Alternating renewal process models for behavioral observation: Simulation methods, software, and validity illustrations

James E. Pustejovsky and Christopher Runyon

The University of Texas at Austin

Author Note

James E. Pustejovsky, Department of Educational Psychology, University of Texas at Austin. Christopher Runyon, Department of Educational Psychology, University of Texas at Austin.

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

Direct observation recording procedures produce reductive summary measurements of an underlying stream of behavior. Previous methodological studies of these recording procedures have employed simulation methods for generating random behavior streams, many of which amount to special cases of a statistical model known as the alternating renewal process. This paper describes the alternating renewal process model in its general form, demonstrates how it provides an organizing framework for most past simulation research on direct observation procedures, and introduces a freely available software package that implements the model. The software can be used to simulate behavior streams as well as data from many common recording procedures, including continuous recording, momentary time sampling, event counting, and interval recording procedures. Several examples illustrate how the software can be used to study the validity and reliability of direct observation data and to develop measurement strategies during the planning phases of empirical studies.

Keywords: behavioral observation; partial interval recording; alternating renewal process; simulation

Alternating renewal process models for behavioral observation: Simulation methods, software, and validity illustrations

Systematic direct observation is an important method for collecting measurements of human behavior in both between-subjects and within-subjects research contexts. Direct observation plays a particularly prominent role in single-case research, where researchers often use measures of overt behavior as dependent variables due to their scientific and social relevance (Ayres & Gast, 2010; Kazdin, 2011). Commonly used direct observation recording procedures include continuous recording, event counting, momentary time sampling, and interval recording (Ayres & Gast, 2010; Kahng, Ingvarsson, Quigg, Seckinger, & Teichman, 2011). Each of these direct observation recording procedures entails starting from a detailed stream of behavior as perceived by the observer (e.g., individual episodes of self-injurious behavior displayed by an autistic child during a therapy session) and reducing it into a simpler summary measurement (e.g., the proportion of session time that the child engages in self-injury).

Just as with other types of assessments, the validity of direct observation measurements hinges on their construct interpretation (Messick, 1988). Specifically, their validity depends on whether the direct observation recording procedure summarizes the desired characteristics of the underlying behavior stream in an interpretable way, capturing essential details while filtering out superfluous ones. The reliability of direct observation measurements depends on whether they quantify characteristics of the behavior stream in a consistent fashion, minimizing random variability in the measurement process.

The validity and reliability of direct observation measurements are crucial considerations across the stages of the research process (Martin & Bateson, 2007; Thompson, Symons, & Felce, 2000). For example, during the planning stages of a single-case study, researchers must weigh the validity and reliability of alternative direct observation recording systems and make choices between shorter or longer observation sessions. Sound interpretation of the results from a completed study rests on valid

measurement of the dependent variable, without which it is not possible to infer that an intervention actually affected the behavior of interest. Likewise, unreliable measurements make it more difficult to identify functional relationships (whether via visual inspection or statistical analysis); consequently, reliable measurement of the dependent variable is an important indicator of overall study quality in single-case research (Horner et al., 2005). Behavioral researchers must therefore have a strong understanding of the validity and reliability of direct observation procedures.

But how are researchers to build such an understanding? Clinical field experience is one route. Another is to draw on extant methodological research on direct observation procedures, which comes from two sources (Lane & Ledford, 2014). One source of evidence is empirical comparison studies, which apply several different recording procedures to a common set of real data and assess the extent to which the resulting measurements differ from known characteristics of the behavior stream (e.g., Gardenier, MacDonald, & Green, 2004; Murphy & Goodall, 1980; Powell, Martindale, & Kulp, 1975; Rapp et al., 2007). For example, Meany-Daboul, Roscoe, Bourret, and Ahearn (2007) used continuously recorded observations from an intervention study of several autistic children who exhibited vocal or motor stereotypy; after completing the study, they considered whether their conclusions would have differed if they had instead used partial interval recording or momentary time sampling. Empirical comparison studies have the advantage of closely emulating field practice, but have limited generality because usually only a small number of behavior streams are examined.

The other main source of evidence comes from simulation studies. Rather than working with real-life behavior stream data, simulation studies involve generating hypothetical behavior stream data using a theoretical model (e.g., Powell & Rockinson, 1978; Rapp, Colby-dirksen, Michalski, Carroll, & Lindenberg, 2008). For example, Harrop and Daniels (1986) used a purpose-written computer program to simulate behaviors of different durations; using the simulated behavior streams, they compared the sensitivity of

partial interval recording and momentary time sampling for detecting changes in behavior. Simulation studies have the advantage that sample sizes are essentially unlimited (constrained only by computing resources), so that a much wider variety of behavior streams can be examined. However, simulation studies involve theoretical models and usually abstract away from factors such as observer error.

The relevance of findings from and guidance based on methodological studies depends on the extent to which the behaviors (whether real or simulated) and procedures considered actually resemble those that researchers encounter in practice. A recent review by Lane and Ledford (2014) indicated that both empirical comparison studies and simulation studies may be lacking in this regard, because little past work has focused on the types of behaviors and recording procedures commonly encountered in the fields of early childhood special education and early intervention research. Thus, a closer examination of the assumptions on which extant research is needed.

Though it has not been widely recognized, much of the past simulation research on direct observation procedures is premised on a common set of modeling procedures. The basic approach in these studies is to simulate behavior streams by using random number generators to sequentially determine the length of each unique behavioral event and the lengths of time between behavioral events. Such procedures amount to specific cases of a general model known as the alternating renewal process (ARP). The ARP therefore serves as an organizing framework for most of the previous simulation research on direct observation recording procedures. However, past studies have typically adopted a very limited scope relative to the full range of possibilities in the ARP model, and little previous research has considered the flexibility of the model in its general form.[1]

This paper provides an introduction to the ARP model in its general form, demonstrates its connection with previous methodological research, and illustrates how the

[1]To our knowledge, the only exception is Rogosa and Ghandour (1991), who used mathematical and numerical analysis of the ARP to study the reliability of different direct observation recording procedures.

model can be used to further study the validity and reliability of direct observation procedures. By highlighting the model's underlying assumptions, we hope to invite greater scrutiny of their applicability to measurement practices in fields such as behavioral disorders, developmental disabilities, and early intervention research. The paper also demonstrates the use of a freely available software package called `ARPobservation`, written for use in the R statistical computing environment. The package provides a suite of tools for simulating behavior streams and direct observation recording procedures based on the ARP model, thus facilitating several types of useful methodological investigation. One use is to conduct further, systematic simulation research that is more directly motivated by and tailored to the characteristics of specific classes of behavior. Another use is develop and verify measurement strategies during the planning stages of an empirical study. By testing different recording systems or different observation session lengths in advance of applying them in the field, a researcher can ensure the validity and reliability of a study's measurement strategy.

The remainder of the paper is organized as follows. We first provide an overview of the ARP model and examine the extent to which past simulation studies fit into the common, general framework that it provides. We then give an overview of the `ARPobservation` package, explaining its design and functionality and providing several small illustrations of how the package can be used to study the validity and reliability of different direct observation recording procedures. Finally, the discussion section highlights the advantages and limitations of the ARP model.

