Measurement-comparable effect sizes for single-case studies of free-operant behavior

James E. Pustejovsky
The University of Texas at Austin

Single-case research comprises a set of designs and methods for evaluating the effects of interventions, practices, or programs on individual cases, through comparison of outcomes measured at different points in time. Although there has long been interest in meta-analytic technique for synthesizing single-case research, there has been little scrutiny of whether proposed effect sizes remain on a directly comparable metric when outcomes are measured using different operational procedures. Much of single-case research focuses on behavioral outcomes in free-operant contexts, which may be measured using a variety of different direct observation procedures. This article describes a suite of effect sizes for quantifying changes in free-operant behavior, motivated by an alternating renewal process model that allows measurement comparability to be established in precise terms. These effect size metrics have the advantage of comporting with how direct observation data are actually collected and summarized. Effect size estimators are proposed that are applicable when the behavior being measured remains stable within a given treatment condition. The methods are illustrated by two examples, including a re-analysis of a systematic review of the effects of choice-making opportunities on problem behavior.

Keywords: effect size; single-case research; free-operant behavior; alternating renewal process

Single-case research comprises a set of designs and methods for evaluating the effects of interventions, practices, or programs on individual cases. Empirical single-case research appears in many areas of psychology and education, particularly in special education, school psychology, clinical psychology, psychotherapy, social work, and applied behavior analysis (e.g. Horner et al., 2005; Kazdin, 2011; Kennedy, 2004). Single-case research emphasizes individual change, focusing on individual-specific effects of intervention that are identified through comparison of outcomes measured repeatedly on the same case at different points in time. Though most single-case studies actually include multiple cases, it is typical to report separate results for each case, with little emphasis on overall summaries across cases.

Despite this idiographic orientation, there has long been interest in meta-analytic synthesis as an approach for summarizing and generalizing from single-case studies. Just as other fields began to take growing interest in meta-analysis, Gingerich (1984) argued that synthesis of single-case research could improve the precision of individual treatment effect estimates, bolster the internal validity of single studies through replication, and provide a means for studying variation in treatment effects and generalizing from a collection of studies. Others noted the drawbacks of excluding single-case studies from comprehensive syntheses (e.g., Allison & Gorman, 1993). In response, several effect size statistics and meta-analytic approaches have been developed specifically for single-case research (e.g., Busk & Serlin, 1992; Center, Skiba, & Casey, 1985; Scruggs, Mastropieri, & Casto, 1987). With the more recent and growing emphasis on evidence-based practice, fields that use single-case research have articulated standards of scientific evidence and attempted to codify synthesis methods (Chambless & Ollendick, 2001; Gast, 2010; Horner et al., 2005; Kratochwill & Stoiber, 2002; Odom et al., 2005). As interest has grown, systematic reviews and syntheses of single-case research now appear with increasing frequency (Maggin, O'Keeffe, & Johnson, 2011). However, there is still little consensus regarding appropriate statistical methods for such synthesis. Even the most basic question of what effect size metric to use for meta-analysis

James E. Pustejovsky, 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.

remains unresolved, though proposals have proliferated (for reviews of current proposals, see Beretvas & Chung, 2008; Wolery, Busick, Reichow, & Barton, 2010).

Effect sizes are the the basic units of analysis in a meta-analysis, and operational definition of an effect size metric is one of the central questions in any quantitative research synthesis (Cooper, 2009). The choice of effect size metric involves, implicitly or explicitly, an assumption about the comparability of results from different studies that may use various participant inclusion criteria, treatment procedures, outcome measurement instruments, or experimental designs (Hedges, 2008). To account for differences in measurement instruments, it is desirable that the magnitude of an effect size not depend strongly on specific or idiosyncratic features of the instrument; I will call effect sizes that have this property *measurement-comparable*. Without measurement comparability, it becomes very difficult to draw meaningful inferences from averages across or comparisons between effect size estimates because true variation in magnitude is confounded by differences in measurement scales.

Issues of measurement-comparability have received scant attention in the context of single-case research. Two of the main classes of effect sizes proposed for use in single-case research are standardized mean differences and non-overlap measures (Beretvas & Chung, 2008; Wolery et al., 2010). Busk and Serlin (1992) proposed a standardized mean difference effect size where the standardization is based on the within-case sample variance; some have argued that standardizing by the within-case variance makes the resulting effect sizes measurement-comparable (Hershberger, Wallace, Green, & Marquis, 1999; Van den Noortgate & Onghena, 2003). However, this claim is justified only by analogy to the standardized mean difference metric used in between-subjects research, rather than by any explicit model. The second class of effect sizes includes the percentage of non-overlapping data (Scruggs et al., 1987) and a variety of other effect sizes connected to robust statistics and non-parametric tests (Parker, Vannest, & Davis, 2011). Parker, Vannest, and Davis (2011) and Scruggs and Mastropieri (2012) argue that non-overlap measures permit direct comparison across measurement procedures because they are all on a scale ranging from 0 to 100%. Again, the claim of measurement comparability is not motivated by any explicit model.

Another limitation of current effect size proposals for single-case research is the lack of specific attention to measurement scales.[1] Effect sizes in the family of standardized mean differences are appropriate for interval-scale measurements and are often employed in connection with normally distributed outcome measures, yet many of the most common measurement procedures used in single-case research produce measurements that are not interval scaled or normally distributed. Non-overlap measures seek to avoid making any sort of distributional assumptions; consequently, it is largely unknown whether and how these measures are sensitive to variations in the measurement scales.

In this paper, I propose several new effect size measures for single-case research that attend closely to issues of measurement-comparability and scaling. All of the effect sizes are defined in terms of a common underlying model, which permits their measurement comparability (or lack thereof) to be established in precise terms. The model also comports closely with commonly used measurement procedures, leading to effect sizes that are more appropriate for the scales of measurements produced by these procedures.

Rather than addressing the comparability of any and all outcome measurement operations used in single-case research, the scope of this investigation is limited to one particular class: direct observations of behavior in free-operant contexts. Free-operant contexts are defined by a setting or time-frame in which behaviors are free to occur at any time, without prompting or restriction by the investigator.[2] For instance, an investigator might observe the bullying behavior of a child during lunch recess (e.g. Ross & Horner, 2009), recording incidents as they occur over the course of the child's natural interactions with his peers.

Focusing only on free-operant behavior is appropriate for three reasons. First, directly observed free-operant behavior is the most common class of outcome measures used in single-case research (Kazdin, 2011; Pustejovsky, 2013); thus, even methods specialized to this outcome domain will still be widely applicable. Second, several past syntheses of single-case research have employed similar scope limitations, examining only studies of free-operant behavior (e.g., Gage, Lewis, & Stichter, 2012; Hart & Banda, 2009; Shogren, Faggella-Luby, Bae, & Wehmeyer, 2004) or drawing distinctions between outcome domains at the analysis stage (e.g., Machalicek et al., 2008). Finally, a measurement comparability model that is applicable across multiple outcome domains would almost certainly involve stronger and more tenuous assumptions than a model for a single domain. This initial study of measurement-comparable effect sizes therefore develops methods for a single, commonly used domain, where a relatively plausible set of modeling assumptions can be articulated and studied.

[1] I am aware of only one exception: Shadish, Kyse, and Rindskopf (2013) consider binomial models for proportion data and poisson models for count data. However, their focus is purely on developing models for individual single-case studies; they do not consider issues of effect size definition or measurement comparability.

[2] The behavior-analytic tradition defines free-operant behavior slightly differently, as behavior stimulated by its consequences, rather than by antecedent prompts or cues (Johnston & Pennypacker, 1993, p. 366).

Measurement procedures

In order to develop measurement-comparable effect sizes, it is important to understand the range of operational procedures across which comparisons might be made. An array of procedures are used to measure behavior in free-operant contexts; four of the most common are event counting, continuous recording, momentary time sampling, and partial interval recording.[3] Using *event counting*, the observer notes the start of each occurrence of a behavior, either by recording the time of each occurrence or simply by tallying the number of occurrences. Data from an observation session are often summarized by the number of occurrences per fixed unit of time.

Using *continuous recording*, the observer notes the beginning and end of each instance of a behavior. Data from an observation session are often summarized by the proportion of session time during which a behavior is observed.

Using *momentary time sampling*, the observer notes whether a behavior is or is not occurring at each of a set of fixed moments in time, which are typically equally spaced over the course of a session. Data from an observation session are often summarized by the proportion of moments at which the behavior is observed.

Using *partial interval recording*, an observation session is divided into short time intervals, sometimes with a short break in between each interval to allow time for recording observations. The observer scores each interval according to whether the behavior is observed; an interval receives a score of one if the behavior occurs at any point during the interval, and otherwise receives a score of zero.[4] Data from an observation session are typically summarized by the proportion of intervals receiving a score of one (or equivalently, the mean score across the intervals).

In practice, these different procedures can be applied to measure very similar constructs. For example, Shogren et al. (2004) conducted a systematic review of single-case studies examining the effects of providing choice-making opportunities on the problem behavior of disabled children. The authors identified 13 studies meeting search criteria, including a total of 32 individual cases. The primary outcome for each case was problem behavior, but the procedures used to measure problem behavior varied across cases and studies. Table 1 reports the number of studies and number of individual cases where each recording procedure was employed. For the majority of studies and cases, problem behavior was measured using interval recording (following a partial interval recording procedure in all but one case). The next most commonly used method was continuous recording, applied with five cases from a single study by Romaniuk et al. (2002).[5] This systematic review, which attempted to synthesize studies using heterogeneous operational procedures for measuring a common construct, provides further motivation for a careful examination of measurement comparability.

Table 1
Measurement procedures used in studies included in Shogren et al. (2004)

Procedure	Studies	Cases
Event counting	3	3
Continuous recording	1	5
Momentary time sampling	1	1
Partial interval recording	7	19
Other	3	4
Total[a]	13	32

[a] Some studies used more than one measurement procedure.

Modeling approach

To study the comparability of these different measurement procedures, I use a model for the free-operant behavior that is observed over the course of an observation session, or what is sometimes called the "behavior stream" (Hartmann & Wood, 1990; Schoenfeld, 1972). By defining the targets of measurement in terms of the behavior stream, the model distinguishes clearly between the quantities of interest (the parameters of the behavior stream) and the procedures used to measure those quantities–much as a psychometric measurement model separates the definition of true scores from the particular set of items used to measure those scores. The model therefore establishes a common basis for understanding the various procedures that can be applied to record observations.

The approach to modeling the behavior stream follows Rogosa and Ghandour (1991), who used a class of stochastic models called alternating renewal processes to study the reliability of different behavioral observation procedures. Though complex, this model is useful because it is based on a plausible description of the behavior being observed and the

[3] Detailed surveys of measurement methods are available from several sources, including Ayres and Gast (2010) or Kahng, Ingvarsson, Quigg, Seckinger, and Teichman (2011). Terminology varies somewhat across authors. My presentation follows the terminology and the main outline of Ayres and Gast (2010).

[4] Another procedure known as *whole interval recording* is structured in the same way as partial interval recording, but the rule for scoring each interval is different. With whole interval recording, an interval is scored as a one if the behavior occurs for the entire active interval, and is otherwise scored as a zero. Whole interval recording captures the same information as partial interval recording applied to the absence of a behavior rather than its presence. Both because of this structural relationship and because it is much less common than partial interval recording, I do not discuss whole interval recording further.

