_0 < \hat{p}_1$. When this condition holds, the MLEs for ϕ and ζ are given by

$$\hat{\phi}_{MTS} = \frac{\hat{p}_0}{\hat{p}_0 + 1 - \hat{p}_1} \qquad \text{and} \qquad \hat{\zeta}_{MTS} = \frac{-\hat{p}_0 (1 - \hat{p}_1) \log(\hat{p}_1 - \hat{p}_0)}{c (\hat{p}_0 + 1 - \hat{p}_1)^2}. \tag{5}$$

The probability that the MLEs are undefined or fall outside of the parameter space is not trivial, even when K is relatively large. In order for the estimates to fall

Table 1

Proportion of 2000 simulated MTS samples ($K = 40$) in which $0 < \hat{\phi}_{MTS} < 1$ and $0 < \hat{\zeta}_{MTS} < \infty$.

	$\zeta = 0.02$	0.05	0.1	0.2	0.25	0.4	0.5
$\phi = 0.1$	0.43	0.63	0.64	0.45	0.37	0.29	0.26
0.2	0.45	0.80	0.90	0.84	0.79	0.59	0.49
0.3	0.46	0.84	0.96	0.96	0.94	0.73	0.64
0.4	0.43	0.87	0.98	0.99	0.97	0.82	0.71
0.5	0.45	0.88	0.99	1.00	0.98	0.84	0.71

strictly within the parameter space, both a 0-1 transition and a 1-0 transition must be observed, so that $\hat{p}_0 > 0$, $\hat{p}_1 < 1$. Table 1 reports the proportion of 2000 simulated samples in which $0 < \hat{\phi}_{MTS} < 1$ and $0 < \hat{\zeta}_{MTS} < \infty$, with $K = 40$ and ζ scaled in terms of the interval length. Values of $\phi > 0.5$ are omitted because the behavior of the MTS estimators is symmetric about $\phi = 0.5$. The proportion of estimates falling within the parameter space decreases as prevalence becomes more extreme and as incidence becomes very infrequent or very frequent. The high proportion of samples in which the estimate of incidence is undefined represents a drawback to the use of maximum likelihood based on MTS data.

Partial Interval Recording

Consider observing a behavior stream generated by the Alternating Poisson Process and recording observations using partial interval recording. Suppose that one observes K intervals, where each interval includes c seconds of active observation time followed by d seconds of recording time. Let time $t_k = (k-1)(c+d)$ denote the beginning of interval k. Let U_k indicate the PIR score from interval k, corresponding to the time from t_k to $t_k + c$. Following the PIR system, $U_k = 1$ if is the behavior occurs at any point during the active portion of interval, and $U_k = 0$ otherwise. In terms of the behavior stream process,

$$U_k = I\left[0 < \int_0^c Z\left(t_k + s\right) ds\right] \tag{6}$$

for $k = 1, ..., K$, where $\int_0^c$ denote the definite integral over the half-open interval $[0, c)$.

Under the assumptions of the Alternating Poisson Process, the joint distribution of $U_1, ..., U_K$ can be derived as follows. Let $\psi_k, k = 2, ..., K$ denote the probability than the behavior is occurring at time $t_k = (k-1)(c+d)$, given the partial interval record up to that time. Let $\psi_1 = \phi$, which follows from the assumption that the process is in

equilibrium. We show in Appendix A that

$$\begin{aligned}\psi_k &= \Pr\left[Z(t_k) = 1 \,|U_1, ..., U_{k-1}\right] \\ &= \left[\frac{\psi_{k-1}p_1(c+d) + (1-\psi_{k-1})\left[p_0(c+d) - p_0(d)\exp\left(\frac{-\zeta c}{1-\phi}\right)\right]}{1-(1-\psi_{k-1})\exp\left(\frac{-\zeta c}{1-\phi}\right)}\right]^{u_{k-1}} [p_0(d)]^{(1-u_{k-1})}.\end{aligned} \tag{7}$$

Note that $Z(t_k) = 1$ implies that $U_k = 1$ with certainty, while

$$\Pr\left(U_k = 1 \,|Z(t_k) = 0\right) = 1 - \exp\left(\frac{-\zeta c}{1-\phi}\right).$$

It follows from the Markov property of the Alternating Poisson Process that

$$\begin{aligned}\Pr\left(U_k = 1 \,|U_1, ..., U_{k-1}\right) &= \psi_k \Pr\left(U_k = 1 \,|Y(t_k) = 1\right) + (1-\psi_k)\Pr\left(U_k = 1 \,|Y(t_k) = 0\right) \\ &= 1 - (1-\psi_k)\exp\left(\frac{-\zeta c}{1-\phi}\right).\end{aligned}$$

The joint distribution of $U_1, ..., U_K$ can therefore be expressed as

$$\begin{aligned}\Pr\left(U_1 = u_1, ..., U_K = u_K\right) &= \Pr\left(U_1 = u_1\right)\prod_{k=2}^{K}\Pr\left(U_k = u_k \,|U_1, ..., U_{k-1}\right) \\ &= \prod_{k=1}^{K}\left[1-(1-\psi_k)\exp\left(\frac{-\zeta c}{1-\phi}\right)\right]^{u_k}\left[(1-\psi_k)\exp\left(\frac{-\zeta c}{1-\phi}\right)\right]^{(1-u_k)}.\end{aligned}$$

The log-likelihood of ϕ and ζ, given observed PIR data $u_1, ..., u_K$, is

$$l_{PIR}\left(\phi, \zeta\right) = \sum_{k=1}^{K} u_k \ln\left[1-(1-\psi_k)\exp\left(\frac{-\zeta c}{1-\phi}\right)\right] + (1-u_k)\left[\ln\left(1-\psi_k\right) - \frac{\zeta c}{1-\phi}\right]. \tag{8}$$

MLEs $\hat{\phi}_{PIR}, \hat{\zeta}_{PIR}$ are obtained by maximizing l_{PIR} using numerical techniques. Because the conditional probabilities $\psi_1, ..., \psi_K$ are defined recursively, it is cumbersome and computationally expensive to evaluate the score function corresponding to this likelihood. The simulation study reported in a section therefore uses the Nelder-Mead algorithm (Nelder & Mead, 1965), which does not require evaluation of the score function.

Just as with MTS data, MLEs based on PIR data do not always fall within the parameter space. Table 2 reports the proportion of 2000 simulated samples in which the

MLEs based on PIR data are within the parameter space, with $K = 40$, $d = 0$, and ζ scaled in terms of the interval length. Because we use numerical maximization, the results of the maximization routine are never precisely on the borders of the parameter space. We therefore use boundaries of $|\text{logit } \hat{\phi}_{PIR}| < 8$ and $|\log \hat{\zeta}_{PIR}| < 8$ to define the edges of the parameter space. The proportion of estimates falling within the parameter space is highest for moderate values of prevalence and incidence ($0.2 \leq \phi \leq 0.5$ and $0.1 \leq \zeta \leq 0.25$), decreases as prevalence becomes more extreme, and decreases as incidence becomes either less frequent (less than once per ten intervals) or more frequent (more than once per four intervals). Unlike the MTS estimators, the pattern of boundary estimates is asymmetric because PIR tends to reach ceiling levels when prevalence is large.

In addition to returning estimates that are on the edges of the parameter space, the MLEs based on PIR data have the further disadvantage of being somewhat sensitive

Table 2

Proportion of 2000 simulated PIR samples ($K = 40$) in which $|logit(\hat{\phi}_{PIR})| < 8$ and $|log(\hat{\zeta}_{PIR})| < 8$.

	$\zeta = 0.02$	0.05	0.1	0.2	0.25	0.4	0.5
$\phi = 0.1$	0.59	0.83	0.94	0.94	0.90	0.78	0.72
0.2	0.67	0.91	0.98	0.99	0.98	0.92	0.84
0.3	0.74	0.94	0.99	1.00	1.00	0.94	0.87
0.4	0.78	0.96	1.00	1.00	0.99	0.93	0.86
0.5	0.77	0.96	0.99	0.99	0.98	0.89	0.80
0.6	0.75	0.94	0.97	0.96	0.92	0.79	0.67
0.7	0.68	0.87	0.92	0.85	0.78	0.56	0.38
0.8	0.60	0.76	0.77	0.56	0.42	0.16	0.09
0.9	0.47	0.48	0.32	0.08	0.04	0.00	0.00

to initialization values. The likelihood surface becomes very flat when the PIR scores are near ceiling or floor levels, making it difficult to numerically identify the maximum. For the implementation of the estimators in the accompanying R package, the consequence is that drastically different estimates can be returned depending on the initialization values of the algorithm. Together with the possibility of obtaining estimates on the edges of the parameter space, the numerical instability of the MLEs motivates our investigation of alternative estimators that incorporate penalty functions.

Whole Interval Recording

Consider observing a behavior stream generated by the Alternating Poisson Process and recording observations using whole interval recording. As with PIR, suppose that one observes K intervals, where each interval includes c seconds of active observation time followed by d seconds of recording time. Let W_k indicate the WIR score from interval k, corresponding to the time from t_k to $t_k + c$. Following the WIR system, $W_k = 1$ if is the behavior occurs for the duration of the active portion of interval, and $W_k = 0$ otherwise. Formally,

$$W_k = I\left[c = \int_0^c Z\left(t_k + s\right) ds\right] \tag{9}$$

for $k = 1, ..., K$.

Using the WIR system to score a state behavior is logically equivalent to using PIR to score the absence of the behavior. WIR data can therefore be modeled just as PIR data, after an appropriate change of parameters. Specifically, the log-likelihood for WIR data under the Alternating Poisson Process can be written in terms of the log-likelihood for PIR data as

$$l_{WIR}\left(\phi, \zeta | W_1 = w_1, ..., W_k = w_k\right) = l_{PIR}(\,1 - \phi, \zeta | U_1 = 1 - w_1, ..., U_K = 1 - w_k)\,. \tag{10}$$

The equivalence of the two system implies that estimates of prevalence and incidence based on WIR data can be obtained using the algorithms developed for PIR.

Augmented interval recording

Thus far, we have considered conventional and widely used procedures for intermittent, systematic direct observation procedures. We now describe a novel recording procedure that might provide more accurate estimates of both prevalence and incidence. The method, which we call augmented interval recording, involves using MTS, PIR, and WIR systems on each interval. To the best of our knowledge, this procedures has not been previously described in the literature on systematic direct observation of behavior.

Just as with PIR or WIR, suppose that the observation session is divided into K intervals and that the first c seconds of the interval are devoted to observation while the remaining d seconds are devoted to recording or resting; interval k therefore begins at time $t_k = (k-1)(c+d)$.