The alternating renewal process model

The ARP is a statistical model that can be used to describe the characteristics of simple behavior streams, in which a behavior of interest is either occurring or not occurring at a given point in time. We will refer to the length of individual episodes of behavior as *event durations* and the lengths of time between episodes of behavior as *interim times*.[2] In

[2]The interim time is sometimes referred to as the inter-response time or inter-event time.

the ARP framework, variability is introduced into the behavior stream by treating each individual event duration and each interim time as a random quantity, drawn from some probability distribution. The characteristics of the behavior stream–and of direct observational measurements based thereon–are controlled by the mean and shape of the probability distributions from which event durations and interim times are drawn (Rogosa & Ghandour, 1991).

In the ARP model, a random behavior stream is constructed as follows. First, the initial interim time and the initial event duration are generated from certain probability distributions. There are several different ways that these initial values might be generated, and so we defer the details until the next subsection. Next, another interim time and another event duration are generated from specified probability distributions with means λ and μ, respectively. Then a third interim time and a third event duration are generated from the same probability distributions as just used. The process is repeated, with subsequent interim times and event durations generated in sequence to form a behavior stream. The stream is truncated when the sum of the interim times and event durations exceeds the length of the observation session. All interim times and all event durations are generated in a mutually independent manner, which means that the length of a given event is influenced neither by the length of previous events nor by how long it has been since the last event ended.

In its general form, the ARP model accommodates a wide variety of probability distributions for the event durations and interim times; all that is required is event duration and interim time distributions that describe non-negative random variables. For example, an exponential distribution with mean $\mu = 5$ s could be used for the event duration distribution and an exponential distribution with mean $\lambda = 35$ s could be used for the interim time distribution. Other common parametric families of distributions that describe continuous, non-negative random variables include the Weibull, gamma, log-normal, and continuous uniform distributions. Common parametric distributions for

non-negative, integer-valued random variables include the geometric, Poisson, negative binomial, and discrete uniform distributions (Leemis & McQueston, 2008). The ARP model applies even if events have a fixed duration (i.e., each event lasts 4 s) so long as the distribution of interim times is random. The flexibility of the ARP means that it can be used to model behaviors with a wide range of characteristics. The challenge then becomes selecting distributions that well describe the types of behaviors one is trying to model, so that the simulated behavior streams provide a reasonable facsimile for those encountered in practice. We comment further on this challenge in the discussion section.

Initial conditions

The behavior of the ARP depends to some extent on how the initial interim time and the initial event duration are generated, or what we will call the *initial conditions*. From a procedural standpoint, the simplest initial conditions involve generating the initial interim time from the same distribution as the later interim times and generating the initial event duration from the same distribution as the later event durations. This will result in behavior streams that always begin with an interim time; as a result, the probability that an event is occurring right at the start of the session (or soon after) will be zero (or near zero). An alternative, more complex set of initial conditions involves generating the initial interim time and the initial event duration from certain, special probability distributions that create a constant probability that a behavioral event is occurring at any point in time during the observation session.[3] Following the latter approach, the ARP that generates the behavior stream is said to be *in equilibrium*.

Using equilibrium initial conditions has the advantage of simplifying the behavior of the ARP model, because certain characteristics of the behavior stream become less dependent on the length of the observation session and the parametric forms of the event duration and interim time distributions (Rogosa & Ghandour, 1991). However, the equilibrium initial conditions may be a less realistic assumption in certain contexts, such as

[3]Kulkarni (2010, Chp. 8) provides further technical details.

when the start of the observation session coincides with a transition between class periods in a school setting. Given that both advantages and disadvantages exist, the simulation software described in a later section allows the user to control whether the equilibrium initial conditions are used when generating random behavior streams.

Simulating data from direct observation procedures

After generating a behavior stream based on the ARP model, different types of direct observation recording procedures can be applied in order to generate summary measurements. Because the behavior streams display random variation, so too do the summary measurements. The process of applying a given direct observation procedure to a behavior stream can be modeled by a mathematical algorithm, which takes as input a simulated behavior stream and produces as output a single summary measurement. In what follows, we briefly describe the summary measurements produced by the main continuous and discontinuous observation procedures.[4]

Continuous observation procedures include event counting and continuous recording. The summary measurement from event counting is calculated as the number of behavioral events that begin during the observation session; we will denote such a measurement as Y^E. The summary measurement from continuous recording, which we will denote as Y^C, is calculated as the proportion of time that the behavior occurs during the observation session.

Discontinuous observation procedures all involve dividing an observation session into a number of short intervals and scoring each interval as a zero or a one according to some rule. In momentary time sampling, an interval is scored as a one if a behavioral event is occurring at the very last instant of the interval. In partial interval recording, an interval is scored as a one if the behavior occurs at any point during the interval. In whole interval

[4]Precise mathematical descriptions of the algorithms implemented in `ARPobservation` can be found in package documentation, which can be accessed by typing `vignette("Observation-algorithms")` at the R command line after installing the `ARPobservation` package.

recording, an interval is scored as a one only if the behavior occurs for the entire duration of the interval. For all three procedures, a summary measurement is calculated as the proportion of intervals receiving a score of one. We will denote denote momentary time sampling measurements as Y^M, partial interval recording measurements as Y^P, and whole interval recording measurements as Y^W.

Two possible targets of measurement

Under the ARP model for the behavior stream, there are multiple ways of conceptualizing the measurand, or target of measurement, corresponding to a given observation procedure. One conception takes as the measurand some characteristic of an observed behavior stream, such as the percentage duration of the behavior or the frequency of the behavior over the course of an observation session. These quantities are equivalent to the measurements produced by continuous recording (Y^C) and event counting (Y^E), respectively. Under this conception, discontinuous observation procedures yield valid and reliable measurements to the extent that they accurately represent the measurements that would be produced by applying continuous observation procedures to the same observed behavior stream. We will call this the *observed behavior* conception.

An alternative definition of measurands is based on the parameters of the ARP model for the behavior stream. In this *behavioral parameter* conception, the behavior observed during any given session is treated not as the target of measurement, but rather as only a sample from the ARP data-generating model. With this conception, the main behavioral characteristics of interest are prevalence, or the long-term proportion of time spent in the behavior, and incidence, or the long-term rate at which new behavioral events occur. Under the ARP model, these quantities are directly related to the mean event duration (μ) and mean interim time (λ); specifically, prevalence is equal to $\mu/(\mu + \lambda)$ and incidence is equal to $1/(\mu + \lambda)$.

A key distinction between these two approaches is what they imply about continuous recording and event counting measurements. The observed behavior approach takes Y^C

and Y^E to be the measurands, assuming implicitly that these continuously measured quantities do not themselves contain any measurement error. In contrast, the behavioral parameter approach allows for the possibility that even continuous recording and event counting measurements may contain measurement errors because they are based on a sample of behavior over a finite amount of time. Thus, measurements based on longer observation sessions will be more reliable than measurements based on shorter sessions. These two conceptualizations of behavioral measurands lead to different approaches to simulating behavior streams, as we show in the next section.