[5] Four cases were measured using procedures that I have categorized as "other." All of these cases were measured using esoteric procedures that were highly adapted to the individual's context.

process through which measurements are recorded. Furthermore, Rogosa and Ghandour (1991) illustrated that the properties of behavioral observation data differ markedly from what one would predict based on models for interval-scale measures and normally distributed errors. Effect sizes based on this model for the behavior stream may therefore lead to substantially different conclusions than previous proposals in the literature. In the following sections, I refer to the alternating renewal process model as the *within-session* model because it has to do with measurements made and recorded during the course of a single observation session.

In addition to the within-session model, a *between-session* model is needed to describe changes in behavior across subsequent observation sessions and phases of the design. I will consider only the simplest possible such model, assuming that the behavior stream process is stable within a given phase (lacking time trends), and therefore that repeated measurements are identically distributed and uncorrelated. This simplistic model is justified for several reasons. The assumption that the pattern of behavior does not have time trends is employed for simplicity, in order to more clearly define effect sizes that quantify changes between phases. Extensions to models with time trends are possible (Pustejovsky, 2013) and will be explored further in future work. The assumption that repeated measurements are uncorrelated runs against the current consensus that statistical models for single-case studies should allow for auto-correlation (Horner, Swaminathan, Sugai, & Smolkowski, 2012; Wolery et al., 2010). However, auto-correlation is not a clear-cut concept in models with non-normal error distributions, and introducing it would add several layers of complication. The focus of this paper is therefore on models without auto-correlation. I comment further on these issues in the discussion section.

The remainder of this paper is organized as follows. In the next section, I outline a within-session model of the major measurement procedures for direct behavioral observation in free-operant contexts. I then define several different effect size measures for quantifying changes in free-operant behavior and note the relationships among the different measures. Next, I describe simple moment estimators that can be applied when behavior is stable within phases and repeated measurements are uncorrelated. To demonstrate the proposed methods, I re-analyze a single study by Romaniuk et al. (2002) and a systematic review by Shogren et al. (2004). I conclude by discussing the choice between alternative effect sizes, noting limitations, and highlighting avenues for future work.

Within-session model

In a typical single-case study, only one measurement per observation session is reported (often by plotting it in a single-case graph), and it is only this information that will be available for secondary meta-analysis. However, a much more fine-grained model is needed in order to establish the measurement comparability of different procedure–one that describes the entire sequence of behaviors that occur over the course of an observation session. I first describe a parametric model for the behavior stream, known as the equilibrium alternating renewal process model. I then examine the implications of the model for each of the main procedures for measuring free-operant behavior.

Equilibrium alternating renewal process

The equilibrium alternating renewal process (ARP) is a model for the sequence of behavioral events that occur during the course of an observation session. It assumes that these events occur singly and sequentially–that is, one event must end before the next event begins, with some time in between where no event is occurring. Some notation is needed for describing such a sequence of events. Let L denote the length of the observation session. Denote the duration of the first event as A_1, the duration of the second event as A_2, and the duration of event u as A_u, for $u = 3, 4, 5, ...$ Let B_0 denote the length of time until the first behavioral event, with $B_0 = 0$ if event 1 is occurring at the beginning of the observation period. Finally, let B_u denote the length of time between the end of event u and the beginning of event $u + 1$, or what I will call the u^{th} *interim time*, for $u = 1, 2, 3, ...$ In this notation, the values $B_0, A_1, B_1, A_2, B_2, A_3, B_3, ...$ provide a quantitative description of the behavior stream observed during a given session. Figure 1 depicts the behavior stream graphically.

To make this notation more concrete, consider the example of a school psychologist who is interested in measuring a child's out-of-seat behavior during math class. During a given session, the psychologist observes the times at which the child leaves her seat and times at which she returns–it is these observations that comprise the behavior stream. The interim time B_0 is the length of time from the beginning of math class until the child first leaves her seat; the event duration A_1 is the length of time that the child is out of her seat on the first occasion of her getting up; B_1 is the length of time from when she first returns to when she gets up again; A_2 is the length of time she is out of her seat on the second occasion, and so forth.

In the ARP model, the length of each behavioral event and the length of each interim time are treated as random quantities, each following some probability distribution. The model has four key assumptions.[6] First, event durations $A_1, A_2, A_3, ...$ are assumed to be identically distributed random quantities with mean μ and cumulative distribution function $F(t; \mu)$.[7] The parameter μ thus represents the average length of each behavioral event, which is strictly pos-

[6] For an introduction to renewal process models, see Cox (1962) or Kulkarni (2010, Chp. 8).

[7] For a random variable A with mean μ, the cumulative distribution function $F(t; \mu)$ gives the probability that A is less than or equal

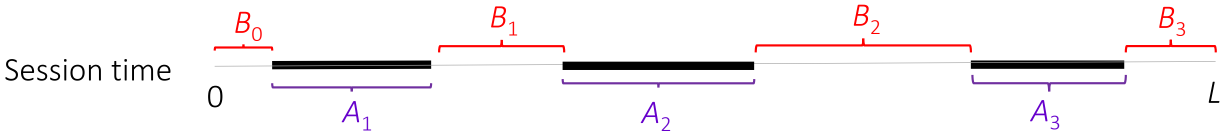

Figure 1. Organization of the behavior stream. The thin grey line represents time during an observation session, ranging from 0 to L. The thick black lines represent behavioral events. Behavioral event durations are labelled A_1, A_2, A_3. Interim times are labelled B_0, B_1, B_2, B_3.

itive and finite. Second, interim times $B_1, B_2, B_3, ...$ are assumed to be identically distributed random quantities with mean λ and cumulative distribution function $G(t; \lambda)$. The parameter λ thus represents the average interim time, which is also strictly positive and finite. Third, all interim times and all event durations are assumed to be mutually independent. This means that the length of a given event is not influenced by the length of previous events, or by how long it has been since the last event ended. Fourth, the process is assumed to be aperiodic and in equilibrium. This final assumption means that events are no more or less likely to occur at the beginning of a session than in the middle or at the end; instead, the probability that an event is occurring any given point in time during the observation session remains constant.

Note that the first and second assumptions are not specific about the precise form of the probability distributions for the event durations and interim times. Instead, the ARP can encompass a very wide variety of parametric distributions. For example, event durations might be log-normally distributed while interim times are exponentially distributed, or event durations and interim times could both follow gamma distributions. The ARP 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. As will be seen, this general formulation of the ARP is useful for understanding the properties of the main measurement procedures.

In the ARP, the main parameters that describe the behavior being observed are the mean event duration μ and the mean interim time λ. In addition to these parameters, two other characteristics of the behavior will also be of interest, both of which are functionally related to μ and λ. First, the incidence of the behavior, denoted ζ, is the average rate at which new events occur; incidence is related to the mean event duration and mean interim time by $\zeta = 1/(\lambda + \mu)$. Second, the prevalence of the behavior, denoted ϕ, is the overall proportion of time that behavioral events occur; prevalence is related to mean event duration and mean interim time by $\phi = \mu/(\mu + \lambda)$.

Implications for measurement procedures

Any of the main procedures could in principle be applied to the same behavior stream, generating a different summary measurement. The ARP model for the behavior stream provides a common basis for understanding the properties of and relationships between the measurements generated by different procedures. After introducing some further notation, I provide the expected value of each measurement procedure, expressed in terms of the parameters of the behavior stream. Derivations and further technical details regarding these expressions are given in the Appendix.

Using event counting, the observer notes the beginning of each new event. The summary measurement generated by this procedure is the total number of new events that begin during the observation session; I denote it as Y^E. In the earlier example, Y^E is the number of times that the child leaves her seat during the course of the session. Under the ARP, the expected value of an event counting measurement is

$$\mathrm{E}\left(Y^E\right) = \zeta L. \tag{1}$$

Intuitively, the expected number of events during the session is equal to the average rate of events per time unit, multiplied by the length of the session. The expectation does not depend on the specific form of the event duration or interim time distributions, a property which makes event counting data quite simple to interpret.

Using a continuous recording procedure, the observer notes the beginning and end of each behavioral event. The summary measurement from a continuous recording procedure is calculated as the proportion of session time during which the behavior occurs; I denote it as Y^C. In the earlier example, Y^C is the proportion of time that the child is out of her seat during the course of that particular math class. Under the ARP, the expected value of a continuous recording measurement is equal to prevalence:

$$\mathrm{E}\left(Y^C\right) = \phi. \tag{2}$$

Intuitively, the expected proportion of time that events occur during a specific session is equal to the long-run average proportion of time that events occur. Like event counting, the expectation of continuous recording data does not depend on the specific form of the event duration or interim time distributions, which makes interpretation simple.

Using a momentary time sampling procedure, the observer divides the session up into many equally-spaced intervals and notes the presence or absence of a behavior at

to a given value t.

the moment each interval ends. Let K denote the number of intervals, so that the length of each interval is L/K. The summary data point, which I will denote Y^M, is calculated as the proportion of these moments (out of K possible) at which the behavior is observed. In the earlier example, suppose that the observer uses 15 s momentary time sampling. The observer would note whether the child is out of her seat at time 15 s, at time 30 s, at time 45 s, etc.; Y^M would then be calculated as the proportion of these moments at which the child is out of her seat. Under the ARP, the expected value of a momentary time sampling measurement is equal to prevalence:

$$\mathrm{E}\left(Y^M\right) = \phi. \tag{3}$$

This is true because the process is in equilibrium, and so the probability that an event is occurring at any given moment is equal to the behavior's prevalence. Thus, momentary time sampling produces measurements of the same behavior stream parameter as does continuous recording, with an expected value that does not depend on the specific form of the event duration or interim time distributions.

In partial interval recording, an observer first divides the session into K intervals, each of length L/K. The first c time units of each interval are devoted to observation, while the remainder $L/K - c$ is used for recording or resting; I call c the *active interval length.* During a given interval, the observer counts a behavior as present if it occurs at any point during the active interval. The summary data point from partial interval recording, which I will denote as Y^P, is calculated the proportion of intervals during which the behavior is observed at any point. In the earlier example, suppose that the observer uses $c = 15$ s intervals, each separated by 5 s of rest time. The observer would first record whether the child was out of her seat at any point between 0 s and 15 s, then record whether the child was out of her seat at any point between 20 s and 35 s, whether the child was out of her seat between 40 s and 55 s, etc.; Y^P would then be calculated as the proportion of the intervals during which the child was out of her seat at some point.

In contrast to the other procedures, partial interval recording produces measurements that have no simple interpretation in terms of the main parameters of the ARP.[8] Instead, it can be shown that the expected value of a partial interval recording measurement is

$$\mathrm{E}\left(Y^P\right) = \phi + \zeta \int_0^c \left[1 - G(t; \lambda)\right] dt, \tag{4}$$

where $\int_0^c$ indicates a definite integral over the interval 0 to c. A detailed derivation can be found in the Appendix. The perplexing implication of Equation (4) is that partial interval recording data are sensitive to many factors, including both prevalence and incidence, the active interval length c used in the procedure, and the specific form of the interim time distribution $G(t; \lambda)$. If one interprets partial interval recording measurements as upwardly biased estimates of prevalence, then the degree of bias will be different depending on whether the interim times follow an exponential distribution, a log-normal distribution, or some other distribution. As I illustrate later, these sensitivities create difficulties for estimating effect sizes from partial interval data.

Between-session model and case-level effect size parameters

The ARP model provides a basis for understanding the relationships among measurements generated by different observation procedures and consequently for establishing the measurement-comparability of different effect sizes. This is because effect sizes defined in terms of the parameters of the ARP can be interpreted in terms of the data from *any* of the measurement procedures, rather than being contingent on the procedure. In this section, I propose several such effect sizes for measuring behavioral changes, then describe the relationships among them.

