A gradual effects model for single-case designs

Daniel M. Swan and James E. Pustejovsky

The University of Texas at Austin

May 15, 2018

Forthcoming in *Multivariate Behavioral Research*

This manuscript is not the copy of record and may not exactly replicate the final, authoritative version. The version of record is available at https://doi.org/10.1080/00273171.2018.1466681

Author Note

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

The research reported in this article was supported by Grant R305D160002 from the Institute of Educational Sciences, U.S. Department of Education. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education. A previous version of this paper was presented at the annual convention of the American Educational Research Association, April 28, 2017 in San Antonio, Texas. The authors are grateful to David Rindskopf, Nicholas Gage, and Daniel Maggin for feedback on earlier drafts of this paper.

Correspondence concerning this article should be addressed to Daniel M. Swan, Department of Educational Psychology, University of Texas at Austin, 1912 Speedway, Stop D5800, Austin, TX 78712-1289. Email: dswan@utexas.edu.

Abstract

Single-case designs are a class of repeated measures experiments used to evaluate the effects of interventions for small or specialized populations, such as individuals with low-incidence disabilities. There has been growing interest in systematic reviews and syntheses of evidence from single-case designs, but there remains a need to further develop appropriate statistical models and effect sizes for data from the designs. We propose a novel model for single-case data that exhibit non-linear time trends created by an intervention that produces gradual effects, which build up and dissipate over time. The model expresses a structural relationship between a pattern of treatment assignment and an outcome variable, making it appropriate for both treatment reversal and multiple baseline designs. It is formulated as a generalized linear model so that it can be applied to outcomes measured as frequency counts or proportions, both of which are commonly used in single-case research, while providing readily interpretable effect size estimates such as log response ratios or log odds ratios. We demonstrate the gradual effects model by applying it to data from a single-case study and examine the performance of proposed estimation methods in a Monte Carlo simulation of frequency count data.

Keywords: effect size; meta-analysis; log response ratio; single-case research; generalized linear model; intervention analysis

A gradual effects model for single-case designs

Single-case designs are a class of experiments involving repeated measurement of an outcome for one or a small number of cases. Cases are often individual participants, although in some applications cases correspond to groups or other aggregate units, while other applications involve a single individual measured across multiple contexts. For each case in the design, the outcome variable is measured repeatedly under two or more treatment conditions or phases (e.g., a baseline phase and an intervention phase), the timing of which is controlled by the researcher. The relative effect of the intervention is inferred by comparing the pattern of observed outcomes under contrasting treatment conditions for each case, such that each case serves as its own control (Horner & Odom, 2014).

Single-case designs play an important role in certain fields within education and psychology. Because they can be conducted with just one or a small number of participants, they are frequently used to evaluate interventions for specialized populations, such as individuals with low-incidence disabilities. Likewise, the designs are useful for evaluating interventions that are tailored specifically for an individual, for iteratively developing and refining interventions, or for gathering pilot evidence needed to warrant a larger study (Barton et al., 2016; Kaiser, 2014).

In practice, well-designed single-case studies involve not just one but several opportunities to test for intervention effects at different points in time. The most common form of single-case design used in applied research is the multiple baseline design (Shadish & Sullivan, 2011). In a multiple baseline design, the baseline phase begins at the same time for each case, but the start of the treatment phase begins at a different time for each case. Staggering the introduction of the intervention helps to guard against some threats to internal validity due to maturation or common history. Treatment phase trajectories that are similar across cases are interpreted as evidence that the outcome responds to intervention (Gast, Lloyd, & Ledford, 2014).

Another common form of single-case design, the treatment reversal (or ABAB) design, involves an initial baseline phase, followed by an initial introduction of an intervention; the intervention is then removed and baseline conditions are restored, followed by a re-introduction of the intervention phase; the process of removing and restoring the intervention can also be repeated further. Thus, each case in a treatment reversal design provides multiple opportunities to observe whether the outcome responds to a change in treatment conditions. Several other forms of single-case designs are also applied in practice (Horner & Odom, 2014), but are less prevalent and beyond the scope of the present article.

In light of the accumulation of evidence from single case designs in certain fields and for certain classes of interventions, there has been growing interest in using systematic reviews and syntheses of single-case studies. Syntheses of single-case research allow researchers to draw broader, more generalizable, and more nuanced conclusions about the effects of interventions than would be warranted from individual studies. Syntheses are potentially a useful tool for informing evidence-based practice in fields where single-case designs are prevalent (Horner et al., 2005). However, there remain a number of outstanding challenges and areas of contention related to methods for meta-analysis—and, more broadly, statistical analysis—of single-case data.

One of the central methodological questions involved in any meta-analysis is what effect size index should be used. In the context of single-case synthesis, effect size indices are quantitative measures of the direction and magnitude of intervention effects (Pustejovsky & Ferron, 2017). To be useful for meta-analysis, an effect size index needs to measure the theoretically meaningful characteristics of an intervention's effects, but do so in a way that allows for comparisons across studies (Hedges, 2008). Consequently, effect size indices should have stable parameter definitions that are not strongly influenced by procedural aspects of a study's design, such as different methods for measuring outcome variables, different sample sizes, or different forms of single-case design, which will often vary across a set of studies on a common topic (Pustejovsky, 2018a).

Recent critical reviews have identified several further criteria that effect size indices should ideally meet in order to be useful for meta-analysis of single-case research. First, an effect size should account for the characteristics of the data pattern, including the presence of time trends (Wolery, Busick, Reichow, & Barton, 2010). Such time trends might be present during baseline phases due to natural growth, as well as during intervention phases—particularly when the response to an intervention occurs gradually or builds up over time. Second, an effect size should be based on all available data, rather than focusing only on analysis of a subset of phases within the design (Horner, Swaminathan, Sugai, & Smolkowski, 2012; Wolery et al., 2010). Third, an effect size index should be based on appropriate assumptions about the distribution of the dependent variable (Shadish, 2014). Fourth, given that the outcomes in single-case designs constitute repeated measurements over time, many scholars have argued that effect size estimation methods should account for the possibility of serial dependence among the outcome measurements (Horner et al., 2012; Shadish, 2014; Wolery et al., 2010). Existing effect size measures and statistical models for single-case designs meet these criteria to varying degrees (Maggin et al., 2011), as we discuss further in the next section.

In this article, we propose a statistical model for single-case data—the *gradual effects model*—that meets many of the desiderata outlined above, while allowing for estimation of several different forms of effect size. The major novel feature of the model is that it allows for non-linear time trends during the treatment phase (or phases), which occur gradually and build up as the intervention remains in place. Drawing on the intervention analysis framework proposed by Box and Tiao (1975), the model posits a structural relationship between the intervention and the outcome, providing a parsimonious description of how the outcome variable responds in the presence or absence of an intervention. Consequently, the model can be applied to estimate effect sizes both from multiple baseline designs and from treatment reversal designs with an arbitrary number of phases.

In much of single-case research, the primary dependent variables are measures of

behavior assessed through systematic direct observation (Ayres & Ledford, 2014), which typically take the form of a count or proportion and do not conform well to modeling assumptions involving normally distributed error terms (Solomon, 2014). We therefore formulate the gradual effects model as a specific case of a generalized linear model (GLM) in the quasi-likelihood framework (McCullagh & Nelder, 1989). Under this framework, the outcome variable is described in terms of the relationship between its mean and variance, rather than in terms of an exact probability distribution. This provides a way to appropriately model outcome variables that are measured as counts or proportions, without having to assume that the outcome exactly follows a certain distribution (e.g., a Poisson distribution for counts or a binomial distribution for proportions). Furthermore, the GLM framework can be used to estimate several different forms of effect size, including log response ratios and log odds ratios. The model thus extends existing methodology for log-response ratio effect sizes (Pustejovsky, 2018b), which assumes that no time trends are present.

A limitation of the gradual effects model is that it does not explicitly account for serial dependence (autocorrelation) in the outcome data series, but instead assumes that outcome measurements are mutually independent. Although recent recommendations have emphasized the need to account for autocorrelation (Horner et al., 2012; Wolery et al., 2010), others have argued that available evidence on the presence of autocorrelation in single-case data is inconclusive. One argument is that linear or non-linear trends data that have not been modeled appropriately can create the appearance of serial dependence, and that once the trends are appropriately modeled, the errors might be serially independent (Huitema & McKean, 1998; Shadish, Kyse, & Rindskopf, 2013). Furthermore, estimation of even simple forms of autocorrelation is quite challenging given the small number of data points available in most single-case data series (Shadish, Rindskopf, Hedges, & Sullivan, 2013). Our model does not directly account for serial dependence, yet it may nonetheless be useful if it produces effect size estimates that are unbiased (or have only minor biases)

when the true data-generating process involves serial dependence. In simulation studies described subsequently, we find that the point estimates of effect size from the gradual effects model are indeed robust to un-modeled serial dependence, although the corresponding standard errors become systematically biased when the outcomes are not independent.

The remainder of the paper is organized as follows. The next section briefly reviews existing effect sizes and statistical models for single-case designs. The following section introduces the gradual effects model, using the framework of generalized linear models; describes the interpretation of effect sizes that can be estimated with the model; and demonstrates application of the model to data from a single-case design conducted by Thorne and Kamps (2008). In the following section, we examine the performance of the gradual effects model using Monte Carlo simulations across a range of data-generating conditions, including conditions with and without autocorrelation. In the final section, we discuss limitations of the model, directions for further research, and implications for practice.

Review of existing methods

We now briefly review existing effect size metrics and statistical models for single-case designs in order to contextualize the novel aspects of the gradual effects model, focusing in particular on modeling of time trends and applications to treatment reversal designs.

The most widely used effect sizes for single-case research come from the family of non-overlap effect sizes, which includes the percentage of non-overlapping data (Scruggs, Mastropieri, & Casto, 1987), non-overlap of all pairs (Parker & Vannest, 2009), and Tau-U (Parker, Vannest, Davis, & Sauber, 2011). These effect size indices are defined in terms of ordinal comparisons between data points from different phases. Non-overlap effect sizes are viewed as advantageous because they are relatively easy to calculate (many can be calculated directly from a graph of the data) and because they do not impose specific assumptions about the distribution of the dependent variable (Parker et al., 2011).

The non-overlap indices are all defined in terms of comparisons between a single baseline phase and a single treatment phase, and so do not directly address how to summarise data from treatment reversal designs involving more than two phases. Furthermore, most of the non-overlap indices do not account for any sort of time trends, although extensions exist that do account for certain forms of baseline trends (Manolov & Moeyaert, 2017). The main exception is Tau-U, which incorporates adjustments for monotonic trends in the baseline phase, treatment phase, or both (Parker et al., 2011), but the interpretation of these adjustments has been called into question (Tarlow, 2016). Related methods for modeling SCDs with trends include non-parametric techniques such as ECL, split middle, and tri-split (Manolov, 2017), but these only address linear baseline trends—not trends due to gradual treatment effects.