Simulation studies of behavioral observation data

Many studies of the validity and reliability of direct observation procedures have employed simulation methods. Given that generalizations from simulation studies are limited by the models that they employ, it is important to understand the range of modeling approaches considered in previous research. We therefore conducted a systematic review to examine (a) the extent to which past studies used data generation methods that fit within the framework of the the ARP model; (b) what other data generation procedures have been studied; and (c) for both ARP and non-ARP approaches, the range of behavioral characteristics (i.e., prevalence and incidence) examined. To be included in our review, a study had to meet two criteria: it had to focus on procedures for direct observation recording of behavior and it had to use stochastic simulation methods.

We searched the PsycNET and Web of Science research databases using the term "simulation" in combination with any of the following terms: "momentary," "momentary time," "partial-interval," "one-zero," "zero-one," "modified frequency," "Hansen frequencies," "continuous recording," "duration recording," "event counting," "frequency recording," and "tally method." The initial searches returned 1051 results. Removing duplicates and screening studies based on titles and abstracts yielded a smaller set of 62 studies for full-text review. We then conducted forward and backward citation searches of these articles to identify further studies that the initial searches may have missed; this

resulted in 7 additional studies for full-text review. After a detailed review of these 69 studies, we identified 20 studies (in 19 articles) that met our inclusion criteria.[5]

We classified the studies that met our inclusion criteria according to whether the data-generating procedures fit within the ARP framework.[6] A study was classified as fitting the ARP framework if it generated behavior streams by sequentially simulating event durations and interim times from known probability distributions. For the studies that fit in the ARP framework, we extracted information on the form and range of mean values for the event duration and interim time distributions. For the studies that did not fit into the ARP framework, our main goal was to understand the nature of the procedures used to simulate behavior streams. We therefore extracted information about the general form of the procedures employed, as well as specific characteristics that varied across studies, including whether event durations were allowed to overlap or to occur consecutively. For both categories of studies, we extracted the ranges over which model parameters were varied, the length of the simulated observation sessions, and the number of times that each simulated condition was replicated.

Studies using an ARP model

Our literature search identified 14 simulation studies that fit within the ARP framework. Table 1 summarizes the simulation design from each of these studies. Most of the studies used the same family of probability distributions for both the event duration and interim time distributions; the only exceptions were Harrop and Daniels (1985, 1986), who used fixed event durations and geometrically distributed interim times. The most commonly employed form of probability distribution was a discrete uniform distribution or some variant thereon. For example, Rapp et al. (2008) simulated from a sum of discrete

[5]The most common reason for excluding a study was that it used empirical data (rather than simulated data) to study direct observation procedures.

[6]We were unable to classify one study (Green & Alverson, 1978) because the article did not provide sufficient procedural detail. One other study (Wilson, Jansen, & Krausman, 2008) simulated momentary time sampling data, but did so without simulating behavior streams.

Table 1

Simulation studies using an ARP model

Study	Distributional form	Mean event duration (s)	Mean interim time (s)	Session length (s)	Simulation conditions	Replications
Repp et al. (1976)	Fixed event durations, Geometric interim times	0.035	6-600	10800	6	3
Powell and Rockinson (1978)	Discrete uniform	2-18	4-54	1800	11	1
Tyler (1979)	Continuous uniform, reciprocal, or inverse uniform	40-150	not specified	3600	3	50-80
Ary and Suen (1983)	Discrete uniform	12.5-120	45-1155	1800	9	100
Griffin and Adams (1983)	Exponential	36-108	36-396	10800	36	3000
Powell (1984a)	Discrete uniform	2	2-62	1800	5	1
Powell (1984b)	Discrete uniform	2-500	2-162	1800	17	1
Harrop and Daniels (1985)	Fixed event durations, Geometric interim times	1-20	6-60	3600	6	20
Harrop and Daniels (1986)	Fixed event durations, Geometric interim times	1-20	5-180	3600	44	20
Quera (1990)	Weibull	11	15	500	1	500
Engel (1996)	Semi-Markov graph	10-250	not specified	1700	1	20
Rapp et al. (2008, Study 1)	Sum of discrete uniforms	2-18	2-18	600	18	1
Rapp et al. (2008, Study 2)	Sum of discrete uniforms	1	2-79	600	9	1
Devine et al. (2011)	Sum of discrete uniforms	3.5-27.5	3.5-27.5	600-3600	18	6

uniform distributions by rolling several dice and using the sum of the pips as the length of an event duration. Other studies used pseudo-random number algorithms to sample from geometric, Weibull, or exponential distributions.

While the set of studies as a whole employed a diverse set of probability distributions, only Tyler (1979) examined multiple families of distributions within a single article. As a result, most studies were unable to examine whether and how the form of probability distribution might have influenced their findings. Furthermore, only a few studies provided any theoretical or empirical justification for the use of a particular probability distribution (the exceptions being Engel, 1996; Powell, 1984a; Quera, 1990; Tyler, 1979). Consequently, it is difficult to determine the types of real behaviors for which the assumptions employed in these studies provide a good model.

Studies using a non-ARP model

Our literature search identified five studies that simulated behavior streams using procedures that did not fit into the ARP framework. Table 2 summarizes the features of these simulation studies. All five studies used similar procedures for simulating behavior streams. The common approach was to first create a blank array representing an observation session, with one slot per second of observation time. A uniform random number generator was then used to identify the point of onset for a behavioral event. Beginning with that randomly determined location, consecutive slots in the array were filled in to indicate the presence of an event. The process of placing behavioral events at randomly determined locations within the array was repeated until a certain dimensional quantity of the behavior stream (either the sample frequency or the percent duration) attained a specified level. We are not aware of any recognized term for this simulation procedure; we will refer to it as the *random onset* model.

The studies differed in certain details of how the random onset model was implemented. Some studies used fixed lengths for event durations and calculated the number of events necessary to achieve a specific percentage duration (Wirth et al., 2014),

Table 2

Simulation studies using a non-ARP model

Study	Event duration distribution	Mean event duration (s)	Percent duration (%)	Session length (s)	Simulation conditions	Replications
Rhine and Ender (1983)	Random	1-120	.06-10	14400	42	20
Rojahn and Kanoy (1985)	Constant or Random	1-8.5	.33-18.89	900	24	5
Kearns et al. (1990)	Constant	10	20-80	3600	6	1
Edwards et al. (1991)	Constant	10	20-80	3600	4	1
Wirth et al. (2014)	Constant	1-256	1-100	3600	2400	100

whereas other studies used event durations selected randomly from within a specified range and calculated the number of events necessary to achieve a specific frequency (Rhine & Ender, 1983).[7] While all of the studies allowed events to occur consecutively (leaving zero interim time in between two events), they differed in whether and how they allowed event durations to overlap. Overlapping events can occur in the random onset model if the randomly selected time of onset corresponds to a location in the array in the middle of or just prior to a previously assigned event. The precise method for handling overlapping events was generally difficult to ascertain; some studies may have allowed certain configurations of overlapping events (Wirth et al., 2014), whereas others may have prevented this possibility (Rhine & Ender, 1983; Rojahn & Kanoy, 1985).