Before turning to the effect sizes, I first need to introduce a further model for descibing multiple observation sessions. The within-session ARP model describes the behavior stream as observed on a single occasion, but single-case studies involve observing the behavior of an individual across several sessions, under different treatment conditions. A between-session model is needed that describes changes in the behavior stream over this longer time frame. I will focus on a very simple model in which the behavior stream remains stable from observation session to observation session but may change under different treatment conditions.

Consider a study in which a total of n outcome measurements are made using measurement procedure r, where $r = E, C, M$, or P. The observed data are then Y_j^r, for $j = 1, ..., n$. To indicate the treatment condition on occasion j, define the covariate X_j for $j = 1, ..., n$, where $X_j = 1$ if the case is in a treatment phase at time j and $X_j = 0$ otherwise. For the between-session model, I assume that 1) during session j, the behavior of the case follows an equilibrium alternating renewal process; 2) measurements generated during different observation sessions are independent of each other; and 3) the parameters of the alternating renewal process are constant within each treatment phase. Let μ_0 denote the average event duration and λ_0 the average interim time during baseline phases (when $X_j = 0$); let μ_1 denote the average event duration and λ_1 the average interim time during treatment phases (when $X_j = 1$).

Under this stable-phase model, measurement-comparable effect sizes involve comparisons between (μ_0, λ_0) and

[8] It has long been recognized that interval recording data measures neither prevalence nor incidence. See for instance J. Altmann (1974) for a discussion of the origins and arguments regarding interval recording methods.

(μ_1, λ_1). A number of such comparisons are worth considering; I will describe five: the log-duration ratio, the log-interim ratio, the log-incidence ratio, the log-prevalence ratio, and the log-prevalence odds ratio. I focus on logged ratios for two reasons. First and perhaps most crucially, published single-case studies often describe results using measures of percentage change (cf. Campbell & Herzinger, 2010), though they do so without any indication of sampling uncertainty. Log-ratios are very closely related to proportionate changes, and thus have the advantage of aligning to some extent with how applied researchers already think. Second, log-ratios are useful in a purely technical sense when they conform to the scales of the quantities being measured and can be defined without range restriction. All but one of the effect size metrics range from negative infinity to positive infinity, with zero corresponding to no change; this allows some problems with meta-analytic models to be avoided.

Log-duration ratio and log-interim ratio

In the abstract, one of the most informative ways to quantify a change in behavior would be to use separate contrasts between each component of the ARP. For instance, the *log-duration ratio*, defined as $\omega^\mu = \ln(\mu_1/\mu_0)$, measures the proportionate change in the average event duration; the *interim ratio*, defined as $\omega^\lambda = \ln(\lambda_0/\lambda_1)$ measures the proportionate change in the average interim time. These two effect sizes would be particularly useful in intervention contexts if the experimenter's goal is to effect change in one dimension of the behavior but not the other, or to evaluate a detailed hypothesis about the mechanism of an intervention. Unfortunately, these effect sizes may only be of hypothetical interest, because none of the observation procedures that I have described yield measurements of the separate components of duration and incidence.

Log-incidence ratio

The *log-incidence ratio* is defined as the log of the ratio of a behavior's incidence in the treatment condition to the behavior's incidence in the baseline condition; this can be written either in terms of incidence or in terms of average duration and interim times:

$$\omega^\zeta = \ln(\zeta_1/\zeta_0) = \ln(\mu_0 + \lambda_0) - \ln(\mu_1 + \lambda_1), \qquad (5)$$

where ζ_0 is the incidence during baseline and ζ_1 is the incidence during treament. This effect size measures the proportionate change in a behavior's incidence; if observation sessions are all of equal length, then the incidence ratio is also equivalent to the log of the proportionate change in the expected number of behaviors during a session. Event counting directly measures incidence and is a very commonly used procedure. Therefore, the log-incidence ratio should be a very useful effect size for describing changes in behavior. However, as a summary metric, it has the disadvantage of not being sensitive to behavioral prevalence. Rather, an observed decrease in incidence could be the result of either an increase in average interim time or an increase in average event duration, two possibilities which might have different substantive implications.

Log-prevalence ratio

When a behavior has non-negligible duration, the foremost concern of an interventionist will often be its prevalence, or the overall proportion of time that it occurs. One metric for quantifying changes in prevalence is the *log-prevalence ratio*, defined as the log of the ratio of prevalence during the treatment condition to prevalence during the baseline condition; this can be written in terms of either prevalence or average duration and interim times:

$$\omega^\phi = \ln\left(\frac{\phi_1}{\phi_0}\right) = \ln\left(\frac{\mu_1}{\mu_1 + \lambda_1}\right) - \ln\left(\frac{\mu_0}{\mu_0 + \lambda_0}\right), \qquad (6)$$

where ϕ_0 is prevalence during baseline and ϕ_1 is prevalence during treatment. This effect size is comparatively straightforward to interpret in terms of proportionate changes, as with the other log-ratio effect sizes. Compared to the log-incidence ratio, the log-prevalence ratio has the advantage of being sensitive to changes in both event duration and interim time. However, there are two inter-related drawbacks to this effect size. First, because prevalence ranges from 0 to 1, the effect size has a range that depends on the initial level: for a given baseline prevalence ϕ_0, the log-prevalence ratio can never be greater than $-\ln(\phi_0)$. Second, the log-prevalence ratio is not symmetric with respect to how behaviors are defined; re-defining prevalence as the proportion of time that behavioral events do not occur will alter the magnitude of the log-prevalence odds ratio, rather than only affecting the sign. Application of the log-prevalence ratio will therefore require establishing conventions as to how behaviors are defined, such as always defining prevalence in terms of negative or undesirable behavior.

Log-prevalence odds ratio

The *log-prevalence odds ratio* is an alternative metric for quantifying changes in prevalence, and is defined as

$$\psi = \ln\left(\frac{\phi_1/(1 - \phi_1)}{\phi_0/(1 - \phi_0)}\right) = \ln(\mu_1/\lambda_1) - \ln(\mu_0/\lambda_0). \qquad (7)$$

This effect size measures proportionate change in the prevalence odds, or the ratio of the average event duration to the average interim time. It is mathematically equivalent to the difference between the log-duration ratio and the log-interim ratio: $\psi = \omega^\mu - \omega^\lambda$. As a result, the log-prevalence odds ratio weighs a given proportionate increase in duration as equal to a corresponding proportionate decrease in interim time. In contrast to the log-prevalence ratio, the range of the log-prevalence odds ratio is unconstrained by baseline prevalence; instead, it ranges from $-\infty$ to ∞. The ratio is symmetric with respect to how behaviors are defined. These mathematical advantages come at the cost of lessened intuitive appeal because researchers and consumers of research may find odds ratios more difficult to interpret than proportionate changes. The difference between the log-prevalence ratio and the log-prevalence odds ratio parallels that between the log-relative risk ratio and the log-odds ratio effect sizes used in other areas of meta-analysis (for further discussion of the latter effect sizes, see Fleiss & Berlin, 2009).

Relationships among effect sizes

The five effect sizes represent different metrics for quantifying change in a behavioral process, as modeled by an ARP. The effect sizes are closely related to one another because they are all defined in terms of ARP parameters. Under certain conditions, the effect sizes also become either approximately or exactly equivalent, in which case it is reasonable to directly compare estimates of those different effect sizes. Understanding the circumstances under which the different effect sizes are measurement-comparable is important for meta-analytic applications, which will often involve combining information across studies that used different outcome measurement procedures.

Figure 2 displays the inter-relationships among the effect sizes, with arrows between two effect sizes indicating equality under the specified condition. If event duration is constant across the baseline and treatment conditions ($\mu_0 = \mu_1$), then the log-prevalence ratio is equal to the log-incidence ratio and the log-prevalence odds ratio reduces to the log-interim ratio. Similarly, if the average interim time is constant across conditions ($\lambda_0 = \lambda_1$), then the log-prevalence odds ratio is equal to the log-duration ratio. If events have very short average duration relative to the average interim time, and this is true under both conditions so that $\mu_0 << \lambda_0$ and $\mu_1 << \lambda_1$, then the log-prevalence ratio will be approximately equal to the log-prevalence odds ratio and the log-incidence ratio will be approximately equal to the log-interim ratio.

Effect size estimators

Thus far, I have described a within-session model for outcome data generated by various measurement procedures, posited a simple between-session model, and defined several different effect size parameters for measuring changes in directly observed free-operant behavior. This section presents methods for estimating those effect sizes based on data from the different measurement procedures.

All of the estimators described in this section involve sample means and sample variances calculated by treatment condition. I will use the following notation for these quantities. Let n_0 denote the number of observations made in the baseline condition and n_1 denote the number of observations made in the treatment condition. For data collected using measurement procedure $r = E, C, M$, or P, let $\bar{y}_0^r$ denote the sample mean outcome in the baseline condition and $\bar{y}_1^r$ denote the sample mean outcome in the treatment condition. Some of the estimators described below cannot be calculated if the sample means are equal to zero. To account for this possibility, I will use truncated sample means defined as

$$\hat{y}_0^r = \begin{cases} \bar{y}_0^r & \text{if} \quad \bar{y}_0^r > 0 \\ k_0^r & \text{if} \quad \bar{y}_0^r = 0 \end{cases} \qquad \hat{y}_1^r = \begin{cases} \bar{y}_1^r & \text{if} \quad \bar{y}_1^r > 0 \\ k_1^r & \text{if} \quad \bar{y}_1^r = 0 \end{cases}$$

for constants k_0^r, k_1^r given below. Finally, let S_{r0}^2 denote the sample variance of the outcomes in the baseline condition and S_{r1}^2 denote the sample variance of the outcomes in the treatment condition.

Several of the effect size estimators described in this section can be viewed as special cases of the log-response ratio, a well-known effect size used for meta-analysis in ecology and other disciplines (Hedges, Gurevitch, & Curtis, 1999). For sample data collected using measurement procedure $r = E, C, M$, or P, a simple moment estimator for the log-response ratio is given by

$$R^r = \ln\left(\hat{y}_1^r\right) - \ln\left(\hat{y}_0^r\right), \tag{8}$$

with variance estimator

$$V_R^r = \frac{S_{r0}^2}{n_0\left(\hat{y}_0^r\right)^2} + \frac{S_{r1}^2}{n_1\left(\hat{y}_1^r\right)^2}. \tag{9}$$

Hedges et al. (1999) studied the distribution of this moment estimator for the log response ratio under the assumption that the raw data are normally distributed. However, the exact distribution theory and approximations that they reported are not applicable in the present context because the direct observation data under consideration are not normally distributed. Furthermore, in some single-case studies, the within-phase sample sizes n_0 and n_1 can be quite small. I therefore propose an alternative estimator, based on a second-degree Taylor series approximation to the bias of the basic plug-in estimator:

$$R_2^r = \ln\left(\hat{y}_1^r\right) + \frac{S_{r0}^2}{2n_0\left(\hat{y}_0^r\right)^2} - \ln\left(\hat{y}_0^r\right) - \frac{S_{r1}^2}{2n_1\left(\hat{y}_1^r\right)^2}. \tag{10}$$

Based on simulations across a variety of event duration and interim time distributions, the bias-corrected estimators are nearly unbiased and have comparable mean-squared error to the simple moment estimators.[9] Use of the bias-corrected form given in Equation (10) is therefore recommended, particularly when the design contains only a few measurements

[9] Further details about the simulation design and findings can be found in the supplementary materials for this article.