Other effect size measures for single-case research are defined in terms of parametric, distributional models for the data from a single case (for more extensive reviews, see Manolov & Moeyaert, 2017; Pustejovsky & Ferron, 2017). Basic parametric effect size measures include the within-case standardized mean difference described by Gingerich (1984), the log-response ratio, and the log-odds ratio. Pustejovsky (2015) introduced the latter two measures, arguing that log-response ratios and log-odds ratios are particularly appropriate for single-case designs with behavioral outcome measures measured as counts or proportions. Pustejovsky (2015) also argued that log-response ratios have a clearer interpretation than other effect sizes because their magnitude is less strongly influenced by operational details of procedures used to measure outcomes. As originally formulated, estimates of log-response ratio and log-odds ratio effect sizes assume that the data do not exhibit time trends and are serially independent. The gradual effects model described in the next section is one way of extending these effect sizes to account for non-linear forms of time trend.

A further class of effect sizes was introduced by Hedges, Pustejovsky, and Shadish (2012, 2013), who proposed a between-case standardized mean difference index that is

constructed to be comparable to the standardized mean difference estimated in a between-groups experimental design (see also Swaminathan, Rogers, and Horner 2014, who proposed Bayesian estimation methods for this effect size, and Shadish, Hedges, Horner, and Odom 2015, who discussed application of these methods in Special Education research). This index is distinctive because it is an *average* effect across all cases in a study, whereas other indices provide an effect size estimate for each case in a study. Although the index has been extended to handle linear time trends in multiple baseline designs (Pustejovsky, Hedges, & Shadish, 2014), available methods for treatment reversal designs do not account for time trends, instead assuming that the introduction or removal of treatment leads to immediate change in the outcome. Also, the effect size index is based on a model with normally distributed errors, which might not be appropriate for outcomes measured as counts or proportions.

Some recent work has focused specifically on statistical models for treatment reversal designs. Moeyaert, Ferron, Beretvas, and Van den Noortgate (2014) and Shadish, Kyse, and Rindskopf (2013) both explored several different methods of parameterizing multi-level regression models for treatment reversal designs, potentially with changes in time trend in addition to changes in level (see also Moeyaert et al. 2015 for an empirical application of these methods). Both studies focused largely on models with normally distributed errors and linear trends, but also included briefer explorations of non-normal error distributions and logistic regression specifications. A potential limitation of these models is that they do not provide an overall summary effect of treatment, instead using separate effects for each phase of the design. As a result, they require an increasing number of parameters as the number of phases in the design increases.

Little existing work has examined models with non-linear time trends, with two notable exceptions. Rindskopf (2014) proposed a model that used a logistic time trend, which is inherently non-linear in the parameters, to capture a gradual effect of treatment. Hembry, Bunuan, Beretvas, Ferron, and Van den Noortgate (2015) proposed a similar

model to capture non-linear trends in the treatment phase of a multiple baseline design. However, neither of these studies addressed how to link the proposed model to an effect size metric or how to apply the model to treatment reversal designs.

The gradual effects model described in the next section thus provides several features that are absent from existing methods: it can be used to estimate existing parametric effect sizes; it can be applied to a multiple baseline design or to a treatment reversal design with an arbitrary number of phases; and it is appropriate for outcomes measured as counts or proportions, both of which are common in single-case data. In the next section, we describe the technical details of the model.

The gradual effects model

The gradual effects model applies to the data series for a single case, which might be one of several within a treatment reversal or multiple baseline design. Suppose that an outcome Y is measured repeatedly across J occasions, or sessions, which occur under different treatment conditions. Let T_j be an indicator for the treatment condition in place on occasion j, so that $T_j = 1$ if session j is during an intervention phase and $T_j = 0$ if session j is during a baseline (or return-to-baseline) phase; let Y_j denote the observed outcome measurement from session j; and let $\mu_j = E(Y_j)$ denote the expected value of the outcome from session j, all for $j = 1, ..., J$. In what follows, we shall assume that the outcome measurements $Y_1, ..., Y_J$ are mutually independent.

The gradual effects model consists of three components: a link function, a variance function, and a functional specification. We describe each component in turn. The link function specifies the relationship between the mean of the outcome and the functional specification of the model, and thus determines the scale on which the outcome is modeled. Let $g()$ denote a generic link function. Specific examples of link functions include the identity link, where $g(x) = x$; the natural log link, where $g(x) = \ln(x)$; and the logistic link, where $g(x) = \ln(x) - \ln(1 - x)$. The choice of link function determines the form of the effect size estimated by the gradual effects model, as we explain in the next subsection.

Next, the variance function specifies the relationship between the expected value and the variance of the outcome from a given measurement occasion. Assuming that the variance of the outcome is constant with respect to the mean, $\text{Var}(Y_j) = \sigma^2$, yields an ordinary regression model, in which deviations from the mean are homoskedastic. Other variance functions are more appropriate for outcomes that are frequency counts or proportions. A basic model for frequency counts assumes that the outcome from a given session j follows a Poisson distribution, in which case $\text{Var}(Y_j) = \mu_j$. A basic model for proportions assumes that the outcome follows a multiple of a binomial distribution, in which case $\text{Var}(Y_j) = \mu_j(1 - \mu_j)$.

While conventional, the Poisson and binomial models entail quite strong assumptions about the mean-variance relationship. In a theoretical study of the psychometrics of behavioral observation, Rogosa and Ghandour (1991) showed that frequency counts generated from systematic direct observation of behavior may be over- or under-dispersed relative to the Poisson distribution, and similarly that proportion measures based on momentary time sampling may be over- or under-dispersed relative to the binomial. Fox (2008) argued that over- or under-dispersion relative to a Poisson-distributed variable is so common with count data that analysts should allow for extra dispersion as a matter of course. We therefore focus on flexible, quasi-likelihood variance functions, which assume

$$\text{Var}(Y_j) = \sigma^2 V(\mu_j) \tag{1}$$

for unknown scale parameter σ^2 and known function $V()$. Specifically, we use the quasi-Poisson variance function $V(\mu_j) = \mu_j$ and the quasi-binomial variance function $V(\mu_j) = \mu_j(1 - \mu_j)$.

The final component of the gradual effects model is the functional specification, which describes the relationship between the pattern of treatment conditions and the mean of the outcome on each measurement occasion. Let $\eta_j = g(\mu_j)$ for $j = 1, .., J$; in the GLM

framework, η_j is known as the linear predictor. The gradual effects model posits that:

$$\eta_j = \beta_0 + \beta_1 \frac{(1-\omega)}{(1-\omega^m)} \sum_{i=1}^{j} \omega^{j-i} T_i. \tag{2}$$

Equation (2) is a special case of the intervention analysis model for time series, introduced by Box and Tiao (1975). Here, β_0 represents the level of the outcome in the absence of treatment and β_1 represents the effect of treatment after m consecutive treatment sessions (note that m is a constant set by the analyst, as explained in more detail below). The parameter $0 \leq \omega < 1$ determines the delay in reaching the full effect of treatment. When $\omega = 0$, the effect of the treatment is immediate (no delay) and the model reduces to $\eta_j = \beta_0 + \beta_1 T_j$. As ω increases towards 1, the full effect of the treatment is increasingly delayed. A value of $\omega = 1$ would correspond to infinite delay; hence, the parameter space of ω excludes the value of 1. If the intervention is applied at a given time t, the predictor η_t changes by $\beta_1(1-\omega)/(1-\omega^m)$, but the effect of any previous instances of intervention also decay geometrically by a rate of ω. As a result, the net effect of repeated intervention sessions grows by smaller and smaller increments, eventually approaching a new equilibrium level. On the other hand, removing the intervention leads to gradual reductions in the level of the outcome. With further sessions in the absence of intervention, the outcome Y would eventually approach the original baseline level.

Figure 1 depicts several examples of functional specifications that are possible under the gradual effects model, using values of ω equal to 0.0, 0.3, 0.6, or 0.9. Each panel plots the linear predictor for a treatment reversal design with 4 phases and 10 sessions per phase. When there is no delay, the model is identical to a simple change-in-level model. As the delay increases, the intervention and return-to-baseline phases begin to exhibit curvilinear time trends. When there is a great deal of delay, the model more closely resembles one with linear slopes in each phase. Note that when $\omega = 0.9$, it takes more than 10 intervention sessions to reach the full effect of treatment and more than 10 sessions in the absence of intervention to fully return to baseline.

The gradual effects model in Equation (2) is formulated in structural terms, such

that the mean of the outcome on a given occasion j depends explicitly on the current and all prior treatment conditions. This feature makes it possible to apply the gradual effects model to treatment reversal designs with an arbitrary number of baseline and treatment phases of any length. Of course, the model can also be applied to the data from a case within a multiple baseline design, in which a given case is observed under only one baseline and one intervention phase. The functional specification can then be expressed more simply as

$$\eta_j = \beta_0 + \beta_1 \frac{(1 - \omega^{U_j})}{(1 - \omega^m)}, \tag{3}$$

where $U_j = \sum_{i=1}^{j} T_i$ is the number of sessions since the beginning of the intervention phase (including the current session), with $U_j = 0$ during the baseline phase.

Effect sizes under the gradual effects model

In the gradual effects model, the interpretation of the effect size parameter β_1 is determined by two analytic decisions: the choice of the scaling time m and the choice of link function $g()$.

The scaling time is incorporated into the model through the denominator of the second term in Equation (2), as $1 - \omega^m$. This allows the analyst to specify the definition of the effect size as corresponding to the effect of m consecutive intervention sessions. Taking $m = \infty$ means that the effect size corresponds to the effect of continuing the intervention indefinitely, or what we call an equilibrium treatment effect. However, if the effect of treatment is very gradual, focusing on the equilibrium treatment effect may entail extrapolating far beyond the observed data. We therefore recommend that the analyst choose a value of m within—or at least not much larger than—the range of intervention phase lengths observed in the data. For example, imagine an analyst is interested in examining a set of a dozen ABAB series with treatment phase lengths ranging from 8 to 16, with a mean of about 12. If the analyst wanted to compare effect sizes across all of the series, they might select $m = 10$ for all series. This value does not entail extrapolating much beyond the range of the shortest phases, while still capturing a large portion of

treatment phases with more than 10 observations.

The specific link function employed determines the effect size metric produced by the gradual effects model. With the identity link, β_1 represents an additive effect, or an unstandardized mean difference. With the log link, β_1 represents a log response ratio and $\exp(\beta_1)$ represents a proportional change in the outcome due to m consecutive intervention sessions. To see this, note that in the absence of any intervention, the expected level of the outcome during session m would be $\mu_m = \exp(\beta_0)$. If the intervention were in place for the first m sessions, then the expected level of the outcome during session m would be $\mu_m = \exp(\beta_0 + \beta_1) = \exp(\beta_0)\exp(\beta_1)$ (see Equation 3 with $U_m = m$). Thus, the ratio of the expected outcome under intervention to the expected outcome in the absence of intervention is $\exp(\beta_1)$. Finally, with an outcome that is a proportion and a logit link, β_1 represents a log odds ratio and $\exp(\beta_1)$ corresponds to the proportionate change in the odds of the outcome due to m consecutive intervention sessions. In contrast to the unstandardized mean difference, the log response ratio and log odds ratio are scale-free metrics. They are thus more suitable for summarizing intervention effects across a set of studies that measure outcomes using different operational procedures, such as longer or shorter observation sessions.