There are three key distinctions between the random onset model and the ARP model. First and foremost, the random onset model is designed to exactly control dimensional quantities of the observed behavior stream, by simulating behavior streams that all exhibit a specified frequency of behaviors or percentage duration. As a result, the random onset model implicitly adopts an observed behavior conception of measurands. Second, the random onset model fills in individual event durations anachronically: it is possible for an event near the end of the session to be filled in prior to an event earlier in the observation session. If events are not allowed to overlap, then the probability of an event occurring at the present moment in time might depend on the occurrence of future events. In contrast, the ARP model describes event durations and interim times that occur sequentially in time, similar to what an observer would actually perceive. Third, the random onset model does not simulate interim times directly (as is done in the ARP model); instead, interim times are represented by those elements of the array that have not been filled after all of the events had been modeled.

[7] A variant of this latter method was also used to simulate conditions where events were restricted to occur in clusters near one another (Rojahn & Kanoy, 1985).

Simulation design considerations

The studies varied in the range of behavioral characteristics used. Some studies based on the ARP model examined large ranges for the mean event duration and mean interim time, aiming to be comprehensive (e.g., Harrop & Daniels, 1986), while other studies looked only at limited ranges (e.g., Quera, 1990). Similarly, studies based on the random onset model varied in the range of event durations and percentage durations studied, from limited (e.g., Edwards et al., 1991) to comprehensive (e.g., Wirth et al., 2014). However, few studies explicitly justified the range of behavioral characteristics used in the simulation in terms of the specific type of behavior to be modeled (one exception was Tyler, 1979).

Across both ARP and random onset studies, there was large variation in the scope of the simulations, with individual studies examining as few as 1 to more than 44 conditions; Wirth et al. (2014) went even further, modeling 2400 distinct conditions. The simulation designs typically included several combinations of parameter values, such as different ranges of event durations or different percentage durations. Many studies also varied the interval lengths for discontinuous recording methods. Some studies did not use fully factorial designs, choosing instead to examine only a subset of all possible combinations of the different factors, based on either theoretical considerations or empirical limitations. Simulated behavior streams ranged in length from 500 s to 14400 s, with the most common used lengths of 600s, 1800s, and 3600s.

A common limitation was the use of only a small number of replications, leaving open the possibility that findings are due in part to random chance. Across the ARP and random onset studies, the majority used 20 or fewer replications per condition. This may be due in part to the complexity of simulating behavior streams, or to the amount of time required with physical randomization devices (Devine et al., 2011; Rapp et al., 2008, e.g.). To address this limitation of previous research, we now describe a software package that can rapidly simulate behavior streams based on the ARP model, making it possible to efficiently conduct larger simulation studies with a greater number of replications per

condition.

The ARPobservation package

We have created a software package called `ARPobservation` that provides a set of tools for simulating direct observation data, based on a flexible ARP model. The software is written for the R statistical computing environment (R Core Team, 2014), which is freely available.[8] The package is available on the Comprehensive R Archive Network and can be installed and loaded from the R command line by typing

```
install.packages("ARPobservation")
library(ARPobservation)
help(package="ARPobservation")
```

The first line needs to be run only once; the second must be run each time the package is used; the third line is optional, and can be used to view the package documentation.

The package can be used to simulate many different direct observation recording procedures. It works by first simulating behavior streams based on an ARP, using specified distributions of event durations and interim times. Different procedures for recording data can then be applied to the simulated behavior streams. In the remainder of this section, we provide an overview of the package's design and functionality.

Simulating behavior streams

The first step in simulating direct observation data is to simulate full behavior streams. This is accomplished using the function `r_behavior_stream`, which takes several arguments.[9] The user must input (a) the number of behavior streams to generate, (b) the average event duration, (c) the average interim time, (d) the forms of the parametric

[8] For a basic introduction to the language and logic of computing in R, see Teetor (2011), among others.

[9] To access the documentation for this function, type `?r_behavior_stream`. Documentation for each of the other functions described in this section can be accessed by typing `?` followed by the function name. Many of the functions have further, optional arguments beyond those described in this article.

distributions to use for generating random event durations and random interim times, (e) the total length of the behavior stream, and (f) whether to use equilibrium initial conditions. The following code generates a simulated behavior stream of length 600 s from an equilibrium ARP, where the event durations follow an exponential distribution with mean $\mu = 10$ s and interim times follow an exponential distribution with mean $\lambda = 30$ s:

```
r_behavior_stream(n = 1, mu = 10, lambda = 30, F_event = F_exp(),
      F_interim = F_exp(), stream_length = 600, equilibrium = TRUE)
```

The parametric form of the event duration distribution can be changed by specifying a different function in the `F_event` argument; similarly, the interim time distribution can be changed using the `F_interim` argument. For example

```
r_behavior_stream(n = 1, mu = 10, lambda = 30, F_event = F_const(),
      F_interim = F_gam(4), stream_length = 120, equilibrium = TRUE)
```

generates a behavior stream of length 120 s in which event durations are constant, each having a length of exactly 10 s, and interim times are generated from a gamma distribution with shape parameter 4. As of this writing, the package includes functions for exponential distributions, gamma distributions, mixtures of two gamma distributions, Weibull distributions, uniform distributions, and constant values;[10] more probability distributions may be added in the future. Finally, behavior streams with non-equilibrium initial conditions can be simulated by setting the option `equilibrium = FALSE`.

The output of the `r_behavior_stream` function is a list of simulated behavior streams. Each behavior stream is stored as two components, an initial state (called `start_state`) and a list of transition times between states (called `b_stream`). For instance, the previous example produces the following output:[11]

[10] Type `?eq_dist` for further details on these distributions.

[11] Here and following, the output of R commands will be indicated with two pound signs (`##`).

```
## $start_state
## [1] 0
##
## $b_stream
## [1]   9.285  19.285  43.013  53.013  69.750  79.750 110.295
```

This simulated behavior stream begins in `start_state = 0`, meaning that the behavior is not occurring at the beginning of the session. A new behavioral event starts at time 9.285 and lasts until time 19.285; another behavioral event begins at time 43.013 and lasts until time 53.013. As expected, each behavior lasts exactly 10 seconds, while the time in between events is random. Also, note that a new behavior begins at time 110.295 but lasts beyond the end of the observation session, so that the time at which it ends is not recorded.

Simulating direct observation procedures

In most cases, one will have little need to inspect the output of the `r_behavior_stream` function. Instead, the user will store the simulated behavior streams generated by the function, and will then apply a direct observation procedure to these stored results. The package provides several functions for turning simulated behavior streams into summary measurements. Each function takes a list of simulated behavior streams and applies a specific recording procedure. For example, to generate four continuous recording observations, one first simulates the behavior streams, then applies the `continuous_duration_recording` function:

```
BS = r_behavior_stream(n = 4, mu = 10, lambda = 30,
        F_event = F_exp(), F_interim = F_exp(), stream_length = 600)
continuous_duration_recording(BS)
```

The first line stores the output of the `r_behavior_stream` function in a new object called `BS`; the second line applies continuous recording to these simulated behavior streams. The

result is four observations, each of which represents the continuous recording summary measurement from one behavior stream:

```
## [1] 0.2374 0.3092 0.1592 0.3075
```

Each value represents the proportion of the observation session during which the behavior occurred. One could instead apply the `event_counting` function to the same set of simulated behavior streams:

```
event_counting(BS)
## [1] 14 20 11 15
```

The result is four observations, each of which represents the number of behavioral events that began over the course of one observation session.