Consider an observer who uses the combination of MTS, PIR, and WIR scoring rules for each interval during an observation session. Doing so requires that the observer record sufficient data so that the values of the MTS, PIR, and WIR variables (X_{k-1}, U_k, W_k) for each interval. Figure 1 depicts the sequence of questions to be answered during interval k in order to completely determine these values. For each interval, the observer begins by noting the presence or absence of the behavior at time t_k and recording the MTS score. If the behavior is present ($X_{k-1} = 1$), then the partial interval record is also determined ($U_k = 1$), and it only remains to determine whether the behavior occurs for the duration of the interval ($W_k = 1$) or ends before the start of the next interval ($W_k = 0$). Similarly, if the behavior is absent at the start of the interval ($X_{k-1} = 0$), then the whole interval record is also determined ($W_k = 0$), and it only remains to determine whether a behavioral event begins before the start of the next interval ($U_k = 1$) or is absent for the entire interval ($U_k = 0$).

The AIR procedure requires only marginally more effort on the part of the observer than an interval recording method used alone. One measure of effort is the

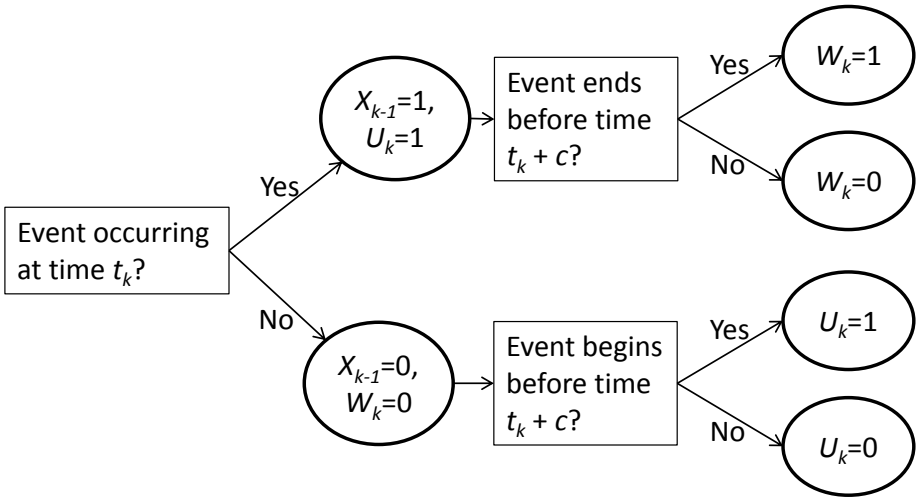

Figure 1. Procedure for combining MTS and interval recording

level of sustained attention required on the part of the observer. Because the sustained attention needed for interval recording also entails the attention needed for momentary time sampling, the additional effort is minimal in this respect. Another measure of effort is the amount of data that must be recorded during the observation period. Because $W_k = 0$ is implied when $X_{k-1} = 0$ and $U_k = 1$ is implied when $X_{k-1} = 1$, AIR requires twice as much data as one of the single methods (rather than three times as much, might be supposed). Thus, for a fixed interval length, simultaneous use of all three methods entails at most twice as much effort as interval recording alone. Furthermore, using longer time-intervals with fewer intervals per observation period would mitigate the effort required.

Under the assumptions of the Alternating Poisson Process, the data generated by the AIR system can be modeled using a discrete-time Markov Chain, from which estimates of prevalence and incidence can be obtained. The Markov property of the Alternating Poisson Process implies that the joint distribution can be written as

$$\begin{aligned}\Pr\left(X_0 = x_0, X_1 = x_1, U_1 = u_1, W_1 = w_1, ..., X_K = x_K, U_K = u_K, W_K = w_K\right) \\ = \Pr\left(X_0 = x_0\right) \prod_{k=1}^{K} \Pr\left(X_k = x_k, U_k = u_k, W_k = w_k | X_{k-1} = x_{k-1}\right). \quad (11)\end{aligned}$$

Denote the transition probabilities $\Pr\left(X_k = b, U_k = c, W_k = d | X_{k-1} = a\right) = \pi_{a|bcd}$ and let $m_{a|bcd} = \sum_{k=1}^{K} I\left(X_{k-1} = a, X_k = b, U_k = c, W_k = d\right)$ for $a, b, c, d = 0, 1$. Conditional on X_0, the log-likelihood of the observed AIR data is given by

$$l_{AIR}\left(\phi, \zeta\right) = \sum_{a=0}^{1}\sum_{b=0}^{1}\sum_{c=0}^{1}\sum_{d=0}^{1} m_{a|bcd} \log \pi_{a|bcd} \tag{12}$$

where

$$\begin{aligned}
\pi_{0|000} &= \left[1 - p_0(d)\right] \exp\left(\frac{-\zeta c}{1-\phi}\right) \\
\pi_{0|010} &= 1 - p_0(c+d) - \left[1 - p_0(d)\right] \exp\left(\frac{-\zeta c}{1-\phi}\right) \\
\pi_{0|100} &= p_0(d) \exp\left(\frac{-\zeta c}{1-\phi}\right) \\
\pi_{0|110} &= p_0(c+d) - p_0(d) \exp\left(\frac{-\zeta c}{1-\phi}\right) \\
\pi_{1|010} &= 1 - p_1(c+d) - \left[1 - p_1(d)\right] \exp\left(\frac{-\zeta c}{\phi}\right) \\
\pi_{1|011} &= \left[1 - p_1(d)\right] \exp\left(\frac{-\zeta c}{\phi}\right) \\
\pi_{1|110} &= p_1(c+d) - p_1(d) \exp\left(\frac{-\zeta c}{\phi}\right) \\
\pi_{1|111} &= p_1(d) \exp\left(\frac{-\zeta c}{\phi}\right)
\end{aligned}$$

and the remaining transition probabilities are all equal to zero. See Appendix B for the derivation of these quantities. As with PIR data, the MLEs $\hat{\phi}_{AIR}, \hat{\zeta}_{AIR}$ are obtained by maximizing l_{AIR} using the Nelder-Mead algorithm.

Penalized likelihood estimators

The previous section has illustrated that maximum likelihood estimates derived from MTS data or from interval recording data have undesirable operating characteristics when based on a moderate sample size of $K = 40$ intervals. In this section, we consider the use of penalized likelihood estimators (PLEs) to stabilize the behavior of the estimators and improve their performance in moderately sized samples. PLEs are derived by maximizing the sum of the log likelihood and a penalty term that

depends on the parameters. Penalized likelihood estimation has been applied to an array of statistical problems where maximum likelihood methods break down, such as estimation of logistic regression coefficients in small samples (Galindo-Garre, Vermunt, & Bergsma, 2004; Gelman, Jakulin, Pittau, & Su, 2008), estimation of correlation structure in high-dimensional multivariate regression models (Warton, 2008), and estimation of variance components in multi-level models when the number of highest-level units is small (Chung, Rabe-Hesketh, Dorie, Gelman, & Liu, 2013).

PLEs can be derived from a Bayesian perspective, by assigning a prior distribution to the model parameters and taking the posterior mode (given the prior and the data) as a point estimator (Chung et al., 2013). The penalty function is thus equivalent to the log of the prior distribution. An advantage of PLEs is that they provide a convenient and coherent way to incorporate into the estimation process prior information about the behavior.

In most research contexts, we expect that prior knowledge about characteristics of the behavior will be more readily expressed in terms of the average event duration (μ) and average interim time (λ). We therefore consider a class of priors in which μ and λ follow independent gamma distributions, with

$$\mu \sim Gamma\left(\alpha_\mu, c\theta_\mu\right), \qquad \lambda \sim Gamma\left(\alpha_\lambda, c\theta_\lambda\right),$$

with hyperparameters $\alpha_\mu, \alpha_\lambda > 1$ and $\theta_\mu, \theta_\lambda > 0$. Note that the priors on μ and λ are scaled in terms of the active interval length, so that they do not depend on the time unit in which the parameters and interval length are measured. This class of priors has the useful property that the implied priors on prevalence (ϕ) and incidence (ζ) have familiar distributional forms. Specifically, when $\theta_\mu = \theta_\lambda = \theta$, it follows that

$$\phi \sim Beta\left(\alpha_\mu, \alpha_\lambda\right), \qquad \zeta \sim Gamma^{-1}\left(\alpha_\mu + \alpha_\lambda, \frac{1}{c\theta}\right)$$

and that ϕ is independent of ζ.

The penalty function implied by these priors depends on how they are parameterized (i.e., based on priors for μ, λ or priors for ϕ, ζ). We recommend using the (μ, λ) parameterization because it reduces to zero when $\alpha_\mu = \alpha_\lambda = 1$ and $\theta_\mu = \theta_\lambda = \infty$, making the PLEs equivalent to the MLEs. With this parameterization, the penalty function has the form

$$p(\phi, \zeta) = (\alpha_\mu - 1)\log(\phi) + (\alpha_\lambda - 1)\log(1-\phi) - (\alpha_\mu + \alpha_\lambda - 2)\log(\zeta) - \frac{\frac{\phi}{\theta_\mu} + \frac{1-\phi}{\theta_\lambda}}{c\zeta}. \quad (13)$$

Given data based on recording system $s \in \{MTS, PIR, WIR, AIR\}$, the PLEs $\tilde{\phi}_s, \tilde{\zeta}_s$ are defined as the values that maximize $l_s(\phi, \zeta) + p(\phi, \zeta)$.

Application of penalized likelihood estimators requires the analyst to choose values for the hyperparameters of the prior distribution. If one does not have specific prior knowledge regarding the characteristics of the behavior stream, it is prudent to choose hyperparameters that have little influence on the values of the PLEs. These weak priors imply a penalty function that is relatively flat, so that PLEs will correspond closely with the MLEs except when the data contain little information about the parameters. We suggest that $\alpha_\mu = \alpha_\lambda = 1.5$ and $\theta_\mu = \theta_\lambda = 10$ are reasonable default choices for hyperparameters. The priors are highest at $\mu = \lambda = 5c$, or an average event duration and an average interim time of 5 intervals; the inter-quartile ranges are from $6.1c$ to $20.5c$. The implied prior for prevalence is symmetric about $\phi = 0.5$; its use amounts to adding the information from observing one independent moment where the probability of observing behavior is 0.5. We examine the empirical performance of the PLEs with these default hyperparameters in the next section.

In some research contexts, one may have fairly strong prior knowledge about certain behavioral characteristics, which can be used to inform the choice of hyperparameters. For example, prior experience with a class of behavior may suggest that it is unlikely for the average event duration to be more than two intervals in length, and that the most likely value for the average event duration is $\frac{1}{2}$ an interval length. This suggests choosing $\alpha_\mu = 3$ and $\theta_\mu = \frac{1}{4}$, so that the prior mode of μ is

$(\alpha_\mu - 1)\, c\theta_\mu = \frac{c}{2}$ and $\Pr(\mu > 2c) < 0.02$. Absent strong prior information about λ, one might use the default hyperparameters suggested above, taking $\alpha_\lambda = 1.5, \theta_\lambda = 10$.