Estimation

Conventional GLMs, in which the functional specification is linear in the parameters, are typically estimated by maximizing the quasi-likelihood function of the model. Although complicated by the fact that the model includes the non-linear delay parameter ω, we propose to use the same technique—maximum quasi-likelihood estimation—to estimate the parameters of the gradual effects model. Appendix A describes an algorithm for obtaining maximum quasi-likelihood estimates and associated standard errors by profiling the non-linear parameter. We provide software implementing the proposed maximum quasi-likelihood estimation procedure in the form of an R package (Pustejovsky & Swan, 2017) and online web application (Swan & Pustejovsky, 2017).

An example

Thorne and Kamps (2008) used a single-case design to evaluate the effects of a group contingency intervention on academic engagement and levels of problem behaviors among students at risk for developing behavioral disorders. Twelve participating students were drawn from four classrooms, including two third grade and two second grade classes in a suburban community. The group contingency intervention involved giving away small prizes such as pencils or erasers to students contingent on their personal behavior and a party contingent on the behavior of the entire classroom. Twice daily during intervention phases, participating teachers distributed lottery tickets based on desirable behaviors and then immediately drew four to five winners, who were allowed to choose for themselves among the available rewards. The intervention was evaluated using an ABAB design replicated across the twelve participants. A primary outcome, frequency of inappropriate behavior, was recorded via direct observation of each student during 15-minute academic periods. Figure 2 displays the raw data for each of the twelve participants in the study. Observation sessions took place daily at shifting times across the morning and afternoon, in order to capture behaviors of interest under a variety of conditions. Complete raw data as well as code for model estimation can be found in the supplementary materials (Swan & Pustejovsky, 2018).

Before estimating the gradual effects model, it is critical to first consider whether its assumptions are reasonable and appropriate for these data. Two critical assumptions are that the baseline level of behavior is stable and that the outcome responds gradually to introduction and removal of the treatment. Visual inspection of the data in Figure 2 does not reveal systematic time trends during the baseline phases (although all phases are short), but does suggest the possibility of slight downward trends during the initial treatment phases for some cases (e.g., participants 1, 7, 8, and 12) and upward trends during return-to-baseline phases. For other cases, the response to treatment appears to be nearly immediate (e.g., participant 2). The presence of varying time trends during

intervention and return-to-baseline phases suggests that the gradual effects model could be useful for estimating the effects of the intervention.

We fit the gradual effects model to the inappropriate behavior data for each of the twelve cases in the study. Because the inappropriate behavior outcome was measured as a frequency count, we modeled the data using a quasi-Poisson variance function and a log link function, so that the effect size estimates are log-response ratios. Across participants, the treatment phases varied in length from 9 to 16 sessions, with a mean of about 12 sessions. To avoid extensive extrapolation, our primary results are based on effect sizes after $m = 10$ intervention sessions. We examined sensitivity to this choice by also estimating effects after $m = 5$ and $m = 15$ intervention sessions.

For comparison purposes, we also obtained estimates using the R_1 estimator for the log-response ratio, as described in Pustejovsky (2015). In contrast to the gradual effects model, the R_1 estimator assumes that there are no time trends, and so is equivalent to a model with an immediate change-in-levels (i.e., $\omega = 0$). Finally, for both the gradual effects and change-in-levels model, we estimated average log-response ratios across all twelve cases based on a random effects meta-analysis model. We used robust variance estimation methods proposed by Sidik and Jonkman (2006) to calculate standard errors for the average effect size.

A plot comparing the fitted values from the gradual effects model to the observed data can be found in the supplementary materials (. Table 1 reports the log response ratio estimates and standard errors for the change-in-levels model and the gradual effects model. The effect size estimates from the gradual effects model were generally larger in magnitude than the estimates from the change-in-levels model because the former model allows for change over time during the intervention and return-to-baseline phases, whereas the latter model just averages across all sessions within the baseline and intervention conditions. The overall average effect size based on the gradual effects model was -1.34, 95% CI [-1.66, -1.02], which corresponds to a 74% reduction in inappropriate behavior (95% CI

[64%, 81%]). Although the standard errors of the individual effect size estimates were comparable across both models, the standard error of the average effect size estimate from the gradual effects model is 61% larger than the corresponding standard error based on the change-in-levels model. This is because the effect size estimates from the gradual effects model are more heterogeneous than those from the change-in-levels model, with an estimated between-case variance of 0.285 versus 0.092.

It is important to note that, for both models, the standard errors for individual effect size estimates are based on the assumption that the outcomes are mutually independent. If this assumption is mistaken and the outcomes are positively autocorrelated, we would expect the standard errors to under-state the true extent of uncertainty in the effect size estimates. However, using the Sidik and Jonkman (2006) robust variance estimation methods ensures that the standard error and confidence interval for the overall average effect size estimate remain valid even if the outcomes are actually autocorrelated.

Table 1 also reports estimates of the delay parameter and the dispersion parameter for the gradual effects model. In the cases where there is obvious non-linearity to the effect of treatment, such as participant 1 and 12, the delay is relatively high ($\omega > 0.60$), and there are considerable differences between the treatment effect estimates from the two models. When the full effect of treatment happens nearly immediately, such as with participants 3 or 11, the gradual effects model provides estimates similar to the change-in-levels model. Note also that the dispersion estimates (σ^2) vary across cases, with four cases over-dispersed and six cases under-dispersed, which suggests that it would not be appropriate to assume that the outcome is strictly Poisson-distributed.

Finally, the meta-analytic results from the sensitivity analysis are presented in Table 2. Detailed, case-level estimates for the models when $m = 5$ and $m = 15$ can be found in the supplementary materials (Swan & Pustejovsky, 2018). Using $m = 5$, the average effect of the intervention was -1.27, corresponding to a 72% reduction in inappropriate behavior (95% CI [62%, 79%]). This estimate is slightly less than the 74% reduction in inappropriate

behavior using $m = 10$ and $m = 15$. The difference in the estimates arises because the treatment effect has not yet reached equilibrium for all cases after 5 consecutive treatment sessions. By 10 treatment sessions and continuing through 15 consecutive treatment sessions, the treatment effect approaches equilibrium and are therefore no longer influenced by increasing the value of m. As the value of m increases, there are more series where the summary treatment effect estimate involves extrapolating outside the data, the uncertainty around individual effect size estimates increases, and the uncertainty around the meta-analytic treatment effect estimate increases slightly. In cases where the effect of treatment is more delayed, the model might be more sensitive to the specification of m. In this example, however, the gradual effects model was relatively robust to varying specifications for the value of m.

This example illustrates the value of a flexible non-linear model for single-case designs. In the absence of time trends, the model generates estimates that are similar to the simpler change-in-levels model. In contrast, when there is a more gradual change in behavior, the model accounts for the resulting time trends and captures the effect of treatment at a specified point in time. Applying the gradual effects model to each case allows for heterogeneity in the immediacy of effect while yielding an effect size estimate that is comparable across cases. The use of the quasi-Poisson variance function reflects the mean-variance relationship seen in the data, without making strong distributional assumptions. However, this is only a single example, and so it is important to understand the accuracy and bias of effect size estimates from the gradual effects model more generally. In order to investigate this question, we conducted a simulation study.

Simulation study

We conducted a Monte Carlo simulation study to evaluate parameter recovery in the gradual effects model. Our primary concern was the accuracy and bias of effect size estimates generated by the model—especially compared to the accuracy of estimates based on a simpler, change-in-levels model. Additionally, we assessed the relative bias of the

variance estimates under the gradual effects model and the bias of the delay parameter estimates. We evaluated parameter recovery under two different scenarios. In the first scenario, we simulated outcomes independently, so that the data-generating model was fully consistent with the assumptions of the gradual effects model. In the second scenario, we simulated outcomes that had the correct functional specification, but were also autocorrelated following a first-order auto-regressive process. This permitted assessment of whether parameter estimates and standard errors were robust to un-modeled autocorrelation in the outcome data.

The GLM formulation of the gradual effects model is very flexible, in that it can be applied with a variety of different link functions and variance functions. For purposes of simulation, however, we focused on one important use-case: Poisson-distributed frequency count outcomes, modeled using a log link. The log link function produces log response ratio effect size estimates, which are on a metric suitable for many types of behavioral outcome measures and are intuitively appealing because they can be interpreted in terms of percentage change (Pustejovsky, 2018b). Additionally, about half of all single-case designs examined by Shadish and Sullivan (2011) used frequency count outcomes, which indicates that models for frequency counts are important in practice.

Data-generating model

The simulations involved generating Poisson-distributed outcomes that were either mutually independent or serially correlated. We generated serially correlated data using a binomial thinning process (McKenzie, 1988), which reduces to independence when the autocorrelation parameter is zero. Appendix B describes the exact algorithm used to simulate autocorrelated Poisson counts.

A challenge with estimating the log-response ratio is that if no instances of behavior are observed during the treatment phase(s) then the effect size estimate will diverge towards $-\infty$. Data series with no instances of behavior in the treatment phase are relatively more common when the baseline level of the outcome is small and the true

treatment effect is a substantial decrease, and it will not generally be reasonable to apply the gradual effects model to estimate log response ratio effect sizes with such data. In order to focus the simulation results on scenarios where the model is potentially appropriate, we discarded simulation replicates where the total number of behaviors across all treatment phases was ≤ 1.

Table 3 lists the parameter values used to generate simulated data. We expected that the baseline level would have an impact on the bias of the treatment effect, and so examined baseline mean levels of 5, 15, or 25 events per observation session. These rates are consistent with low-frequency, moderate-frequency, and high-frequency behaviors observed in a review of multiple baseline and treatment reversal designs published in 2008 (Shadish & Sullivan, 2011). We selected values for the treatment effect parameter to represent a wide range of proportionate changes. Specifically, we varied the range of β_1 from -1.6 to 1.6 in equal increments across the log scale, which corresponds to proportionate changes ranging from approximately an 80% reduction to a 500% increase. We selected values for the delay parameter to cover most of the range of the parameter. We generated data without autocorrelation to reflect cases where our model correctly captures the functional form and no serial correlation remains in the data. We also generated data with two levels of moderate autocorrelation ($\phi = 0.2$ and $\phi = 0.4$) to reflect model mis-specification. In their review of SCD features, Shadish and Sullivan (2011) reported that the average degree of autocorrelation was 0.2 (based on a random effects meta-analysis), so we selected that level as well as levels both above and below it.

We simulated data following a treatment reversal design with four phases (i.e., an ABAB design). According to the What Works Clearinghouse standards for SCDs (Kratochwill et al., 2013), the intervention must be manipulated at least three times for experimental control to be demonstrated and for a treatment reversal design to meet standards. Furthermore, four phases was the median observed in the studies examined by Shadish and Sullivan (2011). We varied the number of sessions per phase, here denoted as

n_p, between 3 and 10. Three points per phase are required for an SCD to "meet standards with reservations," while five points are required to fully meet the standards. We also included 10 points per phase to examine performance for designs that go beyond these minimum standards. In all cases, the scaling parameter m was set equal to the number of points per phase in order to avoid extrapolating beyond the range of the data.