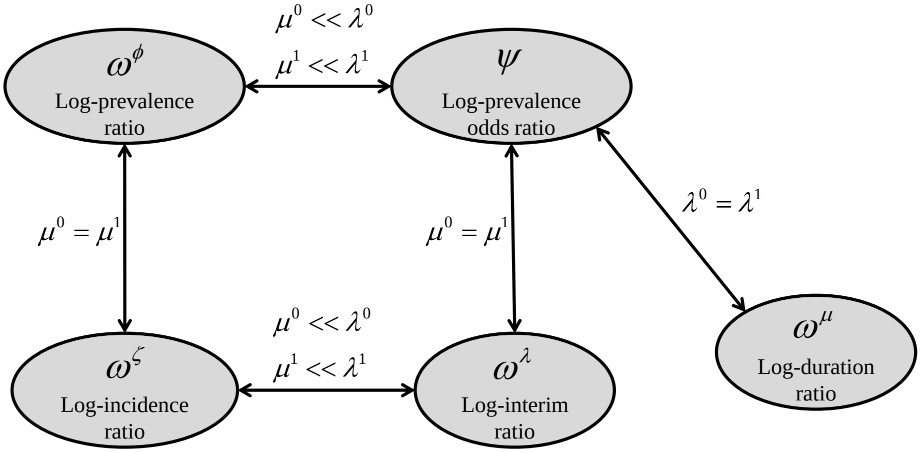

Figure 2. Effect sizes for quantifying change in behavior

in each treatment condition. The variance of R_2^r can also be estimated using V_R^r from Equation (9).

The remainder of this section covers estimation methods for each of the four main measurement procedures. Specifically, I discuss estimators for the log-incidence ratio based on event counting data, for the log-prevalence ratio and log-prevalence odds ratio based on continuous recording data and momentary time sampling data, and for the log-prevalence odds ratio based on partial interval recording data. Empirical applications of the proposed estimators are presented in later sections.

Log-incidence ratio estimators based on event counting

Event counting data measures incidence directly. Assuming that session length is held constant over the duration of the study, the log-incidence ratio can therefore be estimated using the bias-corrected response ratio estimator given in Equation (10), with $r = E$ and constants $k_0^E = 1/(2n_0)$ and $k_1^E = 1/(2n_1)$ to account for the possibility of sample means equal to zero. An approximate $(1 - \alpha)$ confidence interval for ω^{ζ} can be constructed as $R_2^E \pm z_{\alpha/2}\sqrt{V_R^E}$, where $z_{\alpha/2}$ is the $(1 - \alpha/2)$ quantile from a standard normal distribution and V_R^E is calculated using Equation (9) with $r = E$.[10] For ease of interpretation, researchers may find it easier to report percentage changes in incidence, rather than the log-incidence ratio. An approximate $(1 - \alpha)$ confidence interval for the percentage change in incidence can be calculated by transforming the confidence interval for ω^{ζ}, using $\left[\exp\left(R_2^E \pm z_{\alpha/2}\sqrt{V_R^E}\right) - 1\right] \times 100\%$.

Log-prevalence ratio estimators based on continuous recording or momentary time sampling

Log-prevalence ratio estimators based on continuous recording or momentary time sampling can be calculated using Equation (10) with $r = C$ or $r = M$. For continuous recording data, one can use $k_0^C = 1/(2Ln_0)$ and $k_1^C = 1/(2Ln_1)$ as constants for adjusting sample mean outcomes of zero; for momentary time sampling data, one can use $k_0^M = 1/(2Kn_0)$ and $k_1^M = 1/(2Kn_1)$, where K is the number of intervals per observation session. Because both types of data produce direct measurements of behavioral prevalence, these estimators are approximately unbiased for the log-prevalence ratio. A variance estimator is given by Equation (9) with $r = C$ or M, accordingly. Just as with event counting data, researchers may find it easier to report percentage changes in prevalence rather than the log-prevalence ratio. An approximate $(1 - \alpha)$ confidence interval for the percentage change in prevalence can be calculated as $100\%\left[\exp\left(R_2^r \pm z_{\alpha/2}\sqrt{V_R^r}\right) - 1\right]$ for $r = C$ or M.

Log-prevalence odds ratio estimators based on continuous recording or momentary time sampling

The log-prevalence odds ratio ψ offers an alternative metric for measuring differences or changes in prevalence, and can be estimated naturally from continuous recording data or momentary time sampling data. To account for the possibility of sample average prevalences at the ceiling of 100% or

[10] The proposed confidence interval is based on a large-sample approximation, and therefore requires an adequate sample size in each treatment condition. Based on simulation results, an approximate 95% confidence interval has actual coverage rates of better than 92% when $n = 8$ and 93% when $n = 12$. The approximate confidence intervals for other effect size parameters, described below, are based on similar large-sample approximations and have similar coverage rates. Further details can be found in the supplementary materials.

the floor of 0%, a truncated mean is used; let

$$\tilde{y}_0^r = \begin{cases} k_0^r & \bar{y}_0^r = 0 \\ \bar{y}_0^r & 0 < \bar{y}_0^r < 1 \\ 1 - k_0^r & \bar{y}_0^r = 1 \end{cases} \qquad \tilde{y}_1^r = \begin{cases} k_1^r & \bar{y}_1^r = 0 \\ \bar{y}_1^r & 0 < \bar{y}_1^r < 1 \\ 1 - k_1^r & \bar{y}_1^r = 1 \end{cases}$$

for $r = C$ or M. The constants for correcting sample means of zero or one are identical to those used with the log-prevalence ratio: for continuous recording $k_t^C = 1/(2Ln_t)$ and for momentary time sampling, $k_t^M = 1/(2Kn_t)$. Under the assumptions of the stable phase model, a bias-corrected estimator for ψ is then given by

$$\begin{aligned} \hat{\psi}_2^r = \ln(\tilde{y}_1^r) - \ln(1 - \tilde{y}_1^r) - \frac{S_{r1}^2(2\tilde{y}_1^r - 1)}{2n_1\left(\tilde{y}_1^r\right)^2\left(1 - \tilde{y}_1^r\right)^2} \\ - \ln\left(\tilde{y}_0^r\right) + \ln\left(1 - \tilde{y}_0^r\right) + \frac{S_{r0}^2(2\tilde{y}_0^r - 1)}{2n_0\left(\tilde{y}_0^r\right)^2\left(1 - \tilde{y}_0^r\right)^2} \end{aligned} \qquad (11)$$

for $r = C$ or M. Based on simulations reported in the supplementary materials, the bias-corrected estimators are approximately unbiased for the log-prevalence odds and have comparable mean-squared error to a simple moment estimator. An estimator for the approximate variance of the effect size estimate is given by

$$V_\psi^r = \frac{S_{r0}^2}{n_0\left(\tilde{y}_0^r\right)^2\left(1 - \tilde{y}_0^r\right)^2} + \frac{S_{r1}^2}{n_1\left(\tilde{y}_1^r\right)^2\left(1 - \tilde{y}_1^r\right)^2}, \qquad (12)$$

for $r = C$ or M. An approximate $(1 - \alpha)$ confidence interval for the log-prevalence odds ratio can be calculated as $\hat{\psi}_2^r \pm z_{\alpha/2}\sqrt{V_\psi^r}$ for $r = C$ or M.

Log-prevalence odds ratio estimator based on partial interval recording

As discussed previously, partial interval recording data measures neither prevalence nor incidence. One consequence of this is that the response ratio based on partial interval recording data does not estimate any easily interpretable parameter. Still, partial interval recording data can provide some information about interpretable effect size parameters under certain assumptions about the behavior being measured. For instance, S. A. Altmann and Wagner (1970) proposed analyzing partial interval recording data under the assumptions that event durations are negligible and interim times follow an exponential distribution. These assumptions lead to a point estimator for the average interim time, though one that is quite sensitive to violations of the assumptions (Fienberg, 1972). Pustejovsky (2013) described several different analytic approaches for partial interval recording data, based on varying assumptions about the average event durations (μ_0, μ_1) and average interim times (λ_0, λ_1). These approaches lead to bounds (rather than point estimators) for meaningful effect size parameters, yet point estimates will be needed in order to apply conventional meta-analytic methods. Therefore, I now describe a set of assumptions that lead to a point estimator for the log-prevalence odds ratio based on partial interval recording data.

Assume that the average event duration is known a priori and that it is not affected by the intervention, so that $\mu_0 = \mu_1 = \mu^*$ for known constant μ^*.[11] Further suppose that the interim times in each treatment condition follow exponential distributions, so that $G\left(t; \lambda_p\right) = 1 - \exp\left(-t/\lambda_p\right)$ for $p = 0, 1$. It follows that the expected value of the sample mean outcome from phase p is given by

$$\mathrm{E}\left(\bar{y}_p^P\right) = 1 - \frac{\lambda_p \exp\left(-c/\lambda_p\right)}{\mu^* + \lambda_p}. \qquad (13)$$

A closed-form algebraic expression for λ_p in terms of $\bar{y}_p^P$ does not exist (Corless, Gonnet, Hare, Knuth, & Jeffrey, 1996), but a moment estimator for λ_p can nonetheless be obtained by replacing the expected value on left-hand side of Equation (13) with the sample mean and then solving numerically for λ_p. Define $\tilde{\lambda}_0$ and $\tilde{\lambda}_1$ as the solutions to the equations

$$\tilde{y}_0^P = 1 - \frac{\tilde{\lambda}_0 \exp\left(-c/\tilde{\lambda}_0\right)}{\mu^* + \tilde{\lambda}_0} \qquad \tilde{y}_1^P = 1 - \frac{\tilde{\lambda}_1 \exp\left(-c/\tilde{\lambda}_1\right)}{\mu^* + \tilde{\lambda}_1}.$$

Under the stated assumptions, a moment estimator for ψ is given by

$$\hat{\psi}^P = \log\tilde{\lambda}_0 - \log\tilde{\lambda}_1. \qquad (14)$$

A bias-corrected estimator is available but its calculation is quite cumbersome; I therefore forgo reporting it. An estimator for the approximate variance of $\hat{\psi}^P$ is given by

$$\begin{aligned} V_\psi^P = \frac{S_{P0}^2\tilde{\lambda}_0^2(\mu^* + \tilde{\lambda}_0)^2}{n_0(1 - \tilde{y}_0^P)^2[\mu^*\tilde{\lambda}_0 + c(\mu^* + \tilde{\lambda}_0)]^2} \\ + \frac{S_{P1}^2\tilde{\lambda}_1^2(\mu^* + \tilde{\lambda}_1)^2}{n_1(1 - \tilde{y}_1^P)^2[\mu^*\tilde{\lambda}_1 + c(\mu^* + \tilde{\lambda}_1)]^2}. \end{aligned} \qquad (15)$$

An approximate $(1 - \alpha)$ confidence interval for the log-prevalence odds ratio can be calculated as $\hat{\psi}^P \pm z_{\alpha/2}\sqrt{V_\psi^P}$.