Finite-sample performance

We noted in a previous section that maximum likelihood estimates derived from MTS or PIR data are not always well-defined, even when the number of intervals is moderate. To remedy this problem, we have proposed the use of penalized likelihood estimates that are always well-defined and numerically stable. These approaches to estimation should produce equivalent results when based on very long observation sessions with many intervals of data, but they may differ when the number of intervals is more limited.

Both maximum likelihood and penalized likelihood estimation represent alternatives to the standard method of summarizing intermittent behavioral observation data, which is to use the summary proportion of intervals with behavior. For MTS data, the summary proportion is an unbiased estimate of prevalence under a very broad class of data-generating models (Rogosa & Ghandour, 1991); however, the summary proportion may be less efficient than the PLE or MLE for prevalence under the Alternating Poisson Process model. For PIR data, the summary proportion is biased as an estimate of prevalence (Rogosa & Ghandour, 1991), while likelihood-based methods provide approximately unbiased estimates for sufficiently long observation sessions. Still, for a fixed sample size, the variability of the likelihood-based estimates may be worse than the bias in the summary proportion. It is therefore important to compare the accuracy of all of these estimators.

In order to understand the operating characteristics of the MLEs and PLEs in samples with a finite number of intervals, we conducted a computer simulation study. The simulations examined three specific questions:

1. For a given recording procedure (MTS or PIR), how does the accuracy of the PLEs compare to the accuracy of the MLEs (for both prevalence and incidence)

Table 3

Simulation design

Parameter	Definition	Levels	Min.	Step	Max.
s	Recording system	3	MTS, PIR, AIR		
K	Session length	15	10	10	150
ϕ	Prevalence	19	0.05	0.05	0.95
ζ	Incidence	10	0.05	0.05	0.50

and the accuracy of the summary proportion (for prevalence only)?

2. For a given recording procedure, how large a sample is needed to obtain approximately unbiased estimates of prevalence or incidence using penalized likelihood methods?

3. How does the accuracy of estimates based on the novel AIR procedure compare to that of estimates based on MTS or PIR data?

Table 3 summarizes the simulation design. We simulated data based on three different recording systems: MTS, PIR, and AIR; WIR was omitted because it is equivalent to using PIR for the absence of the behavior. For MTS and PIR, we used $c = 1, d = 0$ and varied the length of the observation session from 10 to 150 intervals. This range spans a variety of situations in which intermittent behavioral observation recording might be used, from a quick observation in a classroom where the observer needs to capture the behaviors of several children to an intensive observation of a single child over the course of an entire class period. To provide for a fair comparison with the conventional recording systems, we simulated AIR data using $K/2$ intervals of length $c = 2$ (i.e., twice as long as those for MTS or PIR). We varied the true prevalence of the behavior stream across nearly its entire possible range. Because $c = 1$ for MTS and PIR, incidence is scaled in terms of the length of an interval; for example, $\zeta = 0.1$ corresponds

to an incidence of one new behavioral episode per ten intervals. We varied incidence between $\zeta = .05$ (one new behavior per 20 intervals) and $\zeta = .50$ (one new behavior every 2 intervals) because this represents a range of behaviors where intermittent recording procedures might feasibly be applied; in particular, PIR measurements would quickly approach ceiling levels when behaviors occur more frequently than once per two intervals.

We implemented the simulations using the ARPobservation package (Pustejovsky, 2014) for the R statistical computing environment (R Core Team, 2014). For each combination of parameter values (ϕ, ζ, K) and recording system, we generated 10000 behavior streams of length K from an Alternating Poisson Process, then simulated data based on the specified recording system. For each string of simulated data, we calculated the summary proportion of intervals and then found the MLEs and PLEs using numerical maximization. For the PLEs, we used the default priors of $\alpha_\mu = \alpha_\lambda = 1.5$ and $\theta_\mu = \theta_\lambda = 10$.

Depending on the research context, an analyst using these estimators might want to use the natural parameterization of ϕ and ζ or the transformed parameterization of $\text{logit}(\phi)$ and $\log(\zeta)$, where $\text{logit}(x) = \log(x) - \log(1 - x)$. The latter parameterization puts the scores on a scale from $-\infty$ to ∞, and so might be preferred by an analyst seeking to fit a linear model. Consequently, we studied each research question using both parameterizations. To address the first research question, we examined the root mean-squared error (RMSE) of the prevalence estimates and the relatively RMSE of the incidence estimates; we focused on relative RMSE, defined as $\text{E}\left[\left(\hat{\zeta} - \zeta\right)^2\right]/\zeta$, because incidence is a rate. In the transformed parameterization, we examined the RMSE of $\text{logit}(\phi)$ and $\log(\zeta)$.

In addressing the second research question, regarding required minimum sample sizes to obtain approximately unbiased estimates, we used rather liberal criteria for bias. We believe that doing so is appropriate given the lack of alternative methods,

particularly for estimating incidence. Our criteria for "approximate unbiasedness" are defined as follows. For prevalence, we used absolute bias in $\hat{\phi}$ of less than 0.03 and absolute bias in logit($\hat{\phi}$) of less than 0.05. For incidence, we used absolute relative bias in $\hat{\zeta}$ of less than 0.10 and absolute bias in $\log(\hat{\zeta})$ of less than 0.10. While the criteria for incidence may seem especially liberal, in practical terms estimates that biased might still of interest. If the true incidence of a behavior indicates that behavior is occurring on average once every 20 intervals ($\zeta = 0.05c$), then an estimator with a relative bias of as much as 0.10 could return an estimate that indicates the behavior is happening between approximately once every 18 and once every 22 intervals on average. If the true incidence of the behavior indicates that the behavior is occurring once every 2 intervals on average ($\zeta = 0.50c$), then an estimator with a relative bias of as much as 0.10 could return an estimate suggesting that the behavior occurs between once every 1.8 and once every 2.2 intervals on average.

Are we also (implicitly) arguing that prevalence estimates that are within 3-5 percentage points (not sure precisely how to characterize the logit bias) are still practically of interest?

Results: MTS

Figure 2 illustrates the distribution of RMSE of the MLEs, PLEs, and summary proportions based on MTS data, for varying levels of K; each row of graphs corresponds to a different parameter. Given that the MTS proportions are known to be unbiased estimates of prevalence, those estimates provide a point of reference for the accuracy of the MLEs and PLEs. Across all levels of K, the MLEs of ϕ have comparable RMSEs to the MTS proportions. However, the MLEs perform notably worse in terms of $logit(\phi)$ until $K = 120$ or more, a very large number of intervals. Across the range of K, the PLEs of ϕ and $logit(\phi)$ have RMSE comparable to or smaller than the MTS proportions. In terms of both ζ and $log(\zeta)$, the MLEs perform considerably worse than the PLEs across all levels of K. However, the PLEs for ζ still have rather large relative RMSE (and similarly, the PLEs for $\log(\zeta)$ have large RMSE) for some parts of the parameter space. The PLEs for ζ perform poorly when incidence is very low, or when the prevalence is either very high or very low and incidence is moderate to high.

I added this line at your request, but it's essentially repeating the description in the paragraph below.

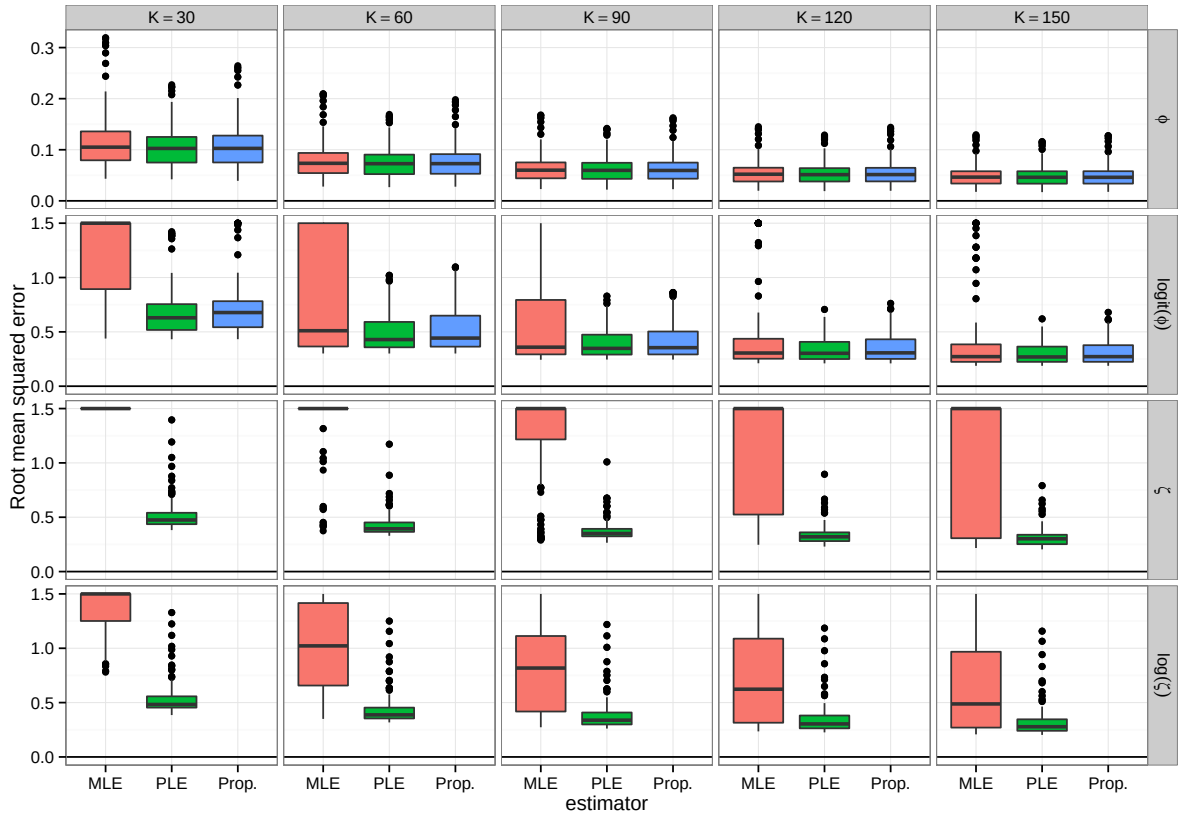

Figure 2. Distribution of root mean-squared error of MLE, PLE, and summary proportion (Prop.) estimates based on MTS data, across range of $0.05 \leq \phi \leq 0.95$ and $0.05 \leq \zeta \leq 0.50$. Values above 1.5 are not displayed. For ζ, the root mean-squared error is relative to the true value.