These parameter values were fully crossed in a $3 \times 9 \times 4 \times 2 \times 3$ factorial design. We simulated 10000 replications per condition. For each replication, we fit the gradual effects model as well as the simpler change-in-levels model (using the R_1 estimator). Following Hoogland and Boomsma (1998), we describe biases of no more than .05 on the log scale or proportionate biases of no more than 5% as "approximately unbiased." Complete code for replicating the simulation, as well as an appendix with more detailed simulation results, can be found in the supplementary materials (Swan & Pustejovsky, 2018).

Results

Accuracy

Figure 3 illustrates the root mean-squared error (RMSE) of the log-response ratio effect size estimates from the gradual effects model compared to the change-in-levels model, when the outcome data are independent. Each box-plot represents the distribution of RMSE across different values of the delay parameter, for specified values of the true treatment effect (horizontal axis), true baseline level (vertical tiles), and phase length (horizontal tiles). Except in scenarios where there are very few points per phase and the baseline level is very low (i.e., $n_p = 3$ and $\exp(\beta_0) = 5$) or when there is no effect of treatment ($\beta_1 = 0$), the gradual effects model produces equally accurate or more accurate estimates than the change-in-levels model. Autocorrelation reduces the accuracy of both models, but does not change this relationship. As delay increases, the gap in accuracy between the intervention analysis model and the change-in-levels model grows larger. The supplementary materials provide further details about these relationships(Swan & Pustejovsky, 2018).

Bias

Figure 4 illustrates the bias of the effect size estimate from the gradual effects model when the outcome data are generated independently. The estimates are approximately unbiased when the design includes at least $n_p = 5$ observations per phase. It is more difficult to estimate the effect when the initial baseline level is low ($exp(\beta_0) = 5$) and when intervention leads to reductions in the outcome ($\beta_1 < 0$). Additionally, the degree of bias increases for smaller values of the delay parameter (that is, treatment effects that are more immediate).

Figure 5 illustrates the bias of the treatment effect estimate when the number of observations per phase is $n_p = 5$ and the outcomes are either independent (left column) or autocorrelated at $\phi = 0.2$ and $\phi = 0.4$ (middle and right columns, respectively). Generally speaking, the patterns of bias when autocorrelation is present are similar to the patterns with independently generated data. However, increases in autocorrelation tend to somewhat magnify the degree of bias under conditions where it is present. For instance, when $\phi = 0.2$ and the baseline level is $exp(\beta_0) = 15$, an immediate treatment effect ($\omega = 0$) leads to biased estimates when the true treatment effect is a reduction. When $\phi = 0.4$, the biased estimates extend to conditions where there is a delay of $\omega = 0.3$.

Variance estimation

Figure 6 illustrates the relative bias of the effect size variance estimator when the outcomes are generated independently. Across baseline levels, the variance estimates are close to unbiased when there there are many observations per phase ($n_p = 10$) and when non-null treatment effects are present. However, the variance estimator performs poorly under all other conditions. Notably, the variance estimator is biased towards zero when there is no effect of treatment ($\beta_1 = 0$).[1] The poor performance of the variance estimator

[1] In additional simulations with $\beta_1 = -0.2, -0.1, 0.1, 0.2$, we found that the variance estimator is also biased when the true treatment effect is close to zero, and that the extent of bias is larger when the baseline level of the outcome is lower.

when the true treatment effect is near zero is a consequence of the fact that the delay parameter ω is not identified when $\beta_1 = 0$, meaning that the quasi-likelihood remains constant for any value ω. Similarly, when β_1 is very close to zero, ω is only weakly identified and so our conventional approach to variance estimation breaks down.

Figure 7 illustrates the relative bias of the variance estimator for the effect size when $n_p = 5$ across all values of autocorrelation (ϕ). Across conditions, the variance estimator always has substantial bias, tending under most conditions to under-state the sampling variance of the effect size estimator. As would be expected, larger degrees of autocorrelation cause the variance estimates to be further under-stated.

In summary, the variance estimator appears to have non-negligible biases under many conditions, even when the assumption of independence holds.

Other aspects of parameter recovery

Plots of the bias of ω can be found in the supplementary materials(Swan & Pustejovsky, 2018). Less biased estimates of ω occur when the true value of ω is closer to the middle of the parameter space, with higher bias observed at the edges of the parameter space. When the true $\omega = 0$, the estimator is biased across most conditions examined. The bias of the ω estimator is reduced when the treatment effect is larger in absolute magnitude, when the baseline level is larger, and when the design includes more observations per phase. Autocorrelation in the outcome leads to more biased estimates, but for the most part this does not alter the conditions where the estimator is approximately unbiased. In general, moderate delay ($0.30 \leq \omega \leq 0.6$), at least a moderate baseline level ($exp(\beta_0) = 15$), and at least a moderate effect ($|\beta_1| \geq 0.80$) are necessary for the estimator of omega to be approximately unbiased.

Discussion

There are three primary findings from the simulation study. First, for designs with reasonable phase lengths and baseline outcome levels, the model provides close-to-unbiased estimates of treatment effects even in the presence of autocorrelation. The exceptions occur

when there is little data, or the absolute magnitude of the outcomes is quite small. In the case of little data, most parametric models will not perform well with only 3 observations per phase and a total of 12 observations for the entire series. In the case of small frequency counts, performance could be ameliorated through the use of longer observation sessions. Second, the variance estimator performs poorly even when the model's assumptions are satisfied and the outcomes are independent. Although of concern, this shortcoming can nonetheless be addressed when conducting meta-analysis of the effect size estimates, as we discuss further in the next section. Third, estimating the delay parameter is difficult. However, we believe this is an acceptable shortcoming because the purpose of ω is primarily to account for non-linear time trends, rather than to measure a feature of primary substantive interest.

One limitation of this study is that we explored a limited set of values for the baseline level. Given that there appears to be sensitivity to the magnitude of the baseline across many of the parameter estimates, further exploration of these relationships is warranted. In addition, we limited our exploration of autocorrelation to just two values, and the parameter estimates that are sensitive to autocorrelation may be more or less robust to larger degrees of autocorrelation. Finally, we did not explore the performance of the model with outcomes measured as proportions (e.g. change in the percentage of time target behaviors occurred) or with logit link functions. Data generating processes for autocorrelated proportional outcomes are more challenging to construct than for counts or normally-distributed outcomes. Additionally, one of the most common observation methods that produces proportional outcomes, partial interval recording, has procedural sensitivities that can produce biased measurements of the underlying behavioral characteristics (Pustejovsky & Swan, 2015). A thorough exploration of the gradual effects model with proportional outcomes would ideally take all of these issues into account, and remains a topic for further research.

General discussion

In this paper, we have introduced the gradual effects model as a tool for estimating effect sizes from single-case data that exhibits non-linear time trends as a result of intervention. Because it is formulated in the framework of generalized linear models (McCullagh & Nelder, 1989), the model provides a way to estimate effect size measures such as log response ratios or log odds ratios, while also appropriately modeling dependent variables measured as counts or proportions. The model is a special case of the intervention analysis framework introduced by Box and Tiao (1975), in which the outcome in a given session is structurally related to the presence or absence of intervention in the current and previous sessions. As a result, the model can be applied not only to multiple baseline designs, but also to treatment reversal designs with an arbitrary number of phases, without the need to introduce further parameters into the model.

We see at least two distinct use cases for the gradual effects model: as a tool for analyzing individual data series and as a tool for estimating effect sizes across multiple cases, possibly drawn from multiple studies. The interactive web-app (Swan & Pustejovsky, 2017) provides facilities for fitting the gradual effects model in both of these scenarios. The gradual effects model has different advantages and limitations in each scenario. In describing guidance for analysts interested in applying the model, it therefore makes sense to consider each scenario in turn.

Analyzing a single data series

Researchers may be interested in applying the gradual effects model to analyze data from a single data series. In this scenario, the model is particularly advantageous for analyzing treatment reversal designs due to its structural form, which readily extends to designs with an arbitrary number of phases while providing a single summary estimate of the effect of treatment. Moreover, the non-linear form of the model, in which the outcome approaches an equilibrium level as treatment continues and returns to the baseline level when treatment is discontinued, is consistent with many of the types of interventions that

could be evaluated with a treatment reversal design. Other parametric models that are directly applicable to treatment reversal designs (e.g., Moeyaert et al., 2014; Shadish, Kyse, & Rindskopf, 2013) do not provide a single summary effect estimate, while existing models that allow for non-linear functional forms (e.g., Hembry et al., 2015; Rindskopf, 2014) are not readily applicable to treatment reversal designs without considerable modification. Thus, the gradual effects model is useful and should be applied when an analyst desires a single summary effect size estimate and when non-linear trajectories in the data series are observed or theoretically plausible.

When applying the gradual effects model to generate an effect size estimate, the analyst must specify a value for m, the number of treatment sessions at which the treatment effect is estimated. Given the non-linear nature of the model, the value chosen for m is important because the treatment effect at time $m + i$ for a given time difference of i is not a simple transformation of the treatment effect at time m. With a single data series, a natural choice for m would be the length of the longest treatment phase; the analyst's choice might also be informed by considering typical intervention durations for the research area.

When applied to an individual data series, an important limitation of the proposed model and estimation methods is that they do not explicitly allow for autocorrelation of the outcomes. Simulation results indicated that, while point estimates of treatment effects have reasonably small biases even with moderate degrees of auto-correlation, the standard errors from the model are strongly affected by auto-correlation and must therefore be interpreted with considerable caution. When reporting effect size estimates for individual series, we therefore strongly recommend emphasizing that the standard errors are based on the assumption of independent outcomes and are likely to under-state the degree of uncertainty if the outcomes are positively auto-correlated. Despite these concerns, we nonetheless believe it is better to report an effect size estimate and corresponding standard error with known sampling properties than to report an effect size estimate without any

indication of uncertainty, or one that is inappropriately assumed to be robust to auto-correlation. For testing the null hypothesis of no intervention effect with a single data series, it may be useful to combine the gradual effects model with a randomization test (Edgington & Onghena, 2007) rather than using parametric standard errors.

Analyzing multiple data series

A second use-case for the gradual effects model is in estimating summary effect sizes across several data series, either drawn from a single study or from multiple studies evaluating the effects of a common class of interventions. An analyst interested in using single-case data must think carefully about the appropriate statistical model to estimate an effect size for a given instance of an intervention. If an analyst is interested in research synthesis combining both between-group and single-case designs, the between-case standardized mean difference (Hedges et al., 2012, 2013; Pustejovsky et al., 2014) may be a more appropriate approach for summarizing the results of the single-case studies. If, in the collection of cases to be analyzed, most treatments seem to reach full effect immediately, it would be appropriate to use a more parsimonious change-in-levels model. The gradual effects model offers advantages for summarizing a collection of cases that include some combination of treatment reversal designs, non-linear trends in many of the treatment phases, or markedly non-normal error distributions. When some combination of these conditions is present or is expected to be present in the data, we would suggest that analysts consider applying the gradual effects model.