The functions for continuous recording and event counting take no arguments other than the list of simulated behavior streams, but the functions for momentary time sampling and interval recording require the user to specify additional procedural details. To use the `momentary_time_recording` function, the user must specify `interval_length`, the time length of each interval. Consider the behavior streams stored in `BS`, all of which have length 600 s. Specifying `interval_length = 15` divides the session into 40 intervals, each of length 15 s:

```
momentary_time_recording(BS, interval_length = 15)
## [1] 0.275 0.375 0.175 0.300
```

The result is again four observations, each of which represents the proportion of instants (out of 40 possible) during which the behavior was observed. The `interval_recording` function works along very similar lines, but also includes an additional option for whether to use partial or whole interval recording. The following example applies 15 s partial interval recording to the `BS`:

```
interval_recording(BS, interval_length = 15, partial = TRUE)
## [1] 0.500 0.725 0.400 0.600
```

If the user instead sets `partial = FALSE`, the function will produce whole interval recording data.

Finally, the package includes a convenience function called `reported_observations`, which applies multiple recording procedures to a list of behavior streams. The user must specify the number of intervals to use for momentary time sampling and the interval recording methods. The following example repeats all of the earlier calculations on `BS`, and also adds the results of whole interval recording:

```
reported_observations(BS, interval_length = 15)
##        C     M  E     P     W
## 1 0.2374 0.275 14 0.500 0.025
## 2 0.3092 0.375 20 0.725 0.075
## 3 0.1592 0.175 11 0.400 0.000
## 4 0.3075 0.300 15 0.600 0.100
```

The result is a matrix with one row per behavior stream and one labelled column per recording procedure. The results in each column are identical to the results of applying the corresponding recording procedure to the same list of behavior streams.

The package is fast enough to simulate thousands of observation sessions in a matter of seconds. For example, generating 10000 behavior streams and applying all five recording procedures (using 40 intervals per session for momentary time sampling and interval recording) takes approximately 1.1 s on a desktop computer with an Intel Core i7-4770 CPU and 8 GB of RAM. We take advantage of this efficiency in the examples presented in the next section.

Validity and reliability analysis

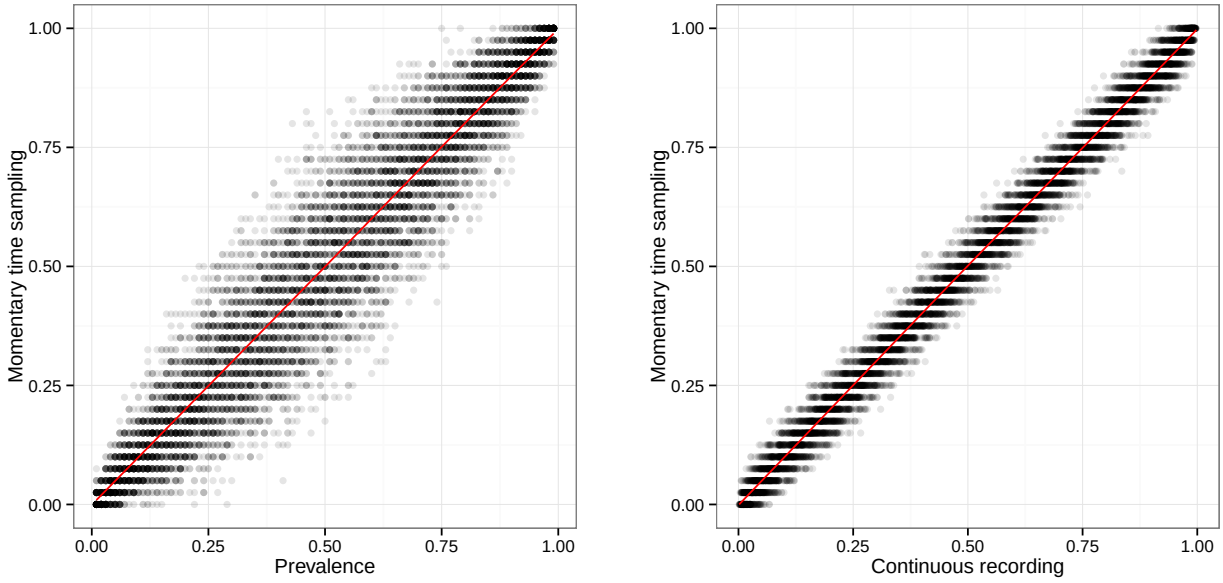
In this section, we provide three brief illustrations of how the `ARPobservation` package can be used to study the validity and reliability of direct observation data. Our intent is to highlight useful types of analysis that can be done with this tool, rather than to provide a exhaustive study of any one topic. The first two illustrations demonstrate the sort of systematic analysis that one might pursue in a methodologically-oriented study, while the final illustration demonstrates a more tailored, smaller scale analysis that could be used to inform the design of an empirical study.

Two measurands for momentary time sampling

As we discussed in a previous section, two possible measurands can be considered in the ARP model: one based on the behavior stream observed during a given session, the other based on the parameters of the underlying data-generating model. The R package can be used in conjunction with either conception, though the latter is somewhat more natural in ARP models. In this example, we show that the two conceptions have somewhat different implications for the reliability of momentary time sampling (MTS) measurements.

In the behavioral parameter conception, MTS data are taken to be measurements of prevalence, as based on the parameters of the ARP model. The bias and variance of MTS data are therefore evaluated with respect to the value of prevalence used to generate the data. To illustrate this, we simulated 600 s behavior streams for varying values of prevalence (ranging from .01 to .99 in steps of .01), while holding incidence fixed at one behavior per 60 s. For sake of simplicity, we arbitrarily chose to use exponential distributions to generate event durations and interim times, assuming equilibrium initial conditions. For each combination of prevalence and incidence, we simulated ten behavior streams, yielding a total of 9900 in all. We then applied 15 s MTS to each simulated behavior stream.[12]

[12]The code used in this simulation can be accessed by typing `demo(MTS_measurands)` at the R command line.



(a) MTS versus prevalence parameter

(b) MTS versus continuous recording

Figure 1. Simulated MTS observations versus two different targets of measurement. Red lines indicate average of MTS datapoints given a value of the measurand.

Figure 1a plots the simulated MTS measurements (Y^M) versus the values of prevalence used to generate the behavior streams. It can be seen that the average values lie along the 45 degree line, which implies that MTS produces unbiased measurements of prevalence. It can also be seen that the reliability of the MTS measurements depends on prevalence: MTS measurements are substantially more variable (less reliable) when prevalence is near .50 than when prevalence is closer to the extremes of zero or one.

In the observed behavior conception, MTS measurements are compared to the continuously recorded measurements (Y^C) from the same observation session, rather than to parametric prevalence. To illustrate this, we applied continuous recording to the same set of simulated behavior on which the MTS measurements were calculated. Figure 1b plots the simulated MTS measurements against the corresponding continuous recording measurements. For a given value of Y^C, it can be seen that the average value of Y^M again lies along the 45 degree line; thus, Y^M is unbiased for Y^C. However, holding Y^C fixed leads

to less variability in the MTS values, and the degree of variability in Y^M depends only weakly on the level of Y^C. It follows that the extent to which the reliability of momentary time sampling data varies with its average level depends on whether one adopts an observed behavior conception or behavioral parameter conception of the measurand.