Figure 3 illustrates the minimum value of K required to obtain PLEs with low bias. In terms of ϕ and logit(ϕ), the estimator requires at most 40 intervals when the behavior is not too infrequent (i.e., $\zeta \geq 0.1c$). Even when ζ drops to $0.05c$, low-bias estimates of $logit(\phi)$ can be obtained when $K \geq 60$. In terms of ζ and log(ζ), low-bias estimates can be obtained when K is at least 80 so long as the behavior has moderate prevalence ($0.30 \leq \phi \leq 0.70$) and incidence that is neither infrequent nor very frequent ($0.10 \leq \zeta \leq 0.45$). However, larger samples are required to obtain good incidence estimates when the behavior has more extreme prevalence and high incidence, or when the behavior has very low incidence.

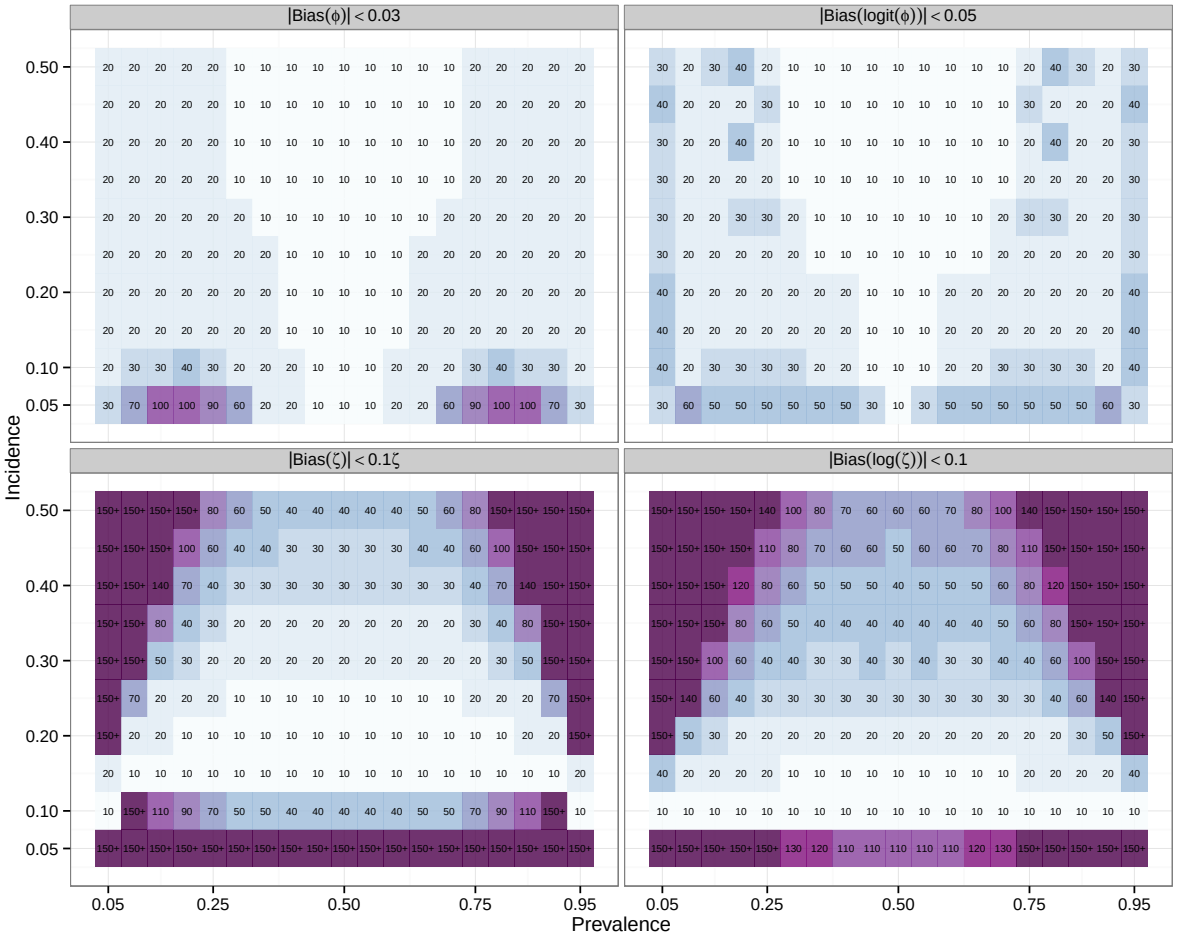

Figure 3. Minimum number of MTS intervals required to obtain PLEs with absolute bias less than given criteria.

Despite its limitations, there appears to be very little practical drawback to using the PLEs in place of the MTS proportion estimates. Unless the behavior is happening very infrequently, the PLE provides comparable estimates of prevalence. Even when the approximately unbiased estimates of incidence are difficult to obtain, the estimates of prevalence are nearly always still approximately unbiased. If an applied researcher is careful and aware of those instances where incidence estimates may not be trustworthy, we would argue that the PLE is superior to other known methods of obtaining MTS estimates.

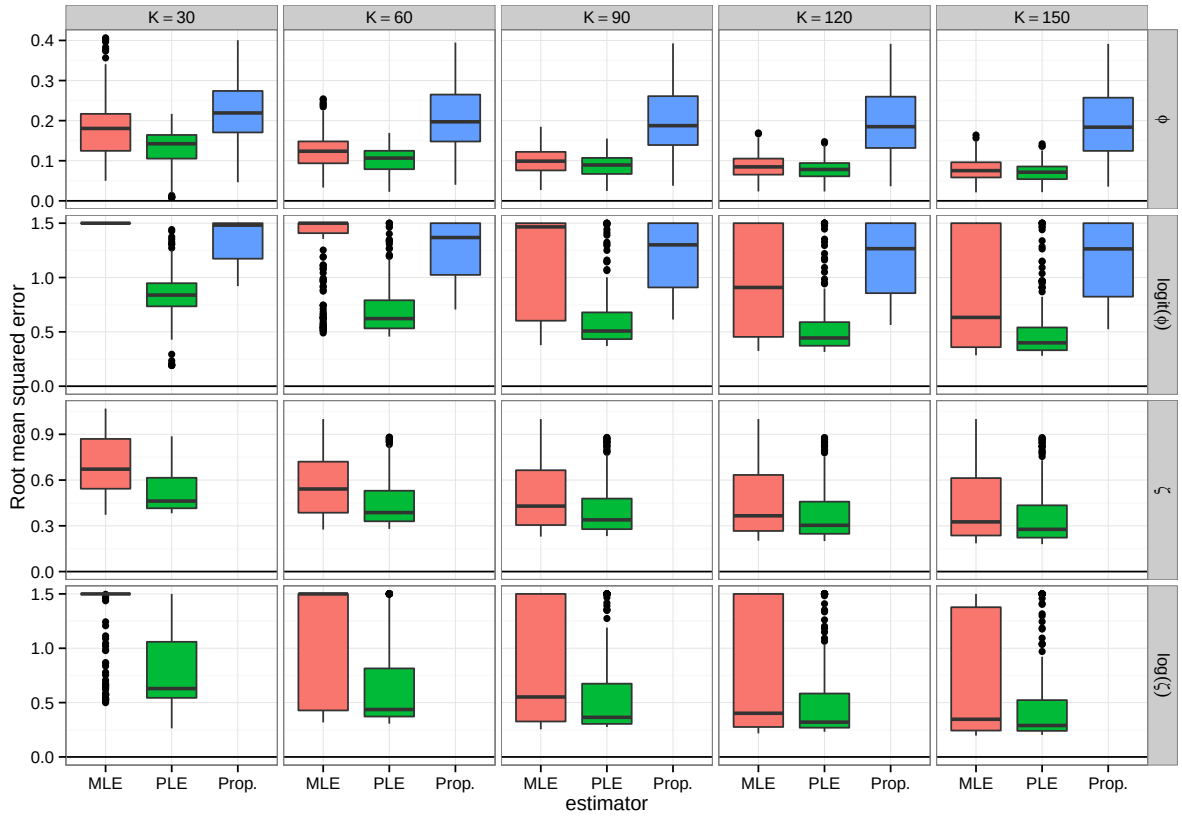

Figure 4. Distribution of root mean-squared error of MLE, PLE, and summary proportion (Prop.) estimates based on PIR data, across range of $0.05 \leq \phi \leq 0.95$ and $0.05 \leq \zeta \leq 0.50$. Values above 1.5 are not displayed. For zeta, the root mean-squared error is relative to the true value.

Results: PIR

Figure 4 illustrates the distribution of RMSE of the MLEs, PLEs, and summary proportions based on PIR data, across varying levels of K and for each parameterization. In contrast to MTS, the PIR summary proportion performs quite poorly as an estimate of prevalence. For prevalence, the PLE performs notably better than both the MLE and the summary proportions across the entire range of K, in terms of both the natural parameterization and the transformed parameterization. For incidence, the PLEs once again provide more accurate estimates than the MLEs, in both parameterizations. As

with PIR, the RMSE (and relative RMSE) of the PLEs is still relatively large.

Figure 5 illustrates the minimum value of K required for the PLEs based on PIR data to have low bias. Estimates of prevalence require a large number of intervals ($K \geq 70$) and estimates of the natural parameterization of incidence require a larger number of intervals ($K \geq$100) to obtain approximately unbiased estimates in even a restricted range. Approximately unbiased estimates can be found when prevalence is known to be relatively low ($0.15 \leq \phi \leq 0.30$) and incidence is moderate to low ($\zeta \leq 0.30c$), or incidence is low ($\zeta \leq 0.15c$) and prevalence is moderate to low ($.15 \leq \phi \leq 0.60$). Approximately unbiased estimates for the natural parameterization of ζ can still be obtained slightly outside these bounds, and approximately unbiased estimates for the natural parameterization of ϕ slightly farther still. Unbiased estimates for $log(\zeta)$ can only be found when the number of intervals is high($K \geq 100$), prevalence is moderate to low ($.10 \leq \phi \leq 0.70$) and incidence is very low ($\zeta \leq 0.10c$).

While these ranges are restrictive, returning to Figure 4 makes clear that our method is superior to the only other summary method in use. In general, PIR is really only appropriate for observing behaviors with moderate to low prevalence and moderate to low incidence. Our estimator works reasonably well in three of the four parameterizations, and no other method can offer even a biased estimate of log incidence. A researcher who only uses PIR when it is appropriate for the behavior of interest can make quite effective use of our PLEs as long as they are careful to restrict their investigation to the three parameterizations where approximately unbiased estimates are reasonably obtained.