When using the gradual effects model for meta-analysis, choosing a value for m for comparability across different series within a study or for estimates across studies is an important consideration. There are at least three defensible ways to choose a value for m. The first is to set m to estimate a treatment effect after a common number of observation sessions, so that all estimates represent a common operationalization of treatment. Here, we recommend that the analyst choose the largest value of m possible without using a value that is much beyond the longest treatment phase in a given series. For instance, in

our empirical example, we picked a value of $m = 10$ because it is not much less than the mean number of observations in the treatment phase (12) and is not much longer than the shortest treatment phase (8). Of course, not all single-case studies use the same frequency of observation, and so a single common m could represent different lengths of time. A second method would be to set the value of m within each study to represent a common length of time, such as a week, a month, or a school semester. Third, in some contexts it might also make sense to set the value of m within each study to represent the typical length of a given intervention—particularly if there is considerable between-study variation in intervention duration. In all cases, the analyst should pay careful attention to variation in the operational details between studies. We also recommend reporting sensitivity analysis for varying values of m, as we have demonstrated in the analysis of the Thorne and Kamps (2008) data.

Finally, we would like to note that the considerable biases present in the variance estimates is of less concern in the context of meta-analysis than it is in the analysis of individual cases. The meta-analysis technique of robust variance estimation (RVE; Hedges, Tipton, & Johnson, 2010; Sidik & Jonkman, 2006) can be used to estimate overall average effect sizes and calculate standard errors and confidence intervals that are robust to the use of inaccurate standard errors, which might arise due to un-modeled autocorrelation. In the case where a researcher is reporting an average effect size across several cases from a single study, the Sidik and Jonkman (2006) methods that we demonstrated will provide valid standard errors in the presence of auto-correlation in the outcome data series. If a researcher is interested in estimating a summary effect across cases from multiple studies, the robust variance estimation methods proposed by Hedges et al. (2010) will provide standard errors that account for dependence among effect sizes from the same study (i.e., cluster-dependence) as well as the presence of autocorrelation in the outcome data series. In each context, we recommend using and reporting RVE standard errors and confidence intervals around average treatment effect estimates when synthesizing effect size estimates

from the gradual effects model.

Future directions

Several extensions to the gradual effects model warrant investigation in further research. First, the current formulation of the gradual effects model applies to data from a single case. However, the proposed case-level model could be used as a building block for a multi-level model that captures variation in its parameters across several cases within a study, or even within and across cases from several studies, following the approach of Moeyaert et al. (2014). Integrating the gradual effects specification into a multi-level model could also provide a way to estimate an effect size parameter that is comparable to the corresponding parameter from a between-groups experimental design, extending the approach described by Pustejovsky et al. (2014) for multiple baseline designs with linear trends.

Second, the gradual effects model might be useful for modeling time trends that arise for reasons other than intervention. One issue in the use of single-case designs is the presence of testing effects, which arise when the method used to assess the dependent variable itself produces changes in the outcome. For instance, repeatedly administering a test of domain knowledge could lead to improved performance, even in the absence of instruction. Modifying the gradual effects model to include a non-linear trend during the initial baseline phase could be one way to account for testing effects in designs where they may be present.

Third, the gradual effects model captures a limited range of functional forms, which will not be appropriate for all forms of time trends encountered in single-case data. We believe that this parsimonious model is useful and appropriate in many cases, particularly given the limited number of observations used in many single-case designs. However, it would nonetheless be useful to investigate how to capture a broader range of functional forms, especially for studies that use more intensive measurement schedules, while still maintaining the structural features of the model. The intervention analysis modeling

framework (Box & Tiao, 1975) is a natural starting point for such extensions.

In visual analysis of single-case designs, there is said to be a *functional relationship* between the independent and dependent variables when the outcome responds to the introduction (and possibly also removal and re-introduction) of treatment. This functional relationship is expressed quite literally in the structure of the gradual effects model, in that the level of the dependent variable is a non-linear function of the manipulation of the independent variable (intervention) up to that point in time. This property makes the model well suited for application to treatment reversal designs, which necessarily involve interventions that can be removed after being put in place and effects that dissipate in the absence of continued treatment. Developing models for the specific forms of functional relationships observed in other classes of single-case designs, such as alternating treatment designs, adapted alternating treatment designs, and multi-element designs, remains an important endeavor for future work.

References

Ayres, K., & Ledford, J. R. (2014). Dependent measures and measurement systems. In D. L. Gast & J. R. Ledford (Eds.), *Single-case research methodology: Applications in special education and behavioral sciences* (pp. 124–153). New York, NY: Routledge.

Barton, E. E., Ledford, J. R., Lane, J. D., Decker, J., Germansky, S. E., Hemmeter, M. L., & Kaiser, A. (2016). The iterative use of single case research designs to advance the science of EI/ECSE. *Topics in Early Childhood Special Education*, *36*(1), 4–14. doi: 10.1177/0271121416630011

Box, G. E. P., & Tiao, G. C. (1975). Intervention Analysis with Applications to Economic and Environmental Problems. *Journal of the American Statistical Association*, *70*(349), 70. doi: 10.2307/2285379

Edgington, E., & Onghena, P. (2007). *Randomization Tests*. Boca Raton, FL: Chapman & Hall.

Fox, J. (2008). *Applied regression analysis and generalized linear models* (2nd ed ed.). Los Angeles: Sage.

Gast, D. L., Lloyd, B. P., & Ledford, J. R. (2014). Multiple baseline and multiple probe designs. *Single case research methodology: Applications in special education and behavioral sciences*, 251–296.

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

Hedges, L. V. (2008). What are effect sizes and why do we need them? *Child Development Perspectives*, *2*(3), 167–171. doi: 10.1111/j.1750-8606.2008.00060.x

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods*, *3*(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

Hembry, I., Bunuan, R., Beretvas, S. N., Ferron, J. M., & Van den Noortgate, W. (2015). Estimation of a Nonlinear Intervention Phase Trajectory for Multiple-Baseline Design Data. *The Journal of Experimental Education*, *83*(4), 514–546. doi: 10.1080/00220973.2014.907231

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*(3), 329–367.

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. doi: 10.1177/001440290507100203

Horner, R. H., & Odom, S. L. (2014). Constructing single-case research designs: Logic and options. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 53–90). Washington, DC: American Psychological Association.

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

Huitema, B. E., & McKean, J. W. (1998). Irrelevant autocorrelation in least-squares intervention models. *Psychological Methods*, *3*(1), 104.

Kaiser, A. P. (2014). Using single case designs in comprehensive programs of research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 309–323). Washington, DC: American Psychological Association.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-Case Intervention Research Design Standards. *Remedial and Special Education*, *34*(1), 26–38. doi: 10.1177/0741932512452794

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), 301–321. doi: 10.1016/j.jsp.2011.03.004

Manolov, R. (2017). Linear trend in single-case visual and quantitative analyses. *Behavior Modification*, *In press*. doi: 10.1177/0145445517726301

Manolov, R., & Moeyaert, M. (2017). Recommendations for choosing single-case data analytical techniques. *Behavior Therapy*, *48*(1), 97–114. doi: 10.1016/j.beth.2016.04.008

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed ed.) (No. 37). London ; New York: Chapman and Hall.

McKenzie, E. (1988). Some ARMA Models for Dependent Sequences of Poisson Counts. *Advances in Applied Probability*, *20*(4), 822. doi: 10.2307/1427362

Moeyaert, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2014). From a single-level analysis to a multilevel analysis of single-case experimental designs. *Journal of School Psychology*, *52*(2), 191–211. doi: 10.1016/j.jsp.2013.11.003

Moeyaert, M., Ugille, M., Ferron, J. M., Onghena, P., Heyvaert, M., Beretvas, S. N., & Van den Noortgate, W. (2015). Estimating intervention effects across different types of single-subject experimental designs: Empirical illustration. *School Psychology Quarterly*, *30*(1), 50.

Parker, R. I., & Vannest, K. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy*, *40*(4), 357–367.

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.

Pustejovsky, J. E. (2015). Measurement-comparable effect sizes for single-case studies of free-operant behavior. *Psychological Methods*, *20*(3), 342–359. doi: 10.1037/met0000019

Pustejovsky, J. E. (2018a). Procedural sensitivities of effect sizes for single-case designs with behavioral outcome. *Psychological Methods*, forthcoming. Retrieved from https://osf.io/p3nuz/

Pustejovsky, J. E. (2018b). Using response ratios for meta-analyzing single-case designs with behavioral outcomes. *Journal of School Psychology*, forthcoming. Retrieved from https://osf.io/4fe6u/

Pustejovsky, J. E., & Ferron, J. M. (2017). Research synthesis and meta-analysis of single-case designs. In J. M. Kaufmann, D. P. Hallahan, & P. C. Pullen (Eds.), *Handbook of special education, 2nd edition* (chap. 12). New York, NY: Routledge.

Pustejovsky, J. E., Hedges, L. V., & Shadish, W. R. (2014). Design-comparable effect sizes in multiple baseline designs: A general modeling framework. *Journal of Educational and Behavioral Statistics*, *39*(5), 368–393. doi: 10.3102/1076998614547577

Pustejovsky, J. E., & Swan, D. M. (2015). Four Methods for Analyzing Partial Interval Recording Data, with Application to Single-Case Research. *Multivariate Behavioral Research*, *50*(3), 365–380. doi: 10.1080/00273171.2015.1014879

Pustejovsky, J. E., & Swan, D. M. (2017). Singlecasees: A calculator for single-case effect size indices [Computer software manual]. Retrieved from https://github.com/jepusto/SingleCaseES (R package version 0.3)

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rindskopf, D. (2014). Nonlinear Bayesian analysis for single case designs. *Journal of School Psychology*, *52*(2), 179–189. doi: 10.1016/j.jsp.2013.12.003

Rogosa, D., & Ghandour, G. (1991). Statistical Models for Behavioral Observations.

Journal of Educational Statistics, *16*(3), 157. doi: 10.2307/1165191

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The Quantitative Synthesis of Single-Subject Research Methodology and Validation. *Remedial and Special Education*, *8*(2), 24–33. doi: 10.1177/074193258700800206

Shadish, W. R. (2014). Statistical analyses of single-case designs: The shape of things to come. *Current Directions in Psychological Science*, *23*(2), 139–146. doi: 10.1177/0963721414524773

Shadish, W. R., Hedges, L. V., Horner, R. H., & Odom, S. L. (2015). The role of between-case effect size in conducting, interpreting, and summarizing single-case research. ncer 2015-002. *National Center for Education Research.*

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*, *18*(3), 385–405. doi: 10.1037/a0032964

Shadish, W. R., Rindskopf, D. M., Hedges, L. V., & Sullivan, K. J. (2013). Bayesian estimates of autocorrelations in single-case designs. *Behavior Research Methods*, *45*(3), 813–821. doi: 10.3758/s13428-012-0282-1

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods*, *43*(4), 971–980. doi: 10.3758/s13428-011-0111-y

Sidik, K., & Jonkman, J. N. (2006). Robust variance estimation for random effects meta-analysis. *Computational Statistics & Data Analysis*, *50*(12), 3681–3701. doi: 10.1016/j.csda.2005.07.019

Solomon, B. G. (2014). Violations of assumptions in school-based single-case data: Implications for the selection and interpretation of effect sizes. *Behavior Modification*, *38*(4), 477–496. doi: 10.1177/0145445513510931

Swaminathan, H., Rogers, H. J., & Horner, R. H. (2014). An effect size measure and bayesian analysis of single-case designs. *Journal of School Psychology*, *52*(2),

213–230.