This illustration is based on a single, common value of the incidence parameter and uses exponential distributions for event durations and interim times. In further analysis, one might examine the extent to which the properties of MTS measurements described here are sensitive to use of other probability distributions for event durations and interim times. One might also investigate whether the properties of MTS measurements change if both Y^C and the observed event frequency (Y^E) are simultaneously held fixed.

Bias in partial interval recording

It has long been recognized that partial interval recording (PIR) measurements over-estimate the prevalence of the behavior (e.g. Altmann, 1974). This systematic bias has been documented and studied with empirical data (e.g. Mann, Ten Have, Plunkett, & Meisels, 1991; Powell et al., 1975), simulations (such as those reviewed in an earlier section), and mathematical analysis (Kraemer, 1979; Rogosa & Ghandour, 1991). PIR measurements are sensitive to several distinct factors, including both the prevalence and incidence of the behavior, as well as the length of the recording interval. Less widely recognized is that the extent of the bias also depends on the generating distribution for interim times (Pustejovsky, 2014).

To illustrate this dependence, we simulated behavior streams from an equilibrium ARP using various interim time distributions. Lacking empirical or theoretical justification for any particular distribution, we examined a diverse (though admittedly arbitrary) range of distributions, including continuous uniform, exponential, gamma, and Weibull distributions; for the gamma and Weibull distributions, we also varied the shape parameter. Holding prevalence fixed at 0.1, we varied incidence over a wide range because the bias depends interactively on both incidence and the interim time distribution. We

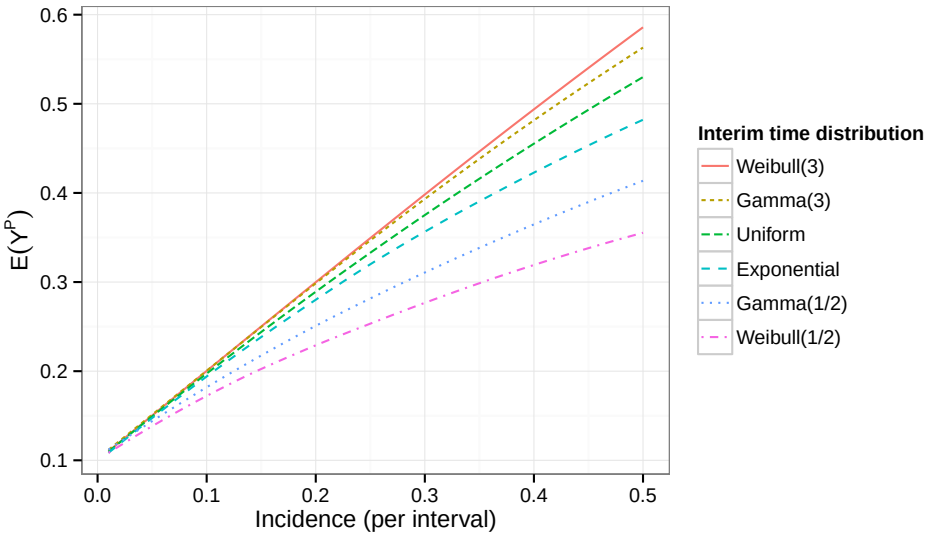


Figure 2. Simulated mean of partial interval recording observations versus incidence for various generating distributions, with prevalence fixed at .10 and equilibrium initial conditions.

simulated behavior streams having a length of 60 intervals, using constant event durations; neither of these factors affects the bias of PIR measurement (Pustejovsky, 2014).[13]

Figure 2 plots the mean value of PIR measurements for a range of incidence values and interim time distributions. It is apparent that the bias of PIR measurements increases with incidence. However, the amount of bias depends as well on the interim time distribution, and the strength of this dependence increases for higher levels of incidence. For instance, when incidence is once per ten intervals, PIR measurements have a marked bias, but the bias is fairly insensitive to the interim time distribution: the difference between the highest and lowest level of bias is only 0.02. When incidence is once per three intervals, the bias not only increases, but also becomes more sensitive to the choice of interim time distributions, with the highest and lowest levels of bias differing by 0.14.

Wirth et al. (2014) have speculated about the possibility of using simulation results to account for the systematic biases of PIR data, either in the planning stages of a study or

[13]The code used in this simulation can be accessed by typing `demo(PIR_bias)` at the R command line.

when analyzing results. The analysis presented here suggests that identifying valid means of doing so may be quite challenging, due to the many factors that influence the level of the bias. In particular, it would seem difficult to make an informed judgement about what type of distribution is an appropriate model for a given behavior, which would be necessary in order to adjust for the bias.[14]

Using simulation to develop a measurement strategy

The previous examples have demonstrated how `ARPobservation` can be used in systematic investigations of the reliability and validity of direct observation procedures, similar to approaches taken in previous, comprehensive simulation studies. Simulation tools can also be useful in more specific, applied contexts as well. Here, we demonstrate how the package might be used to test out alternative measurement strategies during the planning stages of a study.

Imagine that we are planning to evaluate the effect of a non-contingent reinforcement (NCR) on the self-injurious behavior of an 8-year-old boy, Raymond, diagnosed with autism, using an ABAB design. We need to determine whether to use continuous recording or MTS to measure Raymond's self-injurious behavior. We also want to know whether 5 m observation sessions will provide measurements with adequate reliability, or whether longer, 10 m sessions are needed. To investigate these question, we simulate some hypothetical ABAB designs that use the alternative measurement strategies, plot the resulting data, and assess the extent to which the graphs provide a clear basis for determining whether a functional relationship is present.

Before generating hypothetical data, we first need to make some assumptions about the behavior and the effect of intervention. On the basis of initial observations and interviews with his caregivers, we anticipate that Raymond will engage in self-injurious

[14] In other work, we have described methods for analyzing the bias of PIR measurements that circumvent the issue of identifying appropriate interim time distributions through the use of sensitivity bounds (Pustejovsky & Swan, 2014).