Results: AIR

Figure 6 displays the RMSEs for the MLEs and PLEs for AIR data across a range of K for each parameterization. The MLEs appear to perform reasonably well at a moderate number of intervals ($K = 60$) for all parameterizations but $logit(\phi)$. The RMSEs at very large sample sizes for $logit(\phi)$ are much smaller, if still somewhat large.

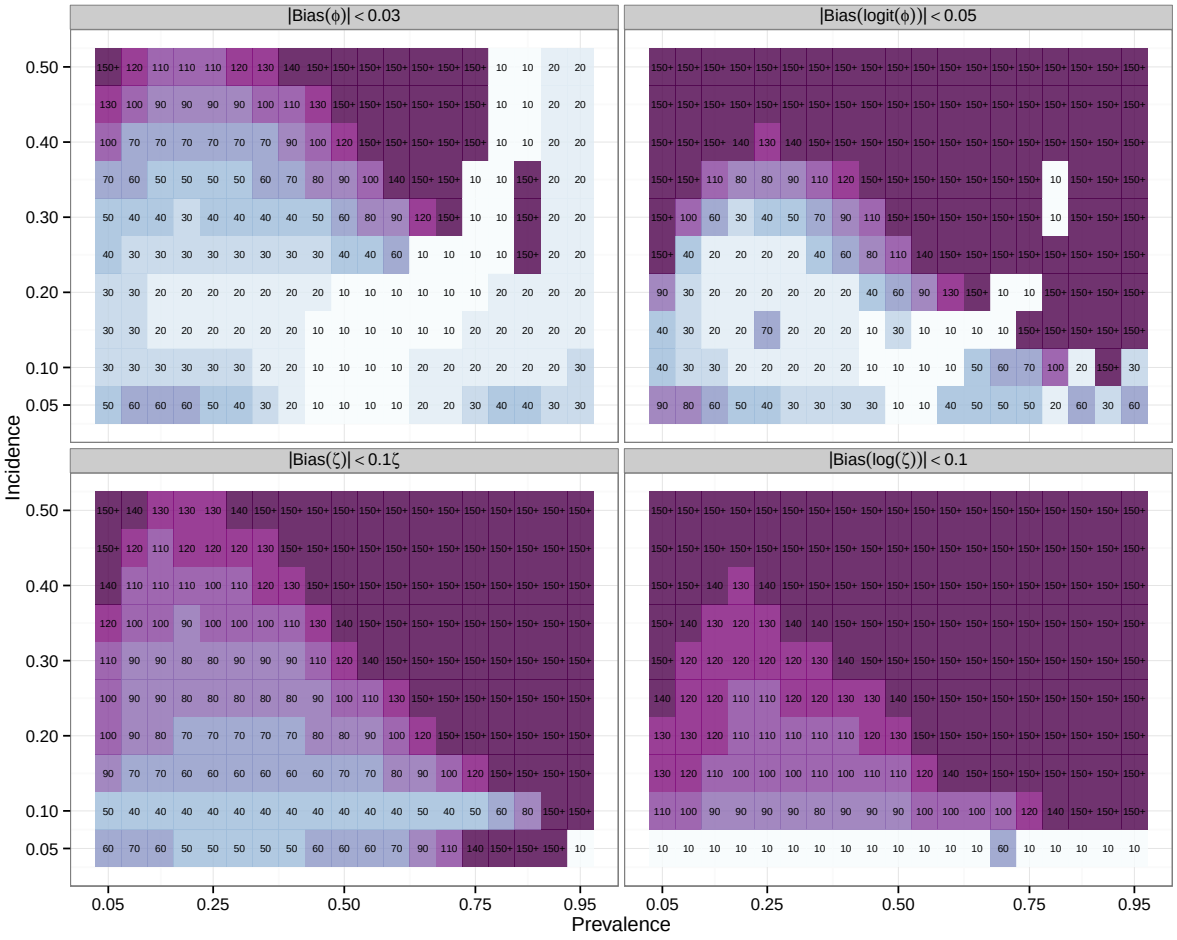

Figure 5. Minimum number of PIR intervals required to obtain PLEs with absolute bias less than given criteria.

For estimates of prevalence and incidence, the PLEs are clearly superior, except sometimes when the number of intervals is very large ($K \geq 120$).

Figure 7 displays the RMSEs for the PLEs for each of the recording methods and each of the parameterizations. For estimates of prevalence, both AIR and MTS are have superior accuracy to PIR. In general, the mean RMSE for MTS is slightly lower than AIR, but the range of the distributions are similar. For estimates of prevalence, AIR is superior to both MTS and PIR, and MTS is superior to PIR. Even at a large number of intervals ($K \geq 120$), the RMSE and relative RMSE can range to very near to or above 1 respectively for both MTS and PIR. In contrast, while the AIR RMSEs cannot be called

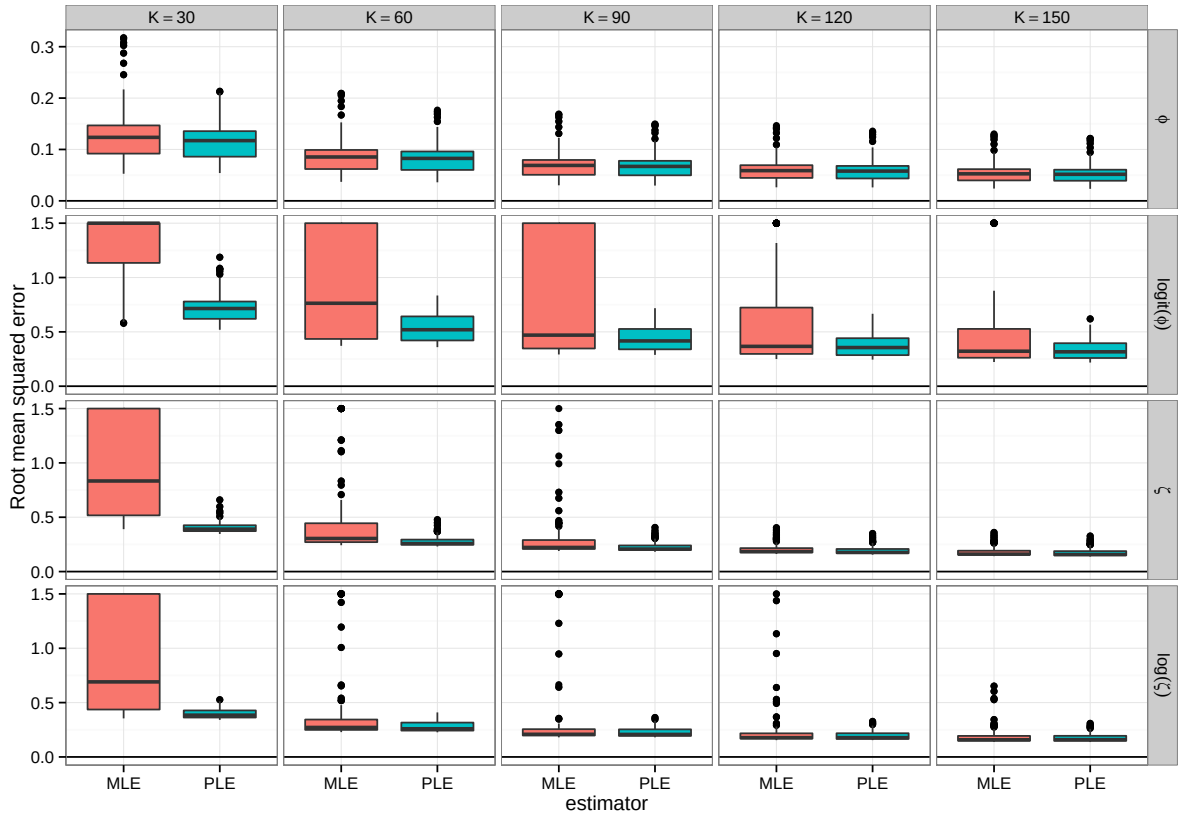

Figure 6. Distribution of root mean-squared error of MLEs and PLEs based on AIR data, across range of $0.05 \leq \phi \leq 0.95$ and $0.05 \leq \zeta \leq 0.50$. Values above 1.5 are not displayed. For ζ, the root mean-squared error is relative to the true value.

small, the range of the distributions is notably lower and also much narrower.

Figure 8 illustrates the minimum value of K required for the PLEs based on AIR data to have low bias. For both of the natural parameterizations, as long as the incidence is not extremely low ($\zeta \geq 0.10c$) approximately unbiased estimates can be obtained at a very reasonable number of intervals ($K \geq 40$). For $logit(\phi)$, as long as prevalence is neither very low nor very high, approximately unbiased estimates can be obtained across the entire range of incidence. Even those portions of the parameter space requiring a large number of intervals ($80 \leq K \leq 100$) are relatively small for all three parameterizations. For $log(\zeta)$, only a modest number of intervals ($K \leq 40$) is required to obtain approximately unbiased estimates across the entire range of the

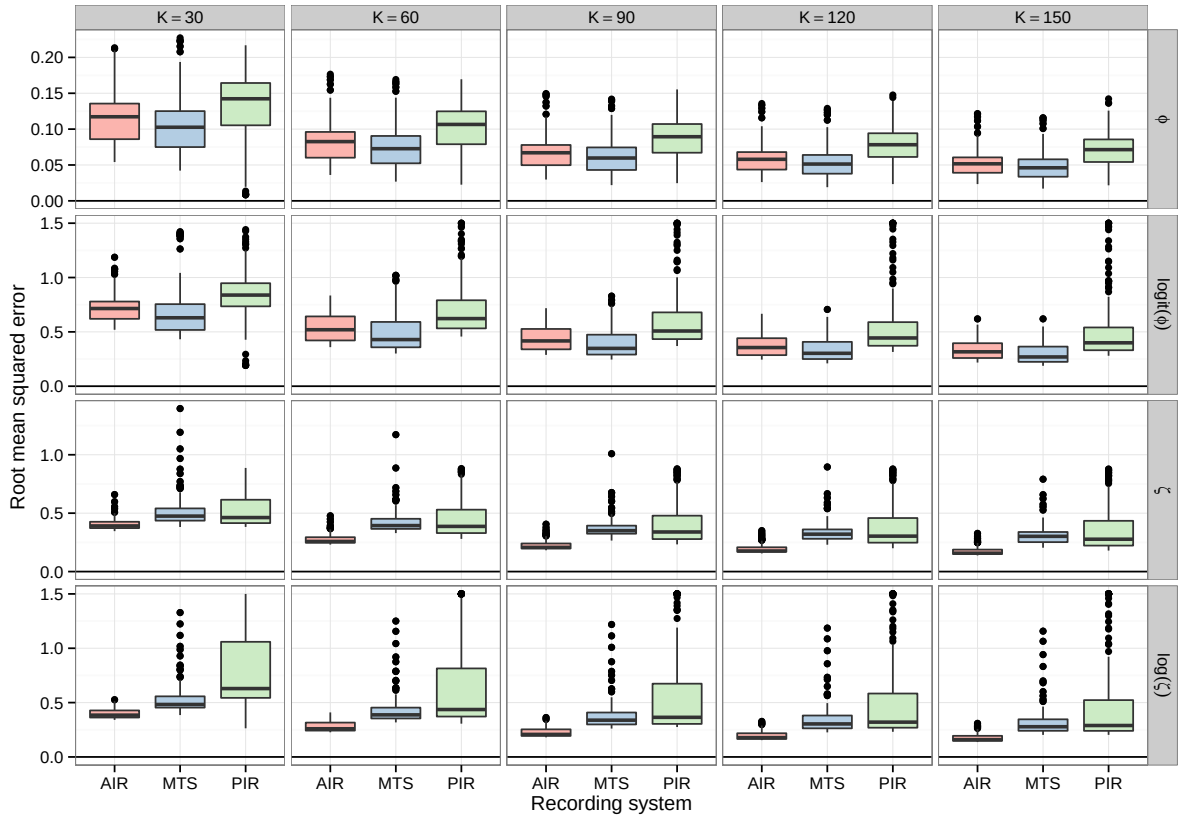

Figure 7. Distribution of root mean-squared error of PLE estimates based on AIR, MTS, and PIR data, across range of $0.05 \leq \phi \leq 0.95$ and $0.05 \leq \zeta \leq 0.50$. Values above 1.5 are not displayed. For zeta, the root mean-squared error for is relative to the true value.

parameter space.