Swan, D. M., & Pustejovsky, J. E. (2017). *gem_scd: A web-based calculator for the gradual effects model.* Retrieved from https://jepusto.shinyapps.io/gem-scd (Web application)

Swan, D. M., & Pustejovsky, J. E. (2018, Jan). *A gradual effects model for single-case designs.* Open Science Framework. Retrieved from osf.io/gaxrv doi: 10.17605/OSF.IO/GAXRV

Tarlow, K. R. (2016). An improved rank correlation effect size statistic for single-case designs: Baseline corrected Tau. *Behavior Modification*, 1–41. doi: 10.1177/0145445516676750

Thorne, S., & Kamps, D. (2008). The effects of a group contingency intervention on academic engagement and problem behavior of at-risk students. *Behavior Analysis in Practice*, *1*(2), 12–18.

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

Table 1

Estimates and standard errors for inappropriate behavior from Thorne and Kamps (2008) based on a change-in-levels model and the gradual effects model

	Change-in-levels model		Gradual effects model			
Case	LRR Est.	SE	LRR Est.	SE	ω	σ^2
Participant 1	-1.22	0.25	-1.91	0.21	0.62	1.47
Participant 2	-1.91	0.24	-2.25	0.22	0.45	1.36
Participant 3	-0.65	0.14	-0.74	0.15	0.35	0.87
Participant 4	-1.17	0.18	-1.38	0.18	0.44	2.18
Participant 5	-1.13	0.20	-1.54	0.25	0.64	1.96
Participant 6	-0.94	0.13	-1.21	0.11	0.49	0.23
Participant 7	-0.63	0.15	-0.70	0.16	0.35	0.66
Participant 8	-0.94	0.21	-1.19	0.24	0.53	0.95
Participant 9	-0.60	0.14	-0.77	0.18	0.55	0.70
Participant 10	-0.94	0.13	-1.31	0.13	0.63	0.53
Participant 11	-0.75	0.12	-0.85	0.18	0.35	0.99
Participant 12	-1.50	0.27	-2.38	0.23	0.66	0.66
Random effects meta-analysis	-0.99	0.10	-1.34	0.16		

Table 2

Meta-analytic results using different values of m in the gradual effects model for inappropriate behavior data from Thorne and Kamps (2008)

m	LRR Est.	SE	Between-case variance
5	-1.27	0.15	0.230
10	-1.34	0.16	0.285
15	-1.34	0.17	0.290

Table 3

Parameter values used to generate simulated data

Parameter	Levels
Baseline frequency (e^{β_0})	5, 15, 25
Treatment effect β_1	-1.6 to 1.6, in steps of 0.4
Delay parameter (ω)	0.0, 0.3, 0.6, 0.9
autocorrelation (ϕ)	0.0, 0.2, 0.4
Observations per phase (n_p)	3, 5, 10

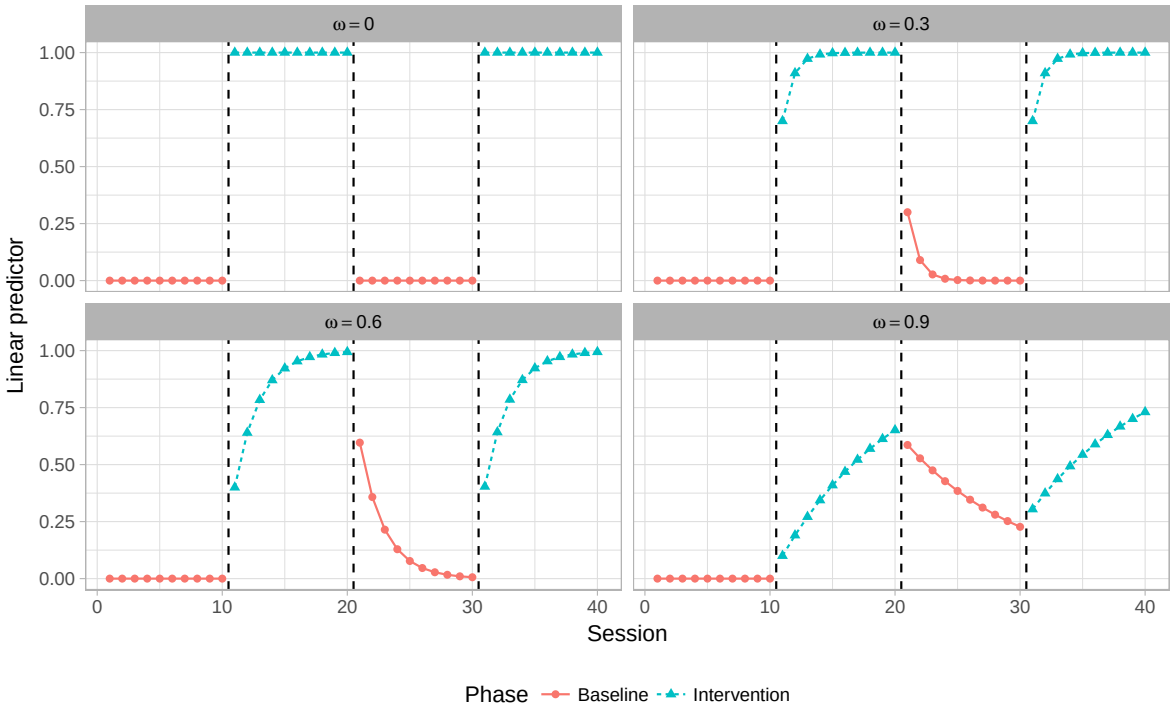

Figure 1. Functional specification of the gradual effects model for differing values of omega, where $\beta_0 = 0$, the equilibrium treatment effect is $\beta_1 = 1$, and $m = \infty$. Each plot depicts an ABAB design with ten sessions per phase.

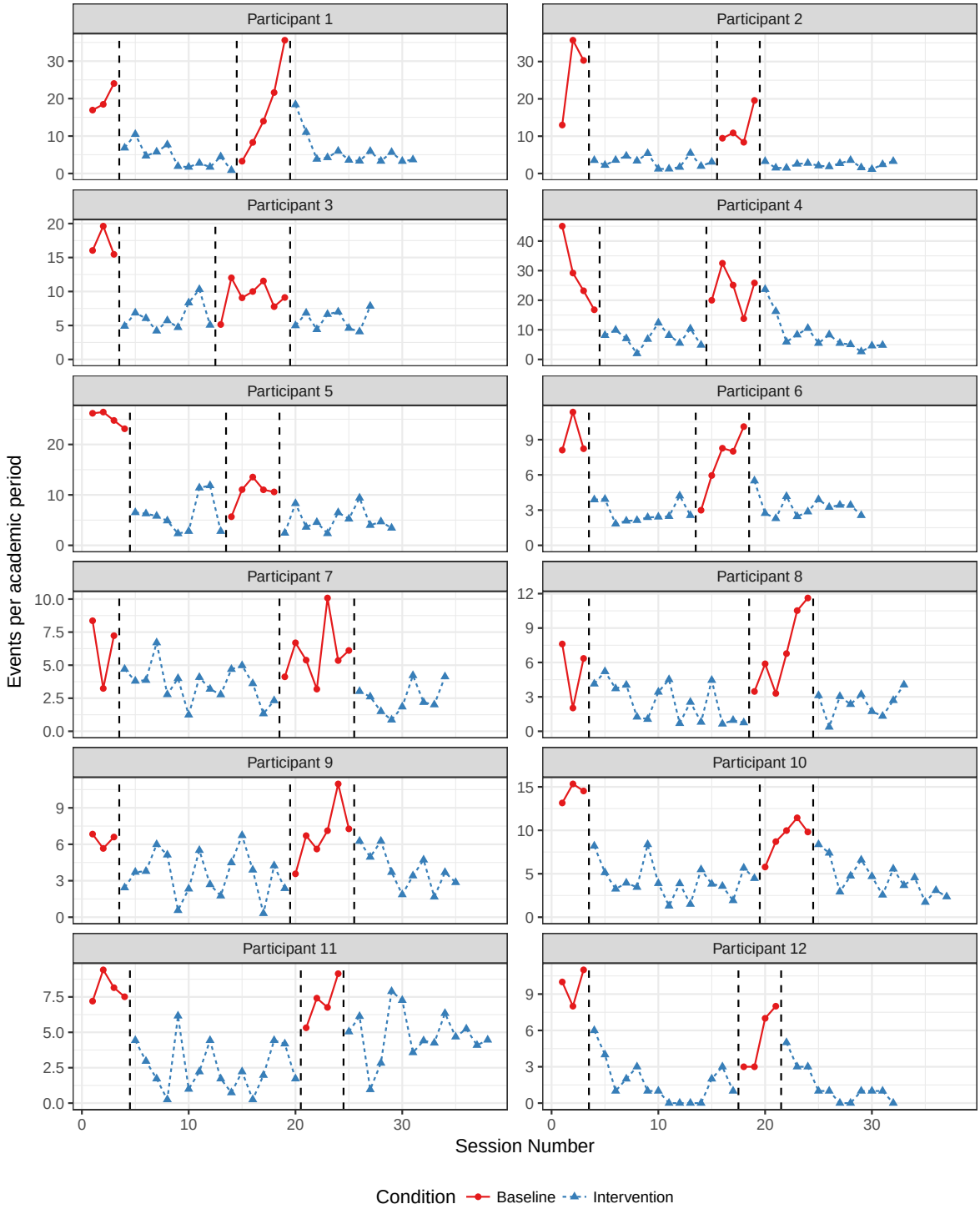

Figure 2. Rates of inappropriate behavior from Thorne and Kamps (2008). Each point represents the number of problem behaviors (vertical axis) during a given observation session (horizontal axis). Point shape and color correspond to treatment condition.