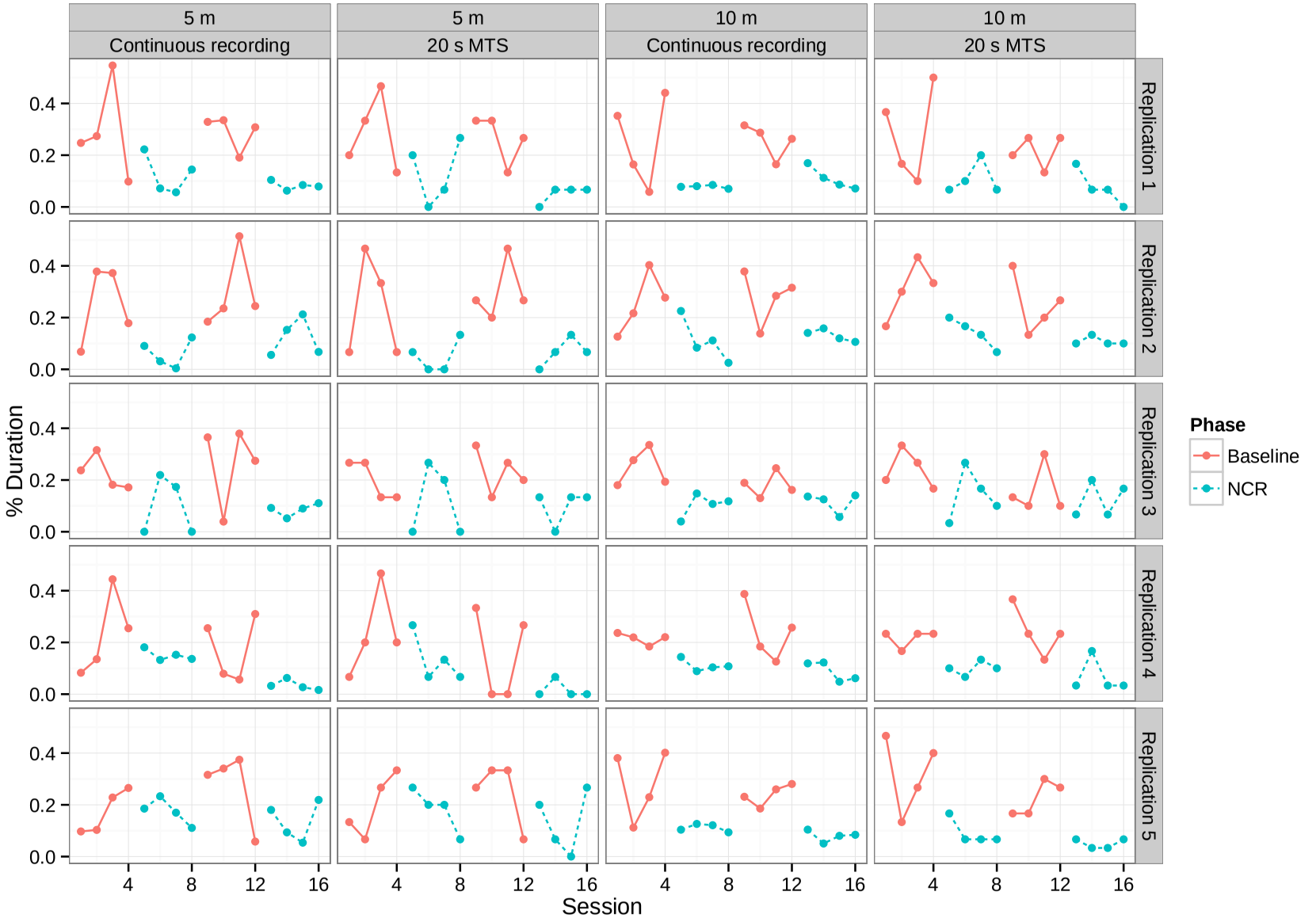


Figure 3. Simulated ABAB designs using 5 m versus 8 m observation sessions and continuous recording versus momentary time sampling.

behavior approximately 25% of the time during the baseline phases, and that bouts occur once per minute, on average. We also anticipate that NCR may reduce the duration of behaviors by 60% or more. We have little understanding of what type of probability distributions would provide good model's for self-injurious behavior, and so we arbitrarily decide to use exponential distributions for the event durations and interim times. Having laid out some tentative assumptions, we use `ARPobservation` to simulate hypothetical behavioral observation data following an ABAB design. We generate 5 studies using each of the following measurement procedures: 5 m of continuous recording, 5 m of MTS, 10 m of continuous recording, and 10 m of MTS.[15]

Figure 3 presents plots of the results, with each panel representing one hypothetical ABAB design; the columns correspond to different measurement strategies, while the rows correspond to hypothetical replications. Careful inspection across the columns suggests that there is little difference between continuous recording and 20 s MTS. Given that the latter procedure is easier to implement, we decide to use MTS. However, regardless of whether continuous recording or MTS is used, measurements based on 5 m sessions are less reliable (more variable) than measurements based on 10 m sessions. The difference between baseline phases and NCR phases is less distinct in the 5 m sessions than in the longer sessions, making it more difficult to detect the presumed functional relationship. If reliability were our sole concern, than we would choose to use 10 m sessions. In practice of course, the benefits of higher reliability need to be weighed against potential risks to Raymond.

Following similar procedures, we might investigate other aspects of our study design, such as whether increasing the number of observations in each phase from 4 to 6 would improve our ability to detect a functional relationship. It would also be prudent to investigate whether using alternative distributional assumptions have any practical

[15]The code used in this simulation can be accessed by typing `demo(study_planning)` at the R command line.

implications for our measurement strategy. Finally, for purposes of illustration we have depicted only 5 designs of each type; systematic assessment across further replications would lead to greater confidence in our conclusions. In this example, we have used visual inspection of the simulated data to inform our design decisions. Alternately, we could have made the judgements based on effect size statistics, inferential statistical tests, or some combination of these as well as visual inspection.

Discussion

The ARP model provides a flexible approach to simulating behavior streams and summary measurements generated by applying direct observation procedures to those streams. Furthermore, its general formulation appears to be a useful organizing framework, insofar as the majority of past simulation research on direct observation procedures is based on specific cases of the ARP. We have introduced `ARPobservation`, a free software package for the R statistical computing environment that can be used to simulate behavior streams and summary measurements based on different observation recording procedures. We have provided a few illustrations of how the package can be used to study the validity and reliability of behavioral observation data, though these were far from comprehensive analyses. We hope that the availability of fast and flexible software may stimulate further inquiry in this area, including both systematic research and informal experimentation. Applied researchers may find `ARPobservation` to be particularly useful for the latter purpose, as a means to build intuition about different observation recording procedures and develop measurement strategies for use in their research.

In our assessment, the ARP model has two key advantages over the random onset model, which is the only other general approach to simulating behavior streams that has been studied in past research. First, the ARP can be used with multiple conceptions of measurands, including both the observed behavior approach or the behavioral parameter approach, whereas the random onset model is premised on the observed behavior conception. Second, the assumptions of the ARP are precisely articulated, making it easier

to reason about them and to test their implications against empirical data. In comparison, the assumptions of the random onset model are less clear, making it more difficult to assess whether they represent a good model for real behaviors.[16] Still, we should emphasize that the drawbacks to the random onset model do not invalidate the findings of methodological studies that have used it. Rather, by studying the validity and reliability of recording procedures using a distinct model, such studies provide a complement to studies based on the ARP model.

Having noted the advantages of the ARP model, it is important to acknowledge its limitations as well. The ARP model describes only simple behavior streams, in which a target behavior is either present or absent at any given point in time. The model does not accommodate behaviors where the intensity of response is of primary interest, nor does it describe contexts in which the inter-relationships between multiple behaviors are the focus.[17] Of course, such contexts also require more sophisticated approaches to data collection, summarization, and analysis (cf. Bakeman & Quera, 2011). Thus, we would argue that the ARP is an appropriate model for contexts in which common observation recording procedures and summary measurements are used. Another limitation is that we have treated the recording procedures as exact algorithms, without allowing for errors on the part of the observer. This approach allows the inherent sampling error of a recording procedure to be isolated from effects of observer errors (Wirth et al., 2014). However, in practice both types of errors will occur and must be weighed when choosing a recording procedure. Development of models that can describe both inherent sampling error and observer error would be useful.