The estimator $\hat{\psi}^P$ has a limited sensitivity to the assumed average event duration, as can be seen by evaluating it at extreme values of μ^*. If μ^* is taken to be equal to zero, then $\tilde{\lambda}_p = -c/\log\left(1 - \tilde{y}_p^P\right)$. The log-prevalence odds ratio estimator will therefore be equal to the complementary log-log ratio

$$\log\left[-\log\left(1 - \tilde{y}_1^P\right)\right] - \log\left[-\log\left(1 - \tilde{y}_0^P\right)\right], \qquad (16)$$

[11] Note that under the assumption that the average event duration does not change, the log-prevalence odds ratio is equivalent to the log-incidence ratio.

which is equivalent to the estimator proposed by S. A. Altmann and Wagner (1970). At the other extreme, for large values of μ^*, the estimator will be approximately equal to the log-odds ratio evaluated using the sample PIR proportions. In other words, $\hat{\psi}^P$ approaches

$$\ln\left(\tilde{y}_1^P\right) - \ln\left(1 - \tilde{y}_1^P\right) - \ln\left(\tilde{y}_0^P\right) + \ln\left(1 - \tilde{y}_0^P\right) \qquad (17)$$

as μ^* increases. For intermediate values of μ^*, $\hat{\psi}^P$ will always lie in between (16) and (17). Consequently, the proposed estimator will often be fairly insensitive to the exact value of μ^*, and sensitivity analysis can be conducted by comparing results based on an assumed value of $\mu^* = 0$ to those based on assuming a very large value for μ^* (e.g., $\mu^* = 1000$ s).

Application: Romaniuk et al. (2002)

This section demonstrates the use of the proposed effect size estimators for analyzing multiple cases from a single study. The results based on measurement-comparable effect sizes are compared to those based on two of the most common effect size measures in single-case research: the within-case standardized mean difference (SMD, Busk & Serlin, 1992) and the percentage of non-overlapping data (PND, Scruggs et al., 1987). The latter measure is calculated as the proportion of outcome measurements in the treatment condition that are lower than the minimum outcome measurement during the baseline condition.

Romaniuk et al. (2002) studied the effects of providing choice between activities on the problem behavior of children with disabilities. A key research question was whether the treatment was differentially effective for children whose problem behavior was maintained by escape versus by attention. Prior assessment identified three cases with escape-maintained problem behavior and three cases with attention-maintained problem behavior. For five cases, the investigators measured children's problem behavior using continuous recording with observation sessions $L = 5$ minutes in length. For a sixth case (called "Riley"), the investigators measured behavior using event counting because the child displayed behaviors that had very brief duration. For all six cases, treatment reversal designs with either three or five reversals were used to assess the effect of providing choice between activities (versus a no-choice condition). Table 2 reports the outcome measurement procedure used for each case along with summary statistics by treatment condition.

In order to synthesize all six cases, an effect size is needed that permits comparisons between the five cases measured using continuous recording and the sixth case measured using event counting; the ARP model provides a basis for establishing that comparability. I assume that the average duration of Riley's problem behaviors was unaffected by the intervention ($\mu_0 = \mu_1$), so that a log-incidence ratio estimated from event counting data is on a comparable scale to that of a log-prevalence ratio estimated from continuous recording. The assumption seems reasonable because Riley's problem behaviors were all very brief, discrete instances that would not be meaningfully altered by intervention.

Table 3 reports estimated log-prevalence ratios R_2^C and standard errors $\sqrt{V_R^C}$, calculated according to Equations (10) and (9), respectively, for the five cases measured using continuous recording. The table also reports Riley's estimated log-incidence ratio R_2^E and corresponding standard error.[12] For the three cases with escape-maintained behavior, the estimated log-prevalence ratios are large and negative, ranging from -0.96 to -2.39; for these cases, providing choice-making opportunites greatly reduces the prevalence of problem behaviors. A fixed-effects meta-analysis of the three cases provides a succinct summary of the average effect of the treatment on these three cases and is reported in the penultimate row of the table.[13] The average log-prevalence ratio for cases with escape-maintained behavior is -1.22, with an approximate 95% CI of [-1.48, -0.95]; the CI corresponds to a reduction of between 61% and 77% in the prevalence of problem behaviors for the children with escape-maintained behavior. In contrast, the estimated log-prevalence ratios for the three cases with attention-maintained behavior are moderately positive, ranging from 0.12 to 0.31. Based on a fixed-effects meta-analysis (reported in the final row of the table), the average log-prevalence ratio for these cases is 0.23, with an approximate 95% CI of [0.11,0.34] corresponding to increases in the prevalence of problem behavior of between 13% and 40%.

An alternative effect size metric that could be applied in this example is the log-prevalence odds ratio. Table 3 also reports log-prevalence odds ratios estimates $\hat{\psi}_2^C$ and associated standard errors $\sqrt{V_\psi^C}$ for the five cases measured using continuous recording. For the case measured using event counting, the estimated log-incidence ratio is used, which is equivalent to the log-prevalence odds ratio under the assumptions that the treatment does not alter average event duration ($\mu_0 = \mu_1$) and that the average event durations are very short. Again, this assumption seems reasonable given the character of the Riley's behaviors; in fact, the short duration of her behaviors was the motivation for using event counting rather than continuous recording (Romaniuk et al., 2002, p. 351).

For four of the five cases measured using continuous recording, the relative magnitudes of the log-prevalence odds ratio estimates are comparable to the log-prevalence estimates. The exception is Christy, whose log-prevalence odds ratio is much larger than the other two cases with attention-

[12] All of the effect size estimates and standard errors can be calculated from the summary statistics reported in Table 2.

[13] A fixed-effects meta-analysis is used due to the small number of cases. I used the `metafor` package in R (Viechtbauer, 2010) for calculation.

Table 2
Summary statistics for Romaniuk et al. (2002)

		No Choice			Choice		
Case	Outcome measure	$\bar{y}_0$	S_0^2	n_0	$\bar{y}_1$	S_1^2	n_1
Brooke	C	0.70	0.079	14	0.06	0.005	11
Gary	C	0.69	0.048	16	0.26	0.048	16
Maggie	C	0.64	0.053	16	0.21	0.010	8
Christy	C	0.71	0.035	15	0.89	0.011	10
Rick	C	0.63	0.064	15	0.71	0.037	14
Riley	E	76.2	794.9	21	103.7	547.2	12

Table 3
Effect size estimates for Romaniuk et al. (2002)

Case	Outcome measure	Behavior function	R_2	$\sqrt{V_R}$	$\hat{\psi}_2$	$\sqrt{V_\psi}$	SMD	$\sqrt{V_{SMD}}$	PND
Brooke	C	escape	-2.39	0.37	-3.50	0.52	-2.95	0.59	1.00
Gary	C	escape	-0.96	0.23	-1.81	0.38	-1.95	0.43	0.50
Maggie	C	escape	-1.09	0.19	-1.86	0.32	-2.16	0.54	0.38
Christy	C	attention	0.22	0.08	1.13	0.40	1.12	0.44	0.00
Rick	C	attention	0.12	0.13	0.37	0.38	0.36	0.37	0.00
Riley	E	attention	0.31	0.10	0.31	0.10	1.03	0.38	0.00
FE meta-analysis		escape	-1.22	0.13	-2.14	0.22	-2.26	0.29	
		attention	0.23	0.06	0.36	0.10	0.81	0.23	

maintained problem behavior, even while her log-prevalence ratio is intermediate between those of the other two cases. This discrepancy is due to the fact that the two effect size metrics are less comparable for prevalences greater than 0.5. Christy's average prevalence is fairly high in both the no-choice ($\bar{y}_0 = 71\%$) and the choice conditions ($\bar{y}_1 = 89\%$), leading to a divergence between the magnitude of the two metrics.

In this example, the log-prevalence odds ratio and the log-prevalence ratio lead to much the same substantive conclusions. Table 3 also reports the SMD effect size and corresponding standard error (the latter is labeled $\sqrt{V_{SMD}}$). For cases with escape-maintained problem behavior, a fixed-effects meta-analysis based on the d effect sizes leads to an average effect estimate of -2.26 (approximate 95% CI: [-2.84, -1.68]); the average effect size for cases with attention-maintained behavior is 0.81 (approximate 95% CI: [0.36,1.25]). At first glance, the magnitude of the SMDs appears remarkably similar to that of the log-prevalence odds ratios ($\hat{\psi}_2$) for all cases but Riley. However, the interpretation is different and more problematic because continuous recording and event counting do not produce interval-scale outcome measurements. To illustrate the problem, consider the hypothetical scenario in which the treatment effect is the same for all participants and is equal to the lower CI bound for escape-maintained problem behavior, that is, $SMD = -2.84$. Based on this information, the level of the outcome during treatment would be predicted as $\bar{y}_0 - 2.84 \times S_0$. For four of the six cases in this example, the resulting prediction is a negative number, yet it is impossible to have a negative proportion of time or a negative number of events. The log-prevalence ratio and log-prevalence odds ratios do not lead to such impossible predictions, because both measure change in proportionate terms.

The final column of Table 3 reports the PND for each case; two problems with this effect size are apparent. First, PND is zero for all three cases with attention-maintained behavior, concealing the fact that the treatment appears to be actively harmful (rather than simply ineffective) and hiding any variation in the magnitude of effects for these cases. Second, there is no way to assess the sampling uncertainty of the PND statistic, which prevents the use of fixed-effects meta-analysis and other conventional techniques for synthesizing multiple effect sizes. In contrast, the log-prevalence ratios and log-prevalence odds ratios can be used in conventional meta-analysis procedures and remain meaningful measures of treatment effect magnitude even when effects are not beneficial.

Application: Shogren, Faggella-Luby, Bae, & Wehmeyer (2004)

The example presented in the previous section is based on one study included in the systematic review by Shogren et al. (2004). This section presents a complete meta-analysis of all studies included in the review. In the original systematic review, the authors used PND and other non-overlap measures as effect size metrics; they used non-parametric tests to examined moderators of the effect sizes (Shogren et al., 2004, Tables 3 and 4). These same data were also analyzed by Van den Noortgate and Onghena (2008) using SMD effect sizes and multi-level meta-analytic models.

Recall from Table 1 that the studies used a variety of measurement procedures, including event counting, continuous recording, momentary time sampling, and partial interval recording.[14] Effect sizes based on each procedure involve distinct assumptions, which I detail separately. I exclude from the analysis the three studies (including 4 cases) that used idiosyncratic measurement procedures (Bambara, Koger, Katzer, & Davenport, 1995; Cole & Levinson, 2002; Dibley & Lim, 1999), from which measurement-comparable effect sizes cannot be derived. I also excluded one study with a single case (Peterson, Caniglia, & Royster, 2001) that reported a functional assessment but did not use an evaluation design. The remaining 9 studies, including a total of 27 cases, are used in the meta-analysis; the number of cases per study ranges from 1 to 6. Across studies and cases, the number of observations per condition ranged from 3 to 21, with a median of 11.

Effect size assumptions

I use the log-prevalence odds ratio as the effect size metric because the majority of included studies used interval recording methods. For cases measured using continuous recording and momentary time sampling, effect size estimates were calculated using $\hat{\psi}_2^C$ and $\hat{\psi}_2^M$, respectively, as given in Equation (11). For cases measured using event counting, I assume that the intervention does not alter the mean event duration ($\mu^0 = \mu^1$) and that the mean event duration is close to zero, so that proportionate changes in incidence are approximately equal to proportionate changes in interim times. Under these assumptions, effect sizes for cases measured by event counting are in the same metric as the log-prevalence odds ratio, and are estimated using R_2^E as given in Equation (10). For cases measured using interval recording, I assume that the average event duration in each phase is equal to $\mu^* = 10$ s and that interim times are exponentially distributed. Together, these assumptions imply that the log-prevalence odds ratio can be estimated using $\hat{\psi}^P$ as given in Equation (14).