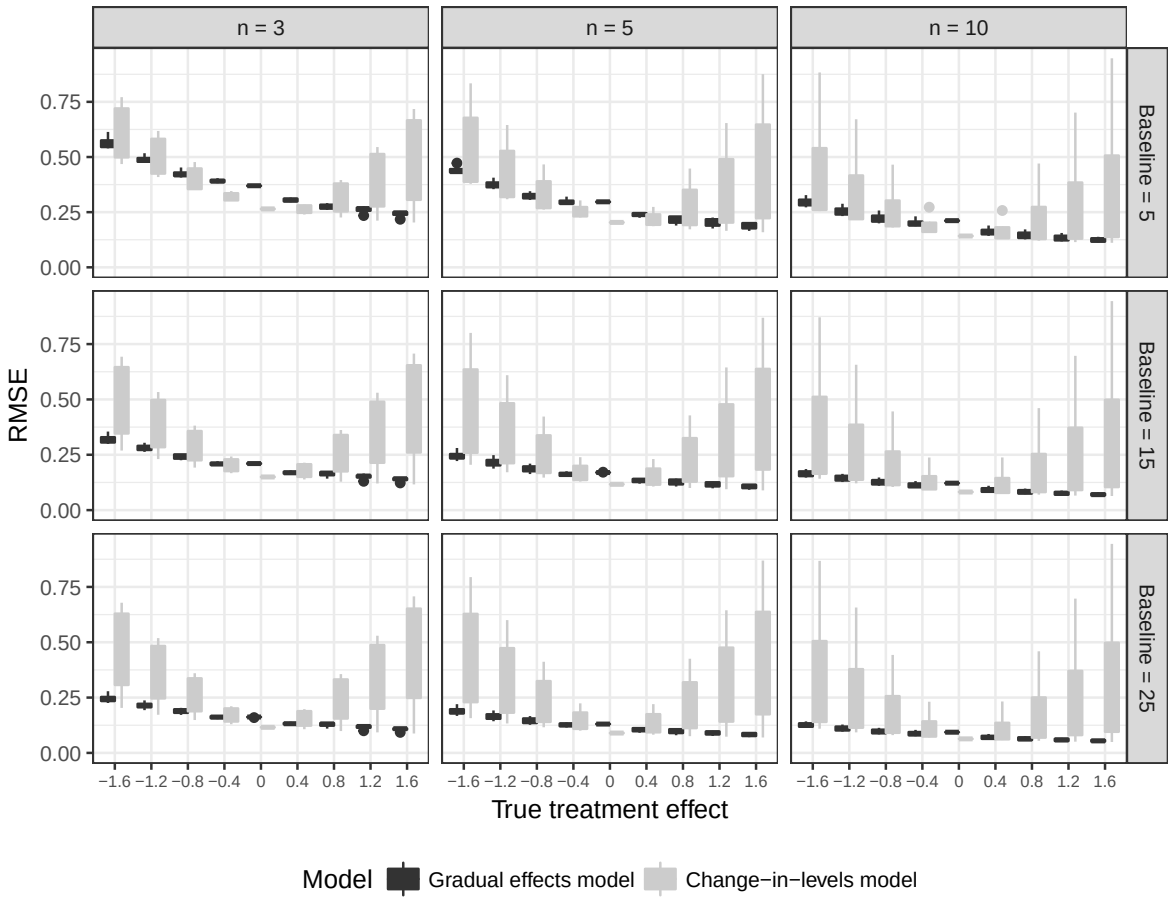

Figure 3. Root mean square error of the treatment effect estimate for the gradual effects model and the change-in-levels model when the outcomes are independent.

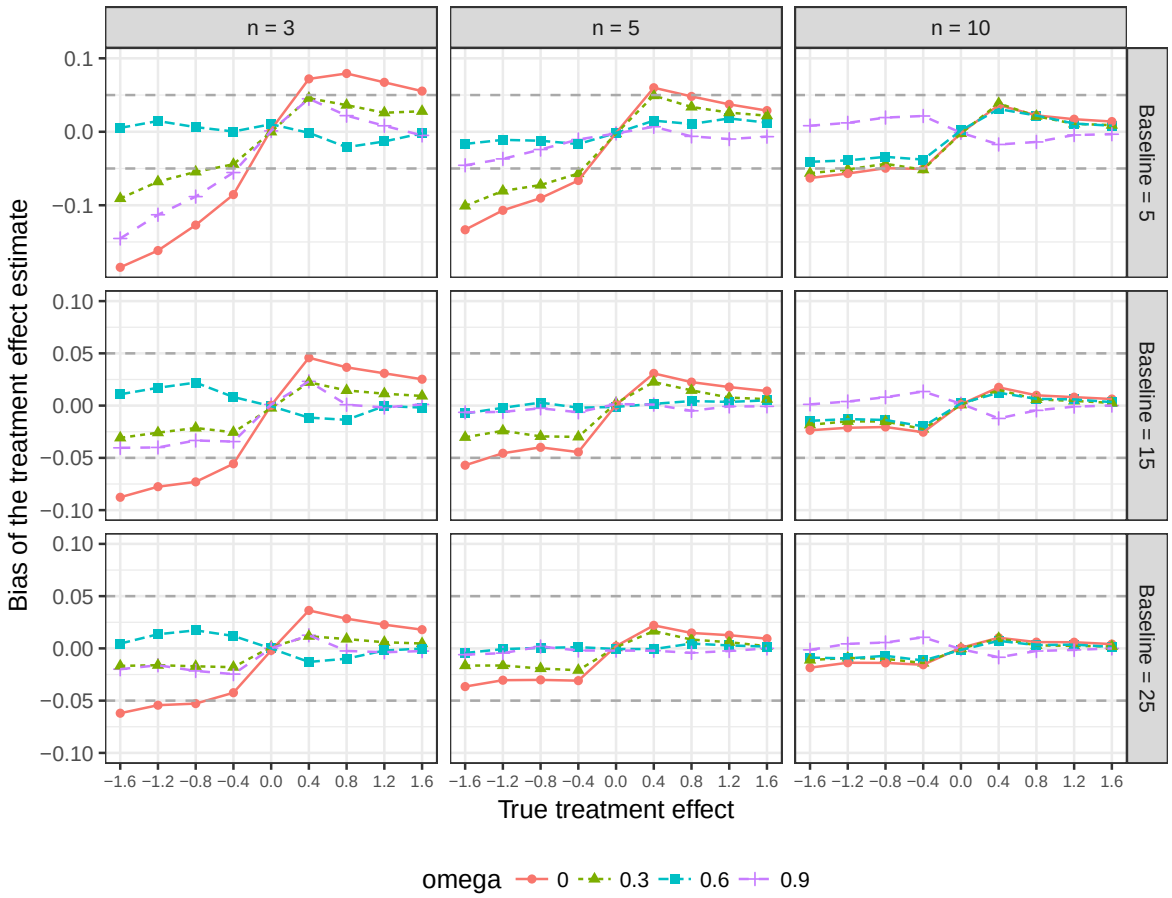

Figure 4. Bias of the treatment effect estimates from the gradual effects model when the outcomes are independent.

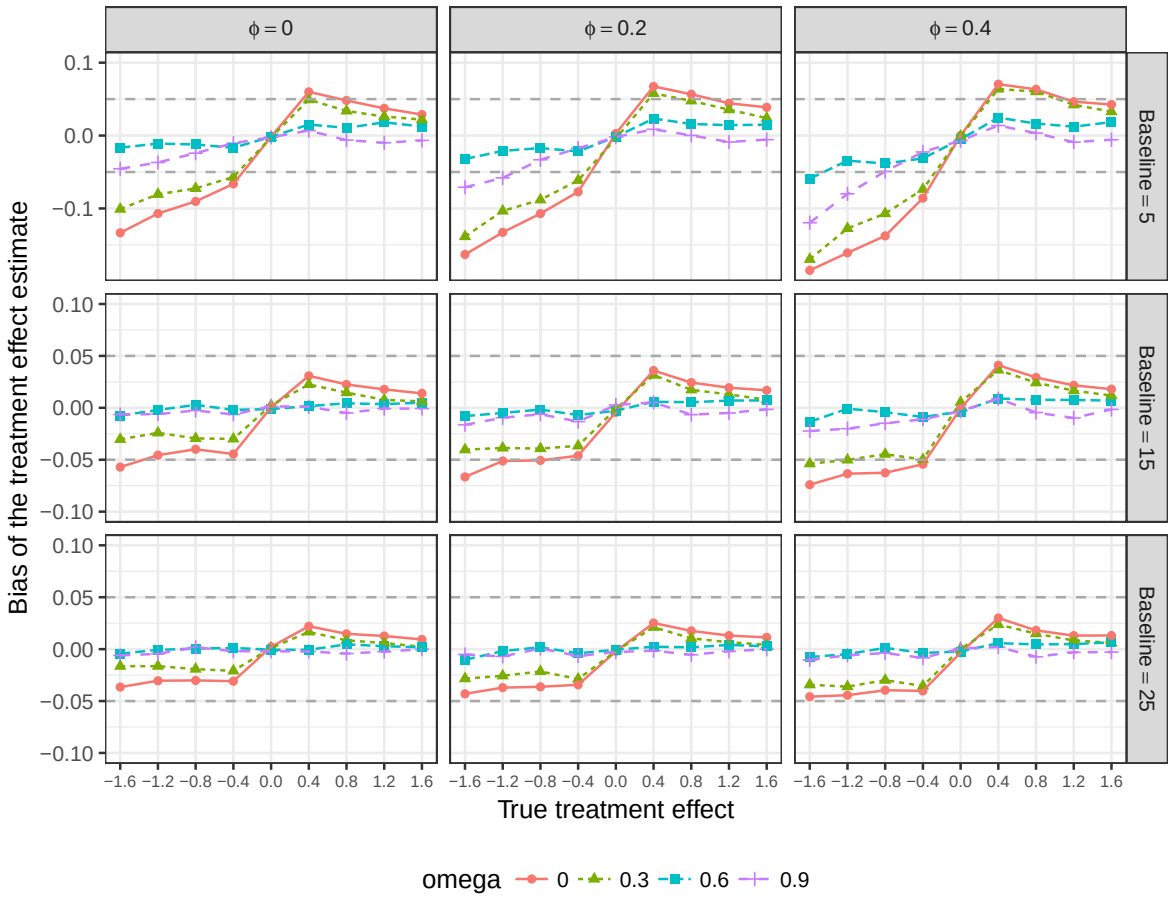

Figure 5. Bias of the treatment effect estimates from the gradual effects model when $n_p = 5$, for varying degrees of of autocorrelation