In our judgement, further research on direct observation recording procedures is

[16]In particular, the random onset model's anachronic process (in which behaviors are assigned to onset times out of chronological order) strikes us as less intuitive than the sequential data-generating process of the ARP, which conforms to how behavior streams are actually perceived in time.

[17]However, it is possible to characterize patterns of multiple behaviors through the use of Semi-Markov models, which are a generalization of the ARP (e.g., Engel, 1996).

warranted, given that few extant simulation studies provided rationales or grounding in real data that is relevant to research in behavioral disorders (cf. Lane & Ledford, 2014). As we noted in the review, most previous ARP studies used only a single class of probability distributions for the event durations and interim times, without attending to whether study results are sensitive to these modeling assumptions. Admittedly, the short simulations that we have presented suffer from the same limitations, in that we have selected probability distributions arbitrarily rather than because they are appropriate for modeling a particular class of real behavior. Unfortunately, we are not aware of existing research that would allow us to make more informed assumptions. To address this gap, future work should investigate the distribution of event durations and interim times in samples of real, continuously recorded data, such as that used in previous empirical studies of direct observation recording procedures (e.g., Gardenier et al., 2004; Murphy & Goodall, 1980). Until such research is available, we recommend that further simulation research based on the ARP model should examine a wide range of possible distributions.

Ultimately, the utility of the ARP model (or, for that matter, the random onset model) depends on the extent to which it provides a good description of real patterns of behavior. Closer integration between theoretical simulations and empirical data would help to ensure that the premises and assumptions of further methodological work are well-grounded in the contexts where direct observation recording procedures are widely used. This, in turn, would help to advance the science of direct observation of behavior.

References

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

Ary, D., & Suen, H. K. H. (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.

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

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

Edwards, R., Kearns, K., & Tingstrom, D. H. (1991). Accuracy of long momentary time-sampling intervals: Effects of errors in the timing of observations. *Journal of Psychoeducational Assessment*, *9*(2), 160–165. doi: 10.1177/073428299100900206

Engel, J. (1996). Choosing an appropriate sample interval for instantaneous sampling. *Behavioural Processes*, *38*(1), 11–17. doi: 10.1016/0376-6357(96)00005-8

Gardenier, N. C., MacDonald, R., & Green, G. (2004). Comparison of direct observational methods for measuring stereotypic behavior in children with autism spectrum disorders. *Research in Developmental Disabilities*, *25*(2), 99–118. doi: 10.1016/j.ridd.2003.05.004

Green, S. B., & Alverson, L. (1978). A comparison of indirect measures for long-duration behaviors. *Journal of Applied Behavior Analysis*, *11*(4), 530.

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

Harrop, A., & Daniels, M. (1985). Momentary time sampling with time series data: A commentary on the paper by Brulle & Repp. *British Journal of Psychology*, *76*, 533–537.

Harrop, A., & Daniels, M. (1986). Methods of time sampling: A reappraisal of momentary time sampling and partial interval recording. *Journal of Applied Behavior Analysis*, *19*(1), 73–77.

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.

Kahng, S., Ingvarsson, E. T., Quigg, A. M., Seckinger, K. E., & Teichman, H. M. (2011). Defining and measuring behavior. In W. W. Fisher, C. C. Piazza, & H. S. Roane (Eds.), *Handbook of applied behavior analysis* (pp. 113–131). New York, NY: Guilford Press.

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

Kearns, K., Edwards, R., & Tingstrom, D. H. (1990). Accuracy of long momentary time-sampling intervals: Implications for classroom data collection. *Journal of Psychoeducational Assessment*, *8*(1), 74–85. doi: 10.1177/073428299000800109

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

Leemis, L. M., & McQueston, J. T. (2008). Univariate distribution relationships. *The American Statistician*, *62*(1), 45–53. doi: 10.1198/000313008X270448

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

Martin, P., & Bateson, P. (2007). *Measuring Behaviour: An Introductory Guide* (3rd ed.). Cambridge, England: Cambridge University Press.

Meany-Daboul, M. G., Roscoe, E. M., Bourret, J. C., & Ahearn, W. H. (2007). A comparison of momentary time sampling and partial-interval recording for evaluating functional relations. *Journal of Applied Behavior Analysis*, *40*(3), 501–514. doi: 10.1901/jaba.2007.40-501

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Powell, J. (1984a). On the misrepresentation of behavioral realities by a widely-practiced direct pbservation procedure: Partial interval (one-zero) sampling. *Behavioral Assessment*, *6*(3), 209–219.

Powell, J. (1984b). Some empirical justification for a modest proposal regarding data acquisition via intermittent direct observation. *Journal of Behavioral Assessment*, *6*(1), 71–80.

Powell, J., Martindale, A., & Kulp, S. (1975). An evaluation of time-sample measures of behavior. *Journal of Applied Behavior Analysis*, *4*(4), 463–469.

Powell, J., & Rockinson, R. (1978). On the inability of interval time sampling to reflect frequency of occurrence data. *Journal of Applied Behavior Analysis*, *11*(4), 531–532.

Pustejovsky, J. E. (2014). Measurement-comparable effect sizes for single-case studies of

free operant behavior. *Psychological Methods*, (In press). doi: 10.1037/met0000019

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.

Quera, V. (1990). A generalized technique to estimate frequency and duration in time sampling. *Behavioral assessment*, *12*(4), 409–424.

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., Michalski, D. N., Carroll, R. A., & Lindenberg, A. M. (2008). Detecting changes in simulated events using partial-interval recording and momentary time sampling. *Behavioral Interventions*, *23*, 237–269.

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.

Repp, A. C., Roberts, D. M., Slack, D. J., Repp, C. F., & Berkler, M. S. (1976). A comparison of frequency, interval, and time-sampling methods of data collection. *Journal of Applied Behavior Analysis*, *9*(4), 501–508.

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

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

Rojahn, J., & Kanoy, R. (1985). Toward an empirically based parameter selection for time-sampling observation systems. *Journal of Psychopathology and Behavioral Assessment*, *7*(2), 99–120.

Teetor, P. (2011). *R Cookbook*. Sebastopol, CA: O'Reilly Media, Inc.

Thompson, F., Symons, F. J., & Felce, D. (2000). Principles of behavioral observation: Assumptions and strategies. In T. Thompson, D. Felce, & F. J. Symons (Eds.), *Behavioral observation: Technology and applications in developmental disabilities* (pp. 3–16). Baltimore, MD: Paul H. Brookes Publishing Co.

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

Wilson, R. R., Jansen, B. D., & Krausman, P. R. (2008). Planning and assessment of activity budget studies employing instantaneous sampling. *Ethology*, *114*(10), 999–1005. doi: 10.1111/j.1439-0310.2008.01544.x

Wirth, O., Slaven, J., & Taylor, M. A. (2014). Interval sampling methods and measurement error: A computer simulation. *Journal of Applied Behavior Analysis*, *47*(1), 1–18. doi: 10.1002/jaba.93