In general, the AIR method performs very well at a modest number of intervals. If a researcher was interested in the natural parameterization of both prevalence and incidence and a behavior they expected to see no less frequently than about once every 7 intervals on average ($\zeta \geq 0.15c$), they would never need to use more than a very modest number of intervals ($K = 30$) to obtain unbiased estimates. The MTS PLEs do require a slightly smaller number of intervals to obtain unbiased estimates of prevalence in a given area of the parameter space. However, the estimates of incidence when using AIR require many fewer intervals and have such superior RMSEs that any researcher interested in estimates of incidence should stronger consider AIR. For a modest increase

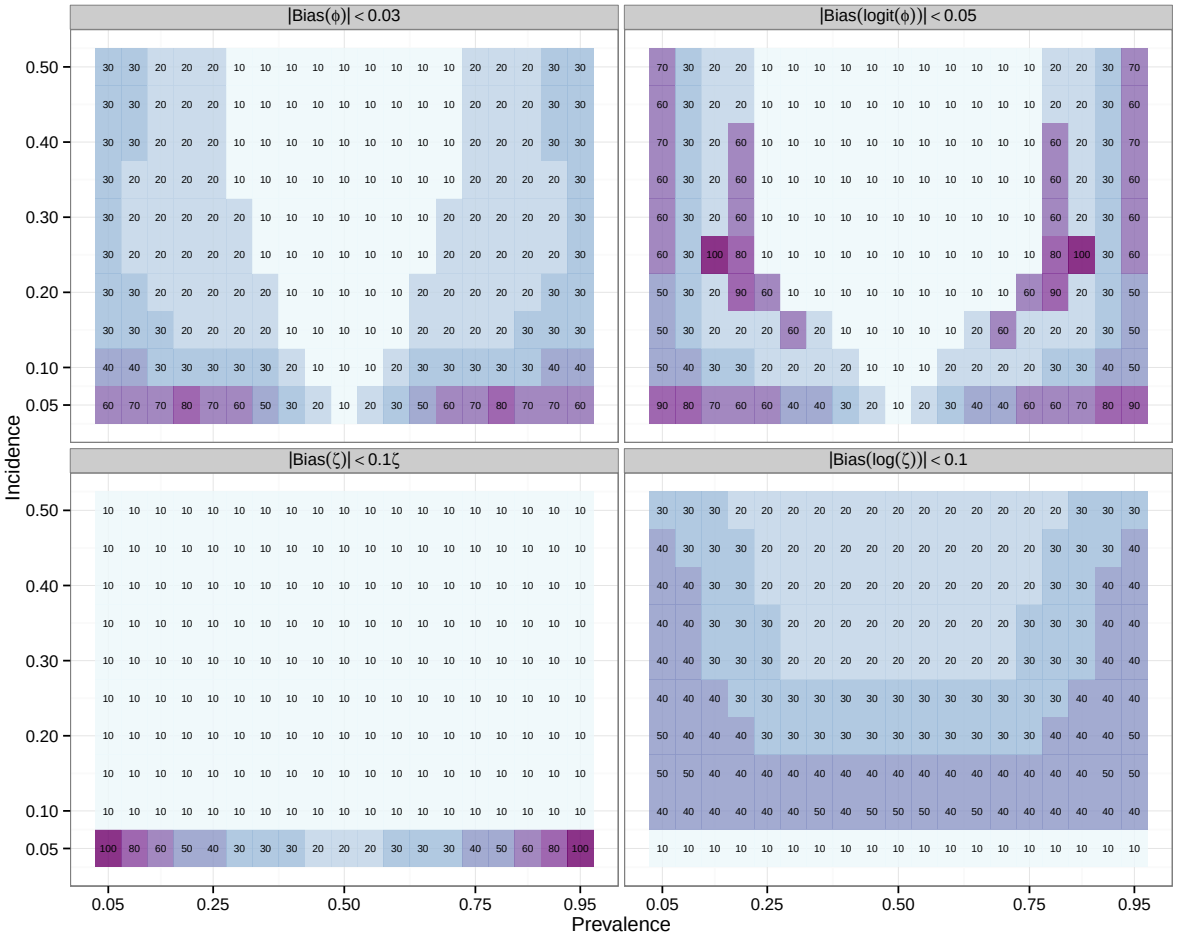

Figure 8. Minimum observation session length required to obtain PLEs based on AIR data with absolute bias less than given criteria.

in effort, a researcher can see serious games in terms of good estimates of both prevalence and incidence.

Application

This section demonstrates the use of the PLEs with empirical behavioral observation data. The data are drawn from **?**. In this study, trained raters watched six scripted, pre-recorded, ten-minute videos of a classroom during a lesson. The script outlined when students in the classroom were to act academically engaged or to act disruptively. For each video, academic engagement behavior was coded using three

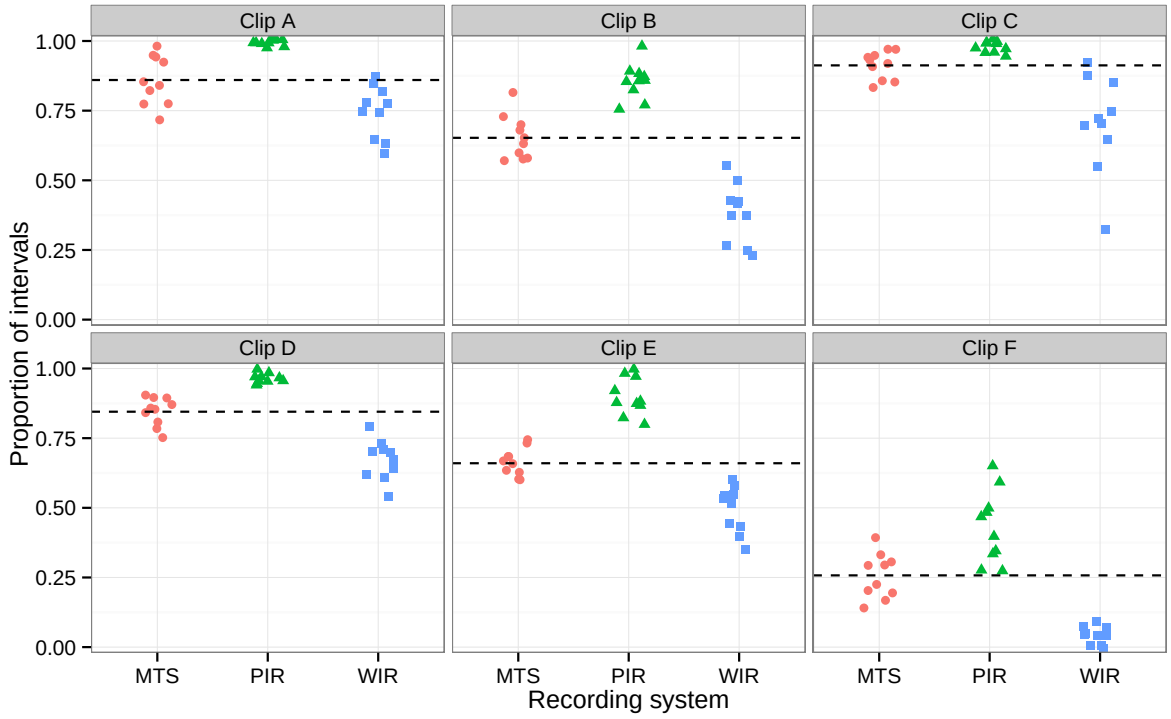

Figure 9. MTS, PIR, and WIR summary proportions for each rater, by clip. Horizontal dashed lines correspond to the average MTS summary proportions.

different recording systems: MTS, PIR, and WIR. All three systems used 15 s intervals, yielding a total of 40 intervals per scored clip. Each video clip was independently coded by ten different raters using each of the recording systems.

The goal of our analysis is to estimate the prevalence and incidence of the behavior observed in each video clip, using the data from each of the three recording systems. Because each video was scored separately using each system, comparing the PLEs obtained from each type of data allows us to characterize the relative strengths and weaknesses of the observation recording methods. The fact that each clip is scored by multiple raters also allows us to observe a source of measurement error that the Alternating Poisson Process model does capture: "inter-rater" error, due the human observers not perfectly following the scoring procedure or not perceiving the behavior stream with perfect accuracy.

Figure 9 plots the overall proportion of intervals with academic engagement for each clip and each recording procedure, with separate points for each rater. The horizontal dashed lines represent the average of the MTS proportions across all ten raters, which we treat as a benchmark for the prevalence estimates based on the Alternating Poisson Process. The PIR proportions are almost all higher than the unbiased MTS proportions, while the WIR proportions are nearly always lower, which is consistent with the fact that the PIR summary proportion is an upwardly biased estimate of prevalence while the WIR proportion is a downwardly biased estimate of prevalence. The prevalence of academic engagement is quite high–above .80–in clips A, C, and D. In these three clips, the PIR proportions are frequently at or very near to the ceiling level.