Meta-analytic model

Having described the underlying assumptions for the estimated effect sizes, I now describe the modeling assumptions I use to meta-analyze those effect sizes. Following Van den Noortgate and Onghena (2008), I adopt a multilevel random-effects model in which case-level effect sizes from the same study are allowed to be correlated. Let Y_{ij} denote the effect size estimate for case j from study i, with estimated sampling variance V_{ij}. I consider two multi-level models for these effect sizes. Model 1 contains no predictors:

$$Y_{ij} = \beta + u_i + v_{ij} + \epsilon_{ij}, \tag{18}$$

where β is the grand-average effect size across cases and studies, u_i is a study-level deviation from the grand-average effect size, v_{ij} is a case-level deviation from the study-level average effect size for study i, and ϵ_{ij} is the sampling error of the estimated effect size for case j in study i. I assume that all errors are mutually independent and normally distributed with means of zero and variances $\text{Var}(u_i) = \tau^2$, $\text{Var}\left(v_{ij}\right) = \sigma^2$, and $\text{Var}\left(\epsilon_{ij}\right) = V_{ij}$. Here, τ^2 captures between-study variation in average effect sizes while σ^2 captures variation in effect sizes across cases from the same study, with the latter source of variation treated as constant across studies. For the effect size estimators under consideration, the assumption that ϵ_{ij} is normally distributed will hold only approximately, given sufficiently large sample sizes for each case and treatment condition.[15] Reliance on such approximations is conventional in random effect meta-analysis of other types of effect sizes (Hedges & Vevea, 1998).

Again following Van den Noortgate and Onghena (2008), the second model adds two moderators of effect size: an indicator for the case's gender and an indicator for the type of choice-making used in the study (choice between alternative activities versus choice of task order).[16] Model 2 is then:

$$Y_{ij} = \beta_0 + \beta_1 \left(Male\right)_{ij} + \beta_3 \left(TaskOrder\right)_{ij} + u_i + v_{ij} + \epsilon_{ij}, \tag{19}$$

with the error terms defined as in the first model. I estimate the variance components τ^2 and σ^2 and the β-coefficients via restricted maximum likelihood.

Finally, it is useful to compare the results based on the log-prevalence odds ratio to other alternative effect size estimators. I therefore replicate the approach of Van den Noortgate and Onghena (2008) by estimating Model 2 using within-case SMD effect sizes. It is not possible to estimate a multi-level meta-analytic model on PND effect sizes, due to the

[14] For one case from a study by Kern, Mantegna, Vorndran, Bailin, and Hilt (2001), a whole interval recording procedure was used to measure task engagement. For purposes of calculating effect sizes, I re-coded the data as a partial interval recording measure of task dis-engagement.

[15] Based on simulation results reported in the supplementary materials, 95% confidence intervals have close to nominal coverage when based on samples with at least 8 observations per condition; this may be taken as indirect evidence that the sampling distribution of the point estimators is very approximately normal. Of the 27 cases included in the meta-analysis, 8 cases have at least one condition with less than eight observations. Despite the small sample sizes, I choose to retain these cases in the meta-analysis because it seems more important to synthesize all relevant data than to select studies to satisfy the technical criteria of approximate normality.

[16] More extensive description of these characteristics can be found in Shogren et al. (2004).

lack of valid sampling variance estimates. Instead, I estimate a simple linear regression model with the same moderators as Model 2; I report robust standard errors, clustered at the study level (Hedges, Tipton, & Johnson, 2010). Because an increased PND corresponds to a decrease in behavior, I reverse the signs of the estimated coefficients in order to maintain comparability with the other effect size metrics.

Results

Figure 3 displays a forest plot of the estimated effect sizes. Based on Model 1, the average log-prevalence odds ratio is estimated as $\hat{\beta} = -1.51$ (95% CI: [-2.11, -0.91]). Odds ratios can be difficult to interpret; as an aid to interpretation, it is helpful to provide translations into proportionate reductions in prevalence at benchmark levels of baseline prevalence.[17] For a baseline prevalence of $\phi^0 = 0.30$, the CI for the average log-prevalence odds ratio corresponds to a reduction of between 51% and 83%; for a baseline prevalence of $\phi^0 = 0.50$, the corresponding reduction is between 43% and 78%. The estimated between-study variance is small ($\hat{\tau}^2 = 0.06$) while the estimated within-study variance is fairly large ($\hat{\sigma}^2 = 1.31$); total heterogeneity, including both between- and within-study variation, is therefore large. For an average log-prevalence odds ratio of -1.51, total heterogeneity of $\tau^2 + \sigma^2 = 1.37$ implies that a quarter of the population has log-prevalence odds ratios of less than -2.30 (a reduction of more than 81% for baseline prevalence $\phi^0 = 0.50$) while another quarter of the population has log-prevalence odds ratios of more than -0.72 (a reduction of less than 35% for baseline prevalence $\phi^0 = 0.50$).

Table 4 reports the results of fitting Model 2 to the estimated log-prevalence odds ratios, as well as the results of comparable models for the SMD and PND. Based on the log-prevalence odds ratio, the treatment is estimated to be more effective (i.e., to produce greater reductions in problem behavior, $\hat{\beta}_1 = -0.21$, 95% CI: [-1.45,1.03]) for male children than for females; however, the difference is estimated very imprecisely and is far from statistically significant. The treatment is estimated to be more effective when participants are allowed to choose the order of tasks versus having a choice between alternative activities ($\hat{\beta}_2 = -0.50$, 95% CI: [-1.74,0.75]), though the difference is also far from statistically significant.

The statistical non-significance of the moderators is consistent with the models based on SMD and PND effect sizes. However, there is some inconsistency regarding the direction of the effects. In the models based on SMD and PND effect sizes, the treatment is estimated to be *less* efficacious for males than for females. Though this inconsistency may be due in part to sampling variation, it nonetheless illustrates how the measurement-comparable effect sizes proposed in this paper may lead to different inferences than other, more widely used effect sizes.

A useful approach to checking the soundness of these meta-analytic models is to calculate the predicted level of the outcome under the treatment condition, based on the baseline scores and the fitted effect size values from Model 2. In the model based on SMD effect sizes, the predicted level of the outcome under treatment would be calculated as $\bar{y}_0 + S\hat{M}D \times S_0$, where $S\hat{M}D$ is a fitted value from Model 2. For 9 out of the 27 cases, the predicted level of the outcome is negative. For outcomes that are calculated as proportions or counts, such predictions are not sensible, and suggest that the SMD is an inappropriate metric for the types of outcome measurements used in these studies. In contrast, similar calculations based on the log-prevalence odds ratio metric produce predictions that remain within the range of the scale.

Discussion

I have presented a model for synthesizing case-level estimates of the effect of providing choice-making opportunities on the prevalence odds of individuals' problem behavior, using data from studies identified by Shogren et al. (2004). In this application, an advantage of using the effect sizes that I have proposed is that the meta-analytic results are interpretable in terms of clear behavioral constructs. For example, the average treatment effect estimate can be translated into a percentage reduction in the prevalence of problem behavior. In contrast, it is difficult to interpret an average of PND or SMD effect sizes because their magnitudes may depend on the measurement procedure used.

A major feature of these data is the large number of cases measured using partial interval recording, which is not a direct measure of prevalence. Strong modeling assumptions are necessary to justify the proposed measurement-comparable effect size estimates based on interval recording data. The results should be interpreted in light of these assumptions, and with considerable caution.

Further analyses would be possible if a larger set of studies could be identified that included a greater variety of measurement methods. Among those identified by Shogren et al. (2004), only one third of the cases used a measurement method other than interval recording. As a result, there is insufficient data to examine whether there are differences in average effect sizes for cases measured by different methods. However, it may be possible to carry out such an analysis in other applications, and meta-analysts are encouraged to do so. Although the goal of using measurement-comparable effect sizes is to reduce irrelevant operational heterogeneity and put different measurement procedures on a comparable

[17] For baseline prevalence ϕ^0 and log-prevalence odds ratio ψ, the proportionate reduction in prevalence is given by

$$\exp\left(\omega^{\phi}\right) - 1 = \frac{\exp(\psi)}{1 - \phi^0\left[1 - \exp(\psi)\right]} - 1.$$

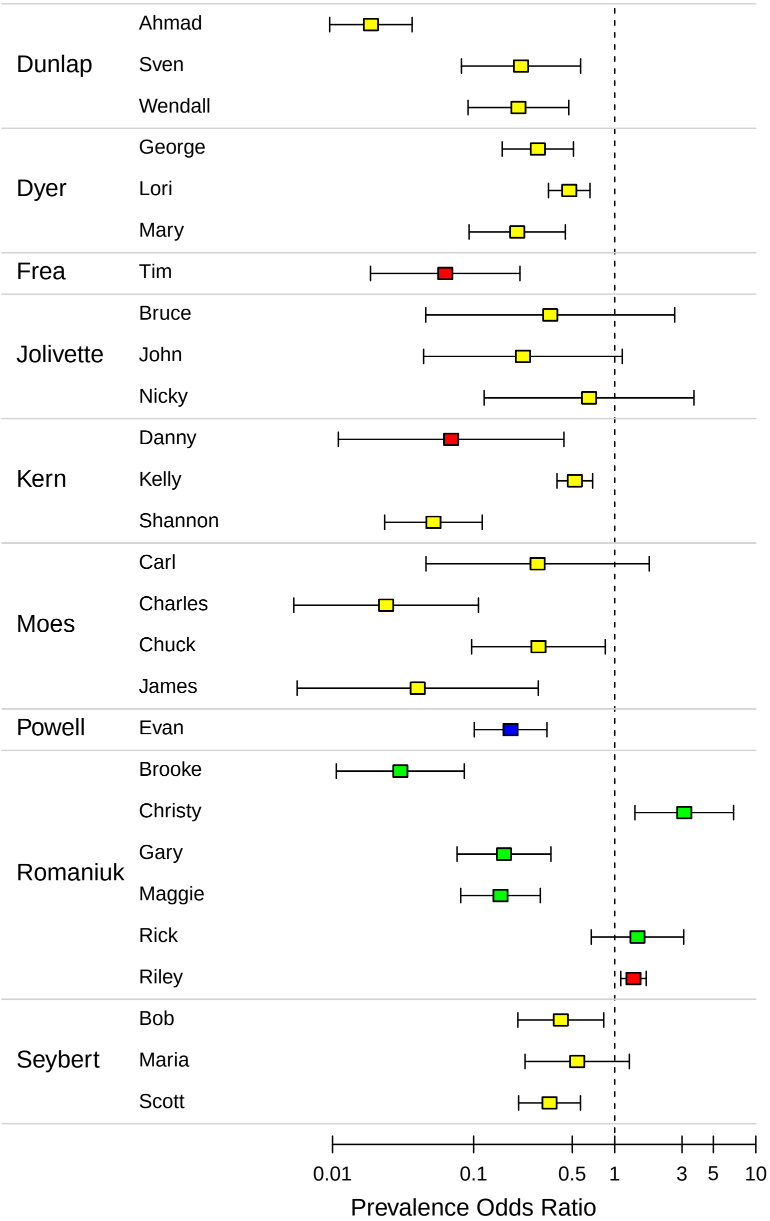

Figure 3. Forest plot of estimated prevalence odds ratios for studies from Shogren et al. (2004). Whisker bars represent approximate 95% confidence intervals for the case-level effect sizes. The color of the symbol corresponds to the measurement method used. Green = continuous recording; blue = momentary time sampling; red = event counting; yellow = interval recording.

basis, it is prudent to test whether the exercise has succeeded. Residual differences between measurement methods may indicate violation of modeling assumptions, which can in turn lead the analyst towards more refined assumptions or signal caution in the interpretation of effect sizes averaged across measurement procedures.