Figure 10 displays the PLEs for prevalence based on the MTS, PIR, and WIR scores for each rater and each clip. The prevalence estimates are plotted on the vertical axis, with the corresponding summary proportion of intervals on the horizontal axis. The quality of the estimates varies substantially by both recording procedure and by clip. For MTS, the PLEs are all close to the raw proportions (which are unbiased estimates of prevalence). The MTS confidence intervals cover the benchmark prevalence estimate in 58 out of the 60 MTS records and are generally narrower than the confidence intervals based on the PIR and WIR data. The PLEs based on PIR and WIR are in roughly in the same range as the estimates based on MTS data. Thus, on a gross level, the PLEs appear to correct the over- or under-estimation of prevalence in the summary proportions. The confidence intervals based on PIR data cover the benchmark prevalence estimate in 47 out of the 60 total records, with most of the discrepancies occurring in clip A, where many of the the PIR records are at ceiling. Similarly, the confidence intervals based on WIR data cover the benchmark prevalence estimate in 54 out of 60 records. However, the PIR and WIR estimates are generally considerably less precise than the MTS estimates, with much wider confidence intervals.

Provide numerical comparison of CI width.

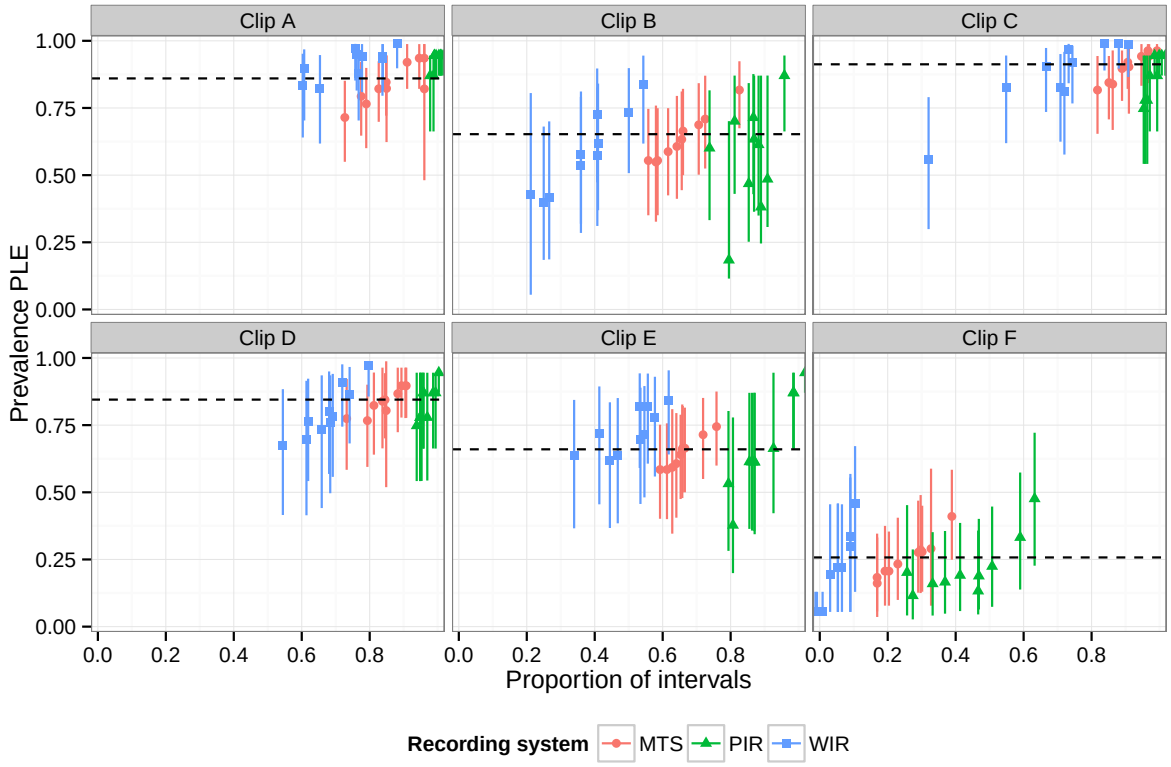

Figure 10. PLE estimates of prevalence versus proportion of intervals with behavior, for each recording system and each rater. Vertical whisker bars correspond to 95% confidence intervals for prevalence, obtained using a parametric bootstrapping procedure. Horizontal dashed lines correspond to the average MTS summary proportions.

Figure 11 displays the penalized likelihood estimates of incidence for each rater and each clip, again plotted against the raw proportion of intervals. Note that the vertical axis of this plot is on the log scale. In clips A through E, the estimates of the incidence of academic engagement are very high, suggesting that the student cycled rapidly between episodes of engagement and disruption. However, based on the simulation evidence presented in the previous section, we do not have great confidence in these estimates. For the moderate sample size of $K = 40$, none of the recording methods provide reasonable estimates of incidence for behaviors with high prevalence

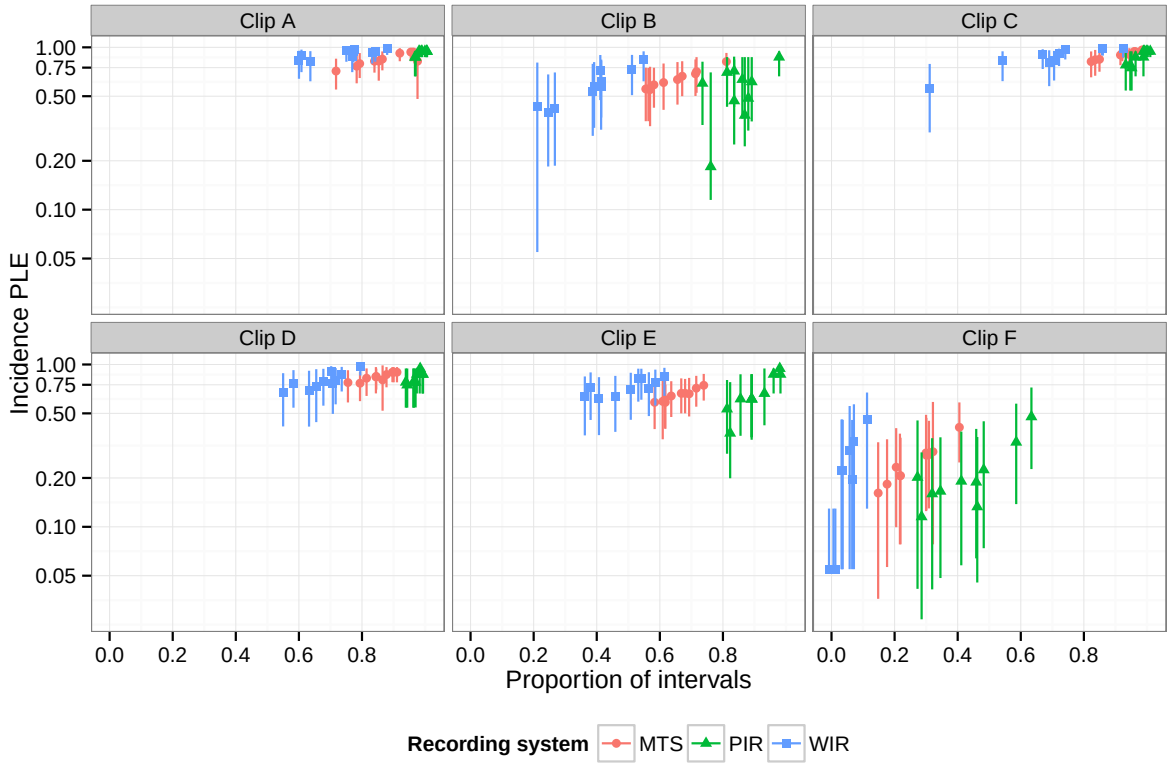

Figure 11. PLE estimates of incidence versus proportion of intervals with behavior, for each recording system and each rater. Vertical whisker bars correspond to 95% confidence intervals for incidence, obtained using a parametric bootstrapping procedure.

($\phi > 0.5$) and high incidence ($\zeta > 0.5$). In contrast, the estimates of incidence in clip F are more moderate, suggesting bouts of engagement occurring about every 4 to 7 intervals, probably sustained for a moderate period of time based on the prevalence estimates. While the PIR and WIR estimates are probably not trustworthy, many of the MTS estimates might be approximately unbiased. Unfortunately, we do not have any benchmark estimate of the true incidence of academic engagement in each of the clips. Lacking a credible point of comparison, we cannot directly evaluate the accuracy of the incidence estimates. We note only that the estimates from all three systems appear to be in rough agreement. However, we do know that the MTS PLEs have the closest agreement to the MTS proportion estimates, that the CIs for the MTS estimates are

generally tighter, and can compare figure 3 and figure 5 to see that MTS general requires fewer intervals for approximately unbiased estimates. If we had to provide a general recommendation of one of the three more "traditional" methods of interval recording, MTS recording using the PLEs would seem to be the obvious choice.

General discussion points?

I know this moderately contradicts your previous point about general agreement but it seems to me that the evidence has sort of stacked up in favor of MTS if we're ignoring AIR for the moment

Discussion

In this paper, we have considered how to estimate the prevalence and incidence of a state behavior from data collected using conventional intermittent observation recording systems, including momentary time sampling, partial interval recording, and whole interval recording. Following earlier work by Brown et al. (1977) on MTS, we used an Alternating Poisson Process to model the behavior stream as perceived by the observer, from which models for PIR, WIR, and AIR could be derived. For estimating the model parameters, simulation evidence indicated that penalized likelihood methods with generic, weak priors generally outperformed maximum likelihood methods–often dramatically so. For PIR data, penalized likelihood estimates of prevalence provided much more accurate estimates than the naivë summary proportion, which is currently widely used.

We have also described a novel recording procedure, augmented interval recording, that involves combining MTS, PIR, and WIR. For a given period of observation, and using intervals twice the length of other procedures, AIR provides estimates of prevalence that are only slightly less accurate than estimates based on MTS, while also providing estimates of incidence that are substantially more accurate than estimates based on any other procedure. Of course, at present these advantages are only theoretical. To determine whether AIR offers any advantage in practice, its feasibility in real-life research contexts will need to be assessed.

Across all of the recording procedures that we have considered, the foremost limitation of the models and estimation techniques that we have studied is the strength

of the assumptions entailed by the Alternating Poisson Process model for the behavior stream. The model posits that the individual episodes of behavior and spans of time in between episodes are exponentially distributed. Whether these distributional assumptions are reasonable–and for what classes of behavior–is an important question requiring further empirical research. Addressing it will likely require measuring the behavior of a large sample of participants using intensive, continuous recording techniques. Another related avenue of further research is to examine the extent to which the proposed estimation techniques are robust to violations of the distributional assumptions (e.g., assuming that event durations follow a gamma distribution that has lower variance than the exponential). It may well be that the robustness of the PLEs depends on which system is used to record the data, and whether prevalence or incidence is of primary interest.

Several other limitations of these models should also be acknowledged. Our approach has treated the recording procedures themselves as essentially mechanical procedures that can be applied without human error, yet in practice the procedures are not perfectly reliable. Indeed, the data from **?** displayed a surprisingly high level of inter-rater variability. The model we have considered could be seen as implicitly accomodating rater error by allowing that the Alternating Poisson Process describes the observer's *perception* of the behavior stream, rather than the true behavior stream. How to extend the model in order to more explicitly account for human error in the recording process remains an open question for further research.

Another limitation of the models is that they are limited to describing measurement error from a single observation session. In practice, systematic behavioral observation data is often collected on a single participant across many sessions (as in a single-case study) or across many participants (as in a between-subjects experiment). In either setting, it would be useful to embed the measurement model that we have proposed in a generalized linear modeling framework, which could be used to describe

changes in prevalence and incidence across time, or in a random-effects framework, which could be used to describe between-subjects variation in the characteristics of behavior streams.

Despite these limitations, the models and estimation methods that we have proposed remain useful. For MTS data, penalized likelihood provides a means to estimate incidence as well as to assess the extent of measurement error in the estimate of prevalence. For PIR data, the penalized likelihood estimates of prevalence represent an improvement over the current standard approach, which is to simply ignore the bias in the summary proportion.

Though not the focus of the present paper, the models that we have described can also be applied to develop better psychometric guidance for behavioral observation data. Through further mathematical analysis or though computer simulation, the models that we have presented could be used to study how choices regarding recording procedures, interval lengths, rest times, and observation session lengths influence the precision of behavioral measurements. Guidance regarding these aspects of study design would be useful to applied researchers designing single-case experiments or between-subjects trials.

Discuss Lane & Ledford (2014) and related work.

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

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.

Bakeman, R., & Quera, V. (2011). *Sequential Analysis and Observational Methods for the Behavioral Sciences*. New York, NY: Cambridge University Press.

Bornstein, M. H. (2002). Measurement variability in infant and maternal behavioral assessment. *Infant Behavior and Development*, *25*(4), 413–432. doi: 10.1016/S0163-6383(02)00143-1

Brown, M., Solomon, H., & Stephens, M. A. (1977). Estimation of Parameters of Zero-One Processes by Interval Sampling. *Operations Research*, *25*(3), 493–505.

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika*, *78*(4), 685–709. doi: 10.1007/s11336-013-9328-2

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

Galindo-Garre, F., Vermunt, J. K., & Bergsma, W. P. (2004, August). Bayesian Posterior Estimation of Logit Parameters with Small Samples. *Sociological Methods & Research*, *33*(1), 88–117. doi: 10.1177/0049124104265997

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of*

Applied Statistics, *2*(4), 1360–1383. doi: 10.1214/08-AOAS191

Griffin, B., & Adams, R. (1983). A parametric model for estimating prevalence, incidence, and mean bout duration from point sampling. *American Journal of Primatology*, *4*(3), 261–271. doi: 10.1002/ajp.1350040305

Haccou, P., & Meelis, E. (1992). *Statistical Analysis of Behavioural Data*. New York, NY: Oxford University Press.

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.

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.

Lane, J. D., & Ledford, J. R. (2014). Using interval-based systems to measure behavior in early childhood special education and early intervention. *Topics in Early Childhood Special Education*. doi: 10.1177/0271121414524063

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

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

Nelder, B. J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*(4), 308–313.

Primavera, L. H., Allison, D. B., & Alfonso, V. C. (1996). Measurement of dependent

variables. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 41–89). Mahwah, NJ: Lawrence Erlbaum.

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

Pustejovsky, J. E., & Swan, D. M. (2014). *Four methods for analyzing partial interval recording data, with application to single-case research.* Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

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.

Suen, H. K., & Ary, D. (1986, March). 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

Suen, H. K., & Ary, D. (1989). *Analyzing Quantitative Behavioral Observation Data.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

Warton, D. I. (2008). Penalized Normal Likelihood and Ridge Regularization of Correlation and Covariance Matrices. *Journal of the American Statistical*

Association, *103*(481), 340–349. doi: 10.1198/016214508000000021

Appendix A

Derivation of PIR model

The joint distribution of PIR observations depends on the conditional probabilities $\psi_k = \Pr[Z(t_k) = 1 \,|U_1, ..., U_{k-1}]$. This appendix provides a derivation of Expression (7) in terms of the parameters of the Alternating Poisson Process. The derivation will make use of the following lemma.

Lemma 1. *The conditional probabilities of* $Z(t_k) = 1, U_{k-1} = 1$ *given* $Z(t_{k-1})$ *are:*

$$\Pr(Z(t_k) = 1, U_{k-1} = 1 \,|Z(t_{k-1}) = 1) = p_1(c+d)$$
$$\Pr(Z(t_k) = 1, U_{k-1} = 1 \,|Z(t_{k-1}) = 0) = p_0(c+d) - p_0(d)\exp\left(\frac{-\zeta c}{1-\phi}\right).$$

Proof. Observe that

$$\Pr(Z(t_k) = 1, U_{k-1} = 1 \,|Z(t_{k-1}) = 1) = \Pr(Z(t_k) = 1 \,|Z(t_{k-1}) = 1) = p_1(c+d)$$

and

$$\begin{aligned}
&\Pr(Z(t_k) = 1, U_{k-1} = 1 \,|Z(t_{k-1}) = 0) \\
&\qquad = \int_0^c \frac{p_1(c-t)\zeta}{(1-\phi)} \exp\left(\frac{-\zeta t}{1-\phi}\right) dt \\
&\qquad = \phi\left[1 - \exp\left(\frac{-\zeta(c+d)}{\phi(1-\phi)}\right) - \exp\left(\frac{-\zeta c}{1-\phi}\right) + \exp\left(\frac{-\zeta(\phi c+d)}{\phi(1-\phi)}\right)\right] \\
&\qquad = p_0(c+d) - p_0(d)\exp\left(\frac{-\zeta c}{1-\phi}\right).
\end{aligned}$$

□

Turning to the derivation of ψ_k, begin by noting that $U_{k-1} = 0$ implies that $Z(t_k + c) = 0$. It follows from the Markov property that

$$\begin{aligned}
&\Pr(Z(t_k) = 1 \,|U_1 = u_1, ..., U_{k-2} = u_{k-2}, U_{k-1} = 0) \\
&\qquad\qquad = \Pr(Z(t_k) = 1 \,|Z(t_{k-1} + c) = 0) = p_0(d).
\end{aligned}$$

Next, Lemma 1 implies that

$$\begin{aligned}&\Pr\left(Z(t_k)=1, U_{k-1}=1\,|U_1,...,U_{k-2}\right)\\&\qquad=\psi_{k-1}p_1(c+d)+(1-\psi_{k-1})\left[p_0(c+d)-p_0(d)\exp\left(\frac{-\zeta c}{1-\phi}\right)\right].\end{aligned}$$

It therefore follows that

$$\begin{aligned}&\Pr\left(Z(t_k)=1\,|U_1=u_1,...,U_{k-2}=u_{k-2},U_{k-1}=1\right)\\&\qquad=\frac{\psi_{k-1}p_1(c+d)+(1-\psi_{k-1})\left[p_0(c+d)-p_0(d)\exp\left(\frac{-\zeta c}{1-\phi}\right)\right]}{1-(1-\psi_{k-1})\exp\left(\frac{-\zeta c}{1-\phi}\right)}.\end{aligned}$$

Thus, ψ_k can be written as a function of ψ_{k-1} and u_{k-1}, as given in (7).

Appendix B

Derivation of AIR model

This appendix provides a derivation of the transition probabilities for the AIR model in terms of the parameters of the Alternating Poisson Process. Begin by noting that, by the definitions of the recording procedures, $X_{k-1} = 0$ implies that $W_k = 0$ and $X_{k-1} = 1$ implies that $U_k = 1$. It follows that $\pi_{0|bc1} = 0$ for $b, c = 0, 1$ and $\pi_{1|b0d} = 0$ for $b, d = 0, 1$. Derivation of the other transition probabilities will make use of the following lemma. (The proof follow the same logic as in Lemma 1, and is therefore omitted.)

Lemma 2. *The conditional probability of* $Z(t_k) = 1, W_{k-1} = 0$ *given that* $Z(t_{k-1}) = 1$ *is*

$$\Pr\left(Z(t_k) = 1, W_{k-1} = 0 \,|Z(t_{k-1}) = 1\right) = p_1(c+d) - p_1(d)\exp\left(\frac{-\zeta c}{\phi}\right).$$

Turning to the eight remaining transition probabilities, note that

$$\begin{aligned}\pi_{0|100} &= \Pr\left(X_k = 1, U_k = 0 | X_{k-1} = 0\right) \\ &= \Pr\left(X_k = 1 | Z(t_k + c) = 0\right) \Pr\left(U_k = 0 | X_{k-1} = 0\right) \\ &= p_0(d)\exp\left(\frac{-\zeta c}{1-\phi}\right).\end{aligned}$$

Similarly,

$$\begin{aligned}\pi_{0|000} &= \Pr\left(X_k = 0, U_k = 0 | X_{k-1} = 0\right) = [1 - p_0(d)]\exp\left(\frac{-\zeta c}{1-\phi}\right) \\ \pi_{1|111} &= \Pr\left(X_k = 1, W_k = 1 | X_{k-1} = 1\right) = p_1(d)\exp\left(\frac{-\zeta c}{\phi}\right) \\ \pi_{1|011} &= \Pr\left(X_k = 0, W_k = 1 | X_{k-1} = 1\right) = [1 - p_1(d)]\exp\left(\frac{-\zeta c}{\phi}\right).\end{aligned}$$

Next, it follows from Lemmas 1 and 2 that

$$\begin{aligned}\pi_{0|110} &= \Pr\left(X_k = 1, U_k = 1 | X_{k-1} = 0\right) = p_0(c+d) - p_0(d)\exp\left(\frac{-\zeta c}{1-\phi}\right) \\ \pi_{1|110} &= \Pr\left(X_k = 1, W_k = 0 | X_{k-1} = 1\right) = p_1(c+d) - p_1(d)\exp\left(\frac{-\zeta c}{\phi}\right).\end{aligned}$$

The two remaining transition probabilities can be obtained by subtraction:

$$\pi_{0|010} = 1 - \pi_{0|000} - \pi_{0|100} - \pi_{0|110} = 1 - p_0(c+d) - [1 - p_0(d)] \exp\left(\frac{-\zeta c}{1-\phi}\right)$$
$$\pi_{1|010} = 1 - \pi_{1|011} - \pi_{1|110} - \pi_{1|111} = 1 - p_1(c+d) - [1 - p_1(d)] \exp\left(\frac{-\zeta c}{\phi}\right).$$