General discussion

I have presented an alternating renewal process model for free-operant behavior that can be used to describe data collected via several common measurement procedures. I used the model to define measurement-comparable effect size metrics and proposed estimators that are applicable under a simple between-session model. Several of the effect

Table 4
Random-effects meta-analysis of Shogren et al. (2004) data

	Log-POR		SMD		PND	
Parameter	Est.	SE	Est.	SE	Est.	SE
β-coefficients						
Intercept (β_0)	-1.19	0.46	-1.57	0.56	-0.56	0.10
Male (β_1)	-0.21	0.54	0.05	0.58	0.26	0.13
Task order (β_2)	-0.50	0.54	-0.26	0.67	-0.20	0.21
Variance components						
Between-study (τ^2)	0.00		0.32			
Within-study (σ^2)	1.42		1.30			

sizes are closely related to the log-response ratio, a well-known effect size used in other areas of meta-analysis. Data based on partial interval recording procedures present special difficulties because they cannot be interpreted as direct measurements of either prevalence or incidence. I have proposed one approach to estimating measurement-comparable effect sizes based on partial interval recording data, involving rather strong modeling assumptions that may be difficult to verify in practice. Elsewhere I have suggested other analytic methods that are based on somewhat weaker assumptions (Pustejovsky, 2013), but these approaches yield bounds instead of point estimates, and are therefore more difficult to meta-analyze.

A working meta-analyst, interested in synthesizing evidence from single-case studies of free-operant behavior, might sensibly question the need for such an elaborate model to define effect sizes. I see three advantages to this model. First, a model that captures the essential features of the outcome measurement procedures improves the interpretability of effect sizes defined with respect to it. The effect sizes that I have proposed are defined in terms of prevalence and incidence, both readily understood aspects of a behavior. In contrast, other effect size proposals such as the within-case SMD and the PND do not correspond closely with clear behavioral constructs.

Second, a measurement-comparability model is all the more necessary when dealing with measurements that are difficult to interpret. Given that interval recording procedures are widely used for measurement of free-operant behavior, meta-analytic methods for single-case research cannot simply ignore them. In the Shogren et al. (2004) application, interval recording was used with two thirds of the cases to be synthesized. Taking a naive approach by treating interval recording data just as other data would compromise the construct validity of the synthesis. On the other hand, restricting the meta-analysis to cases measured using procedures other than interval recording would drastically reduce the sample size and possibly compromise external validity. Methods are therefore needed that retain cases measured using interval recording while also making use of interpretable, measurement-comparable effect sizes.

Third, use of a measurement-comparability model has prospective implications for research practice. Effect sizes defined under such a model allow meta-analysts to formulate research questions in more precise terms, such as whether an intervention affects the prevalence of a behavior, the incidence of a behavior, or both. As previously noted, it will rarely be possible to estimate theoretically interesting effect sizes such as the log-duration ratio and log-interim ratio based only on data collected from published graphs. However, such limitations do not pertain to primary researchers planning future studies; data collection procedures and reporting practices could certainly be adjusted so that effects on event duration and interim time could be separately estimated.

The most widely used effect size metrics for meta-analysis of single-case designs, including the within-case SMD and the PND, do not adequately account for the comparability of effect sizes across studies that use different measurement procedures, yet they suffer from similar drawbacks as the measurement-comparable effect sizes that I have proposed. The within-case SMD is premised on between-session modeling assumptions that are largely parallel to those on which I have relied, including lack of auto-correlation and means that are stable within phases (Busk & Serlin, 1992). Similarly, the PND is known to be inappropriate when there are within-phase trends, and the potential effect of auto-correlation on PND is unknown (Shadish & Rindskopf, 2007). Given these common limitations, together with the advantages of measurement-comparable effect sizes that I have noted, I recommend that meta-analysts consider using the proposed measurement-comparable effect sizes for synthesizing single-case studies of free-operant behavior, rather than the more widely used SMD or PND effect sizes.[18]

[18] However, approaches other than the SMD and PND may offer some advantage over the methods proposed in this paper. Several recently proposed effect size metrics and analytic methods can handle data that display within-phase trends and auto-correlation, but these approaches have yet to be widely applied (e.g., Maggin, Swaminathan, et al., 2011; Parker, Vannest, Davis, & Sauber, 2011;

Choosing a measurement-comparable effect size

Several factors are relevant when choosing an effect size metric to use for summarizing the results of a single-case study or meta-analyzing the results of several studies. Clearly, the choice of a summary effect size will be limited to those that can be estimated from the available data. If a set of cases or studies were measured using different procedures, then it may be necessary to use several different effect sizes and to evaluate whether the those effect sizes are directly comparable. For example, if some studies measure incidence with an event counting procedure while the remainder measure prevalence with continuous recording, each subset of studies might first be summarized and meta-analyzed separately. The meta-analyst could then assess whether it is reasonable to assume that average event durations can be treated as constant for the studies that use event counting, in which case a combined meta-analysis would be warranted.

If the preponderance of included studies focus on prevalence, then the meta-analyst must further determine whether to use the log-prevalence ratio or the log-prevalence odds ratio. In this circumstance, it may be possible to choose an effect size metric by comparing the empirical fit of the meta-analytic models applied to each effect size, with preference given to the effect size metric that exhibits less heterogeneity. Engels, Schmid, Terrin, Olkin, and Lau (2000) used such an approach for meta-analysis of medical studies with binary outcomes (see also Deeks, 2002). These authors found that it was often difficult to distinguish between models based on risk ratios versus those based on odds ratios, particularly when effects were small and the number of studies was limited. However, single-case meta-analyses might have comparatively greater power for determining the more homogeneous effect size metric, due both to the availability of case-level data and the potential for very large effects on prevalence.

Limitations and Extensions

The models and methods that I have presented have several limitations that should be noted, including those related to the within-session model, between-session model, effect sizes, and meta-analytic methods. I address each of these in turn, while also noting potential areas for future research.

The equilibrium alternating renewal process model for describing observed behavior is general in the sense that it is not limited to specific parametric distributions for the event durations and interim times. Still, the assumption that the process is in equilibrium might not be entirely realistic, particularly when observation sessions coincide with the start of some other event (such as the beginning of math class) or entrance into a novel setting (such as a therapist's office). Also, the model does not account for relationships between behavioral events and environmental contingencies, which are of central interest in some behavior-analytic theories. Both of these criticisms are quite reasonable, but one must also consider that the available data (in the form of published graphs) do not provide sufficient information to model such intra-session dynamics. Without the equilibrium assumption, all of the measurement procedures that I have described become somewhat sensitive to initial conditions and to the length of observation sessions–details about which the meta-analyst will have little or no information. The equilibrium assumption implies that the process is uniform over the course of a session, which is consistent with how the data generated by various measurement procedures are typically summarized and reported.

Next, the proposed measurement-comparable effect sizes are defined under the simplest possible model for describing change in behavior over time. The stable-phase model allows for neither trends in the process over time nor for serial dependence of repeated measurements, both of which are prime concerns in quantitative single-case research methodology (Horner et al., 2012; Maggin, Swaminathan, et al., 2011; Wolery et al., 2010). Extensions to the between-session model are possible and will be explored in further research. One possible extension is to introduce auto-correlation into the model, but doing so adds several layers of complication because the measurement errors in an alternating renewal process are non-normal. Consequently, there are several different ways of formulating auto-correlation models, and further research is needed to understand which are reasonable. Moreover, practically all previous research on auto-correlation in single-case time series has been premised on the assumption that the measurements are interval scaled with normally distributed errors. If the modeling approach in this paper is reasonable, then much of the previous research on the presence of auto-correlation in single-case studies may need to be re-examined.

There are two shortcomings to the effect size metrics that I have described. The goal of some behavioral interventions is complete elimination of an undesirable behavior, and a researcher or meta-analyst may be interested in the probability that elimination will be achieved. The effect size estimators that I have proposed, which make use of truncated means, do not distinguish between complete elimination of a behavior versus reduction to very low prevalence or incidence. If such a distinction is of primary interest, specialized meta-analytic models and methods may be needed. The other short-coming is that the proposed effect sizes apply to individual cases, and are not useful for syntheses containing both single-case and between-subjects research designs. Hedges, Pustejovsky, and Shadish (2012, 2013) have proposed design-comparable effect sizes that are on the same

Swaminathan, Rogers, & Horner, 2014). Further methodological research is warranted to understand the relative strengths and weaknesses of these new proposals.

scale as effect sizes identified by between-subjects designs, but only for standardized mean differences. A similar approach could be taken for defining design-comparable effect sizes for the ARP model, and this remains a topic for future research.

A final, important limitation of this work is that it has yet to be fully vetted by applied researchers with experience measuring free-operant behavior. Researchers who are familiar with the profiles of the behaviors being measured will be in a better position to judge the plausibility and utility of the key assumptions. Also, a better understanding of how researchers choose between alternative measurement procedures would be helpful in assessing whether and under what circumstances the assumptions regarding those procedures are reasonable. This final limitation highlights the need for greater collaboration between applied single-case researchers and statistical methodologists, as Campbell and Herzinger (2010) and others have argued.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research and Therapy*, *31*(6), 621–31.

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.

*Bambara, L. M., Koger, F., Katzer, T., & Davenport, T. A. (1995). Embedding choice in the context of daily routines: An experimental case study. *The Journal of the Association for Persons with Severe Handicaps*, *20*(3), 185–195.

Beretvas, S. N., & Chung, H. (2008). A review of meta-analyses of single-subject experimental designs: Methodological issues and practice. *Evidence-Based Communication Assessment and Intervention*, *2*(3), 129–141. doi: 10.1080/17489530802446302

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 187–212). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Campbell, J. M., & Herzinger, C. V. (2010). Statistics and single subject research methodology. In D. L. Gast (Ed.), *Single subject research methodology in behavioral sciences* (pp. 417–450). New York, NY: Routledge.

Center, B. A., Skiba, R. J., & Casey, A. (1985). A methodology for the quantitative synthesis of intra-subject design research. *The Journal of Special Education*, *19*(4), 387–400.

Chambless, D. L., & Ollendick, T. H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Annual Review of Psychology*, *52*, 685–716. doi: 10.1146/annurev.psych.52.1.685

*Cole, C. L., & Levinson, T. R. (2002). Effects of within-activity choices on the challenging behavior of children with severe developmental disabilities. *Journal of Positive Behavior Interventions*, *4*(1), 29–37.

Cooper, H. M. (2009). Hypotheses and problems in research synthesis. In H. M. Cooper, L. V. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 19–35). New York, NY: Russell Sage Foundation.

Corless, R. M., Gonnet, G. H., Hare, D. E. G., Knuth, D. E., & Jeffrey, D. J. (1996). On the Lambert W function. *Advances in Computational Mathematics*, *5*, 329–359.

Cox, D. R. (1962). *Renewal Theory*. Great Britain: Metheun & Co. Ltd.

Deeks, J. J. (2002). Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine*, *21*(11), 1575–1600. doi: 10.1002/sim.1188

*Dibley, S., & Lim, L. (1999). Providing choice making opportunities within and between daily school routines. *Journal of Behavioral Education*, *9*(2), 117–132.

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

*Dyer, K., Dunlap, G., & Winterling, V. (1990). Effects of choice making on the serious problem behaviors of students with severe handicaps. *Journal of Applied Behavior Analysis*, *23*(4), 515–524.

Engels, E. A., Schmid, C. H., Terrin, N., Olkin, I., & Lau, J. (2000). Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Statistics in Medicine*, *19*, 1707–1728.

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

Fleiss, J., & Berlin, J. A. (2009). Effect sizes for dichotomous outcomes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 237–253). New

York, NY: Russell Sage Foundation.

*Frea, W. D., Arnold, C. L., & Vittimberga, G. L. (2001). A demonstration of the effects of augmentative communication on the extreme aggressive behavior of a child with autism within an integrated preschool setting. *Journal of Positive Behavior Interventions*, *3*(4), 194–98.

Gage, N. A., Lewis, T. J., & Stichter, J. P. (2012). Functional behavioral assessment-based interventions for students with or at risk for emotional and/or behavioral disorders in school: A hierarchical linear modeling meta-analysis. *Behavioral Disorders*, *37*(2), 55–77.

Gast, D. L. (2010). Applied research in education and behavioral sciences. In D. L. Gast (Ed.), *Single subject research methodology in behavioral sciences* (pp. 1–19). New York, NY: Routledge.

Gingerich, W. J. (1984). Meta-analysis of applied time-series data. *The Journal of Applied Behavioral Science*, *20*(1), 71–79. doi: 10.1177/002188638402000113

Hart, S. L., & Banda, D. R. (2009). Picture exchange communication system with individuals with developmental disabilities: A meta-analysis of single subject studies. *Remedial and Special Education*, *31*(6), 476–488. doi: 10.1177/0741932509338354

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.

Hedges, L. V. (2008). What are effect sizes and why do we need them? *Child Development Perspectives*, *2*(3), 167–171.

Hedges, L. V., Gurevitch, J., & Curtis, P. (1999). The meta-analysis of response ratios in experimental ecology. *Ecology*, *80*(4), 1150–1156.

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods*, *3*, 224–239. doi: 10.1002/jrsm.1052

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2013). A standardized mean difference effect size for multiple baseline designs across individuals. *Research Synthesis Methods*, *4*(4), 324–341. doi: 10.1002/jrsm.1086

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. doi: 10.1002/jrsm.5

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486.

Hershberger, S. L., Wallace, D. D., Green, S. B., & Marquis, J. G. (1999). Meta-analysis of single-case designs. *Statistical Strategies for Small Sample Research*, 109–132.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S. L., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, *71*(2), 165–179.

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

Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and Tactics of Behavioral Research* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

*Jolivette, K., Wehby, J. H., Canale, J., & Massey, N. G. (2001). Effects of choice-making opportunities on the behavior of students with emotional and behavioral disorders. *Behavioral Disorders*, *26*(2), 131–145.

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 Pres.

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

Kennedy, C. H. (2004). *Single-Case Designs for Educational Research*. Boston, MA: Allyn & Bacon.

*Kern, L., Mantegna, M. E., Vorndran, C. M., Bailin, D., & Hilt, A. (2001). Choice of task sequence to reduce problem behaviors. *Journal of Positive Behavior Interventions*, *3*(1), 3–10.

Kratochwill, T. R., & Stoiber, K. C. (2002). Evidence-based interventions in school psychology: Conceptual foundations of the Procedural and Coding Manual of Division 16 and the Society for the Study of School Psychology Task Force. *School Psychology Quarterly*, *17*(4), 341–389. doi: 10.1521/scpq.17.4.341.20872

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

Machalicek, W., O'Reilly, M. F., Beretvas, S. N., Sigafoos, J., Lancioni, G. E., Sorrells, A., … Rispoli, M. (2008). A review of school-based instructional interventions for students with autism spectrum disorders. *Research in Autism Spectrum Disorders*, *2*(3), 395–416. doi: 10.1016/j.rasd.2007.07.001

Maggin, D. M., O'Keeffe, B. V., & Johnson, A. H. (2011). A quantitative synthesis of methodology in the meta-analysis of single-subject research for students with disabilities: 1985-2009. *Exceptionality*, *19*(2), 109–135. doi: 10.1080/09362835.2011.565725

Maggin, D. M., Swaminathan, H., Rogers, H. J., O'Keeffe, B. V., Sugai, G., & Horner, R. H. (2011). A generalized least squares regression approach for computing effect sizes in single-case research: Application examples. *Journal of School Psychology*, *49*(3), 301–21. doi: 10.1016/j.jsp.2011.03.004

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

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., & Harris, K. R. (2005). Research in special education: Scientific methods and evidence-based practices. *Exceptional Children*, *71*(2), 137–148.

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification*, *35*(4), 303–22. doi: 10.1177/0145445511399147

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy*, *42*(2), 284–299. doi: 10.1016/j.beth.2010.08.006

*Peterson, S. M. P., Caniglia, C., & Royster, A. J. (2001). Application of choice-making intervention for a student with multiply maintained problem behavior. *Focus on Autism and Other Developmental Disabilities*, *16*(4), 240–246.

*Powell, S., & Nelson, B. (1997). Effects of choosing academic assignments on a student with attention deficit hyperactivity disorder. *Journal of Applied Behavior Analysis*, *30*(1), 181–183. doi: 10.1901/jaba.1997.30-181

Pustejovsky, J. E. (2013). *Operationally comparable effect sizes for meta-analysis of single-case research* (Doctoral dissertation, Order No. 3563832, Northwestern University). Retrieved from ProQuestDissertationsandTheses.

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

*Romaniuk, C., Miltenberger, R., Conyers, C., Jenner, N., Jurgens, M., & Ringenberg, C. (2002). The influence of activity choice on problem behaviors maintained by escape versus attention. *Journal of Applied Behavior Analysis*, *35*(4), 349–62. doi: 10.1901/jaba.2002.35-349

Ross, S. W., & Horner, R. H. (2009). Bully prevention in positive behavior support. *Journal of Applied Behavior Analysis*, *42*(4), 747–59. doi: 10.1901/jaba.2009.42-747

Schoenfeld, W. N. (1972). Problems of modern behavior theory. *Integrative Physiological and Behavioral Science*, *7*(1), 33–65.

Scruggs, T. E., & Mastropieri, M. A. (2012). PND at 25: Past, present, and future trends in summarizing single-subject research. *Remedial and Special Education*, *34*(1), 9–19. doi: 10.1177/0741932512440730

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research. *Remedial and Special Education*, *8*(2), 24–43.

*Seybert, S., Dunlap, G., & Ferro, J. (1996). The effects of choice-making on the problem behaviors of high school students with intellectual disabilities. *Journal of Behavioral Education*, *6*(1), 49–65.

Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: New applications and some agenda items for future research. *Psychological Methods*, 1–43.

Shadish, W. R., & Rindskopf, D. M. (2007). Methods for evidence-based practice: Quantitative synthesis of single-subject designs. *New Directions for Evaluation*, *113*(113), 95–109. doi: 10.1002/ev.217

Shogren, K. A., Faggella-Luby, M. N., Bae, S. J., & Wehmeyer, M. L. (2004). The effect of choice-making as an intervention for problem behavior. *Journal of Positive Behavior Interventions*, *6*(4), 228–237.

Swaminathan, H., Rogers, H. J., & Horner, R. H. (2014). An effect size measure and Bayesian analysis of single-case designs. *Journal of School Psychology*. doi: 10.1016/j.jsp.2013.12.002

Van den Noortgate, W., & Onghena, P. (2003). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods*, *35*(1), 1–10.

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment and Intervention*, *2*(3), 142–151. doi: 10.1080/17489530802505362

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *The Journal of Special Education*, *44*(1), 18–28. doi: 10.1177/0022466908328009

Appendix

Expected values of summary measurements

This Appendix provides further mathematical details about how the summary measurements generated by event counting, continuous recording, momentary time sampling, and partial interval recording are related to the behavior stream and ARP model.

Recall that the behavior stream consists of event durations $A_1, A_2, A_3, ...$ and interim times $B_0, B_1, B_2, B_3, ...$ For notational convenience, let $A_0 = 0$. Define $N(t)$ as the number of events that have begun by time t. Formally, this is the counting process

$$N(t) = \sum_{w=1}^{\infty} I\left[\sum_{v=0}^{w-1}(A_v + B_v) \leq t\right],$$

where $I()$ denotes the indicator function, such that $I(Q) = 1$ if Q is true, and $I(Q) = 0$ if Q is false. Finally, define

$$Z(t) = \sum_{w=1}^{\infty} I\left[0 \leq t - \sum_{v=0}^{w-1}(A_v + B_v) < A_w\right],$$

so that $Z(t) = 1$ indicates that an event is occurring at time t and $Z(t) = 0$ indicates that an event is not occurring at time t. With this notation established, the summary measurements can be defined precisely in terms of the behavior stream data and their expectations can be derived from the assumptions of the ARP.

The summary measurement generated by event counting can be defined formally as

$$Y^E = N(L),$$

where L is the length of the observation session. Under the ARP, the fact that $\text{E}\left(Y^E\right) = \zeta L$ as given in Equation (1) is a consequence of Blackwell's Renewal Theorem (Kulkarni, 2010, p. 360).

The summary measurement generated by continuous recording can be defined formally as

$$Y^C = \frac{1}{L}\int_0^L Z(t)dt,$$

Cox (1962, p. 101) demonstrates that $\text{E}\left(Y^C\right) = \phi$, as given in Equation (2).

In momentary time sampling, the observer notes the presence or absence of a behavior at K moments in time. The observations for these moments correspond to the points $Z(L/K), Z(2L/K), Z(3L/K), ..., , Z(KL/K)$. The summary measurement generated by momentary time sampling can therefore be written as

$$Y^M = \frac{1}{K}\sum_{k=1}^{K} Z(kL/K).$$

By the assumption that the process is in equilibrium, $\text{E}[Z(t)] = \phi$ for any fixed t, and so $\text{E}\left(Y^M\right) = \phi$, as given in Equation (3).

In partial interval recording, the observer scores each interval according to whether the behavior is present at any point during the active interval. Let U_k denote the score from interval k; formally,

$$U_k = I\left(0 < \int_0^c Z(t + (k-1)L/K)\,dt\right)$$

for $k = 1, ..., K$. The summary measurement from partial interval recording is then

$$Y^P = \frac{1}{K}\sum_{k=1}^{K} U_k.$$

Rogosa and Ghandour (1991) provided the expectation of a partial interval recording summary measurement for the special case of an alternating Poisson process, in which both A_1 and B_1 follow exponential distributions. Here I derive the expectation under the more general ARP model. First consider that the residual interim time at time-point t can be expressed as

$$R(t) = \sum_{u=0}^{N(t)}(A_u + B_u) - t.$$

In an equilibrium ARP, the conditional distribution of the residual interim time, given that $Z(t) = 0$, is

$$\Pr(R(t) \leq x | Z(t) = 0) = \frac{1}{\lambda}\int_0^x [1 - G(t;\lambda)]\,dt$$

(Kulkarni, 2010, Thm. 9.17). Let $s_k = (k-1)L/K$ be the time at which interval k begins, so that the state of process at the start of interval k is $Z(s_k)$. Conditioning on the initial state, the expectation of U_k is

$$\begin{aligned}\text{E}(U_k) &= \sum_{a=0}^{1}\Pr\left[0 < \int_0^c Z(s_k + t)\,dt \middle| Z(s_k) = a\right] \times \Pr[Z(s_k) = a] \\ &= (1-\phi)\Pr[R(s_k) < c | Z(s_k) = 0] + \phi \\ &= \zeta\int_0^c [1 - G(t;\lambda)]\,dt + \phi,\end{aligned}$$

as given in Equation (4).