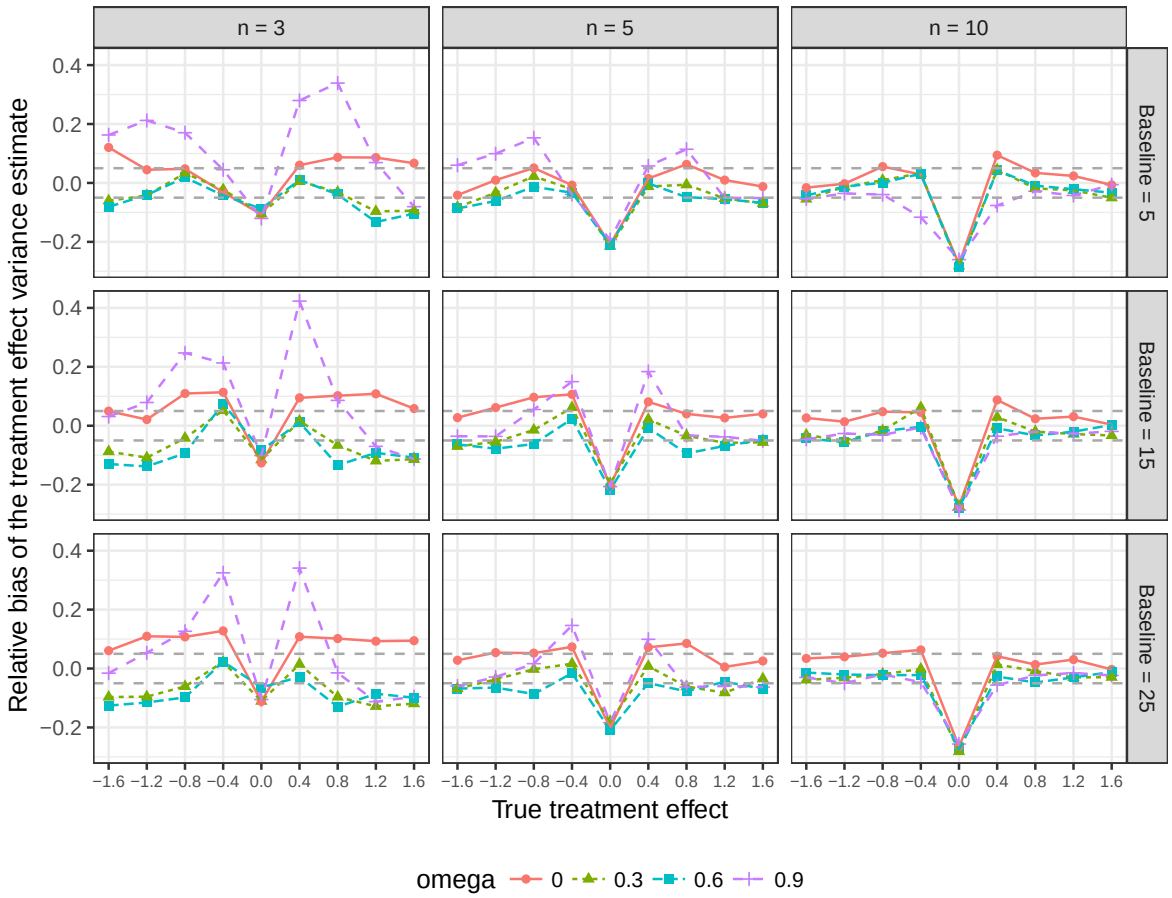

Figure 6. Relative bias of the treatment effect variance estimates from the gradual effects model when the outcomes are independent.

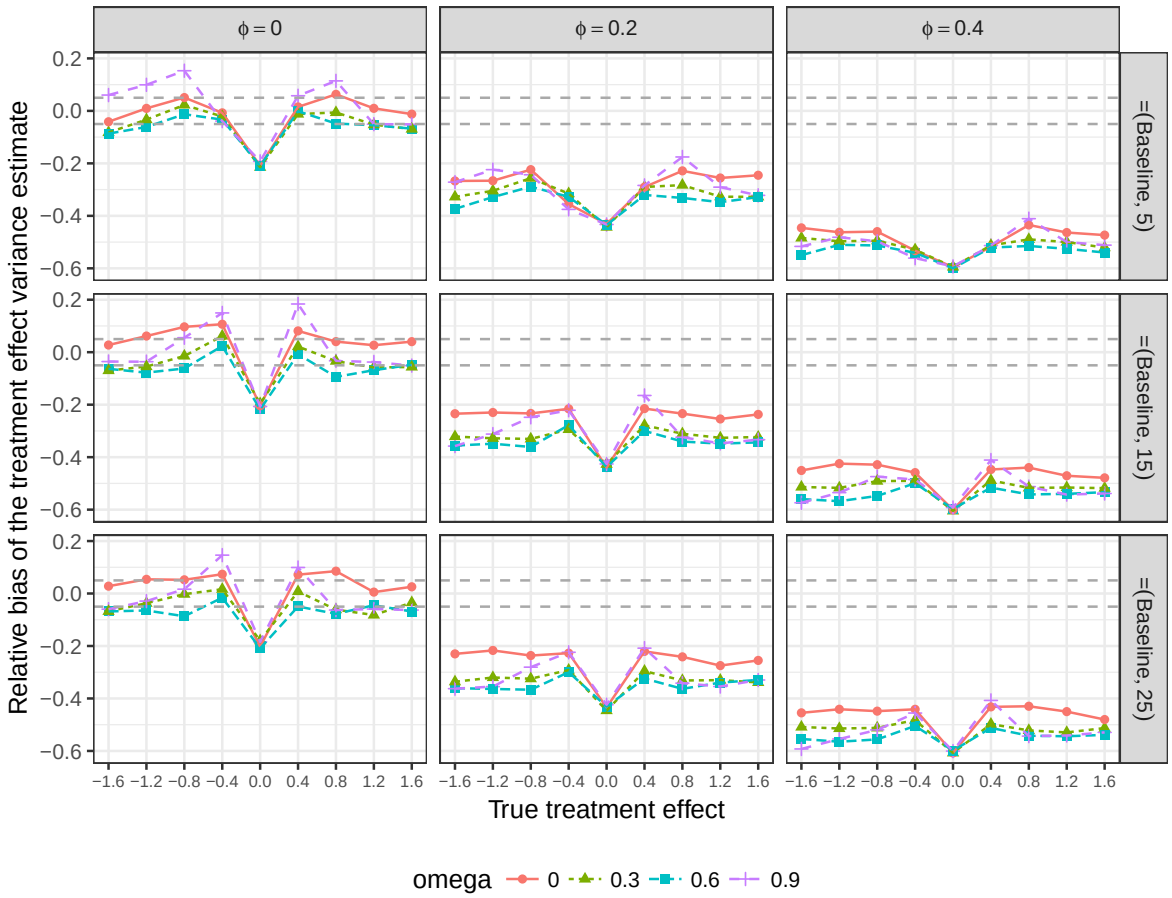

Figure 7. Relative bias of the treatment effect variance estimates from the gradual effects model when $n_p = 5$, for varying degrees of autocorrelation.

Appendix A

Maximum quasi-likelihood estimation

Maximum quasi-likelihood estimates of the parameters in the gradual effects model can be obtained as follows. Given a link function $g()$ and variance function $V()$, the log quasi-likelihood contribution for a measurement Y with expectation μ is given by

$$Q(\mu; Y) = \int_Y^\mu \frac{y - t}{V(t)} dt \tag{4}$$

(McCullagh & Nelder, 1989). Note that μ is determined by the functional specification η, $\mu = g^{-1}(\eta)$, and the functional specification is in turn determined by the model parameters β_0, β_1, and ω. Thus, we can treat the log quasi-likelihood contribution as a function of the model parameters, writing $Q(\beta_0, \beta_1, \omega; Y)$. Let $\mathbf{Y}$ denote the full vector of n measurements. Assuming the measurements are mutually independent, the complete log quasi-likelihood is then

$$Q(\beta_0, \beta_1, \omega; \mathbf{Y}) = \sum_{j=1}^{n} Q(\beta_0, \beta_1, \omega; Y_j). \tag{5}$$

We seek parameter estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\omega}$ that maximize (5).

The functional specification of the gradual effects model can be written as a linear function of the parameters β_0 and β_1 and a non-linear term. Let $\mathbf{T}$ denote the full vector of n treatment indicators and

$$f_j(\mathbf{T}, \omega) = \frac{1 - \omega}{1 - \omega^m} \sum_{i=1}^{j} \omega^{j-i} T_i.$$

Then $\eta = \beta_0 + \beta_1 f_j(\mathbf{T}, \omega)$. Setting the delay parameter to a fixed value ω^*, the functional specification becomes fully linear in β_0 and β_1. We can therefore find maximum quasi-likelihood estimators $\hat{\beta}_0(\omega^*)$ and $\hat{\beta}_1(\omega^*)$ using iteratively re-weighted least squares, as implemented in conventional software such as the `glm()` function in R (R Core Team, 2017). The overall maximum quasi-likelihood estimators can then be identified by maximizing the profile quasi-likelihood, $Q\left(\hat{\beta}_0(\omega^*), \hat{\beta}_1(\omega^*), \omega^*\right)$ as a function of ω^*. In the simulations, we used the `optimize` command in R to find the value of ω^* that maximizes the quasi-likelihood.

Finally, the scale parameter σ^2 is estimated as

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{j=1}^{n}\frac{Y_j - \hat{\mu}_j}{V\left(\hat{\mu}_j\right)} \tag{6}$$

where $\hat{\mu}_j = g^{-1}(\hat{\eta}_j)$ and $\hat{\eta}_j = \hat{\beta}_0 + \hat{\beta}_1 f_j(\mathbf{T}, \hat{\omega})$.

The sampling variance-covariance matrix of the parameter estimators can be approximated using a method described by McCullagh and Nelder (1989). Let $\mathbf{x}_j$ be a 1×3 row vector with entries

$$\mathbf{x}_j = \begin{bmatrix} 1 & f_j(\mathbf{T}, \hat{\omega}) & \hat{\beta}_1 f_j'(\mathbf{T}, \hat{\omega}) \end{bmatrix},$$

where $f_j'(\mathbf{T}, \hat{\omega})$ is the derivative of f_j with respect to ω:

$$f_j'(\mathbf{T}, \hat{\omega}) = \frac{m\omega^{m-1}(1-\omega)}{(1-\omega^m)^2}\sum_{i=1}^{j}\omega^{j-i}T_i - \frac{1}{1-\omega^m}\left[j\omega^{j-1}T_1 + \sum_{i=2}^{j}(j-i+1)\omega^{j-i}\left(T_i - T_{i-1}\right)\right].$$

Let $g'(\eta)$ denote the derivative of the link function with respect to η, and let $w_j = V\left(\hat{\mu}_j\right)\left[g'(\hat{\eta}_j)\right]^2$. The variance-covariance matrix of $\left(\hat{\beta}_0, \hat{\beta}_1, \hat{\omega}\right)$ is then estimated as

$$\mathbf{V} = \hat{\sigma}^2\left(\sum_{j=1}^{n}\frac{\mathbf{x}_j\mathbf{x}_j'}{w_j}\right)^{-1}, \tag{7}$$

with standard errors given by the square root of the diagonal terms in $\mathbf{V}$.

Appendix B

Simulating autocorrelated Poisson outcomes

The Monte Carlo simulations used the following process to simulate autocorrelated Poisson outcomes. Let ϕ denote the first-order autocorrelation between pairs of sequential observations, with $0 \leq \phi < 1$. For a series with a total of n observations, and given the treatment pattern $\mathbf{T}$ and parameter values β_0, β_1, and ω, we first calculated the expectation of each observation, $\mu_1, ..., \mu_n$, using the functional specification of the gradual effects model, as given in Equation (2), and a log link function. Let $\phi_j = \min\{\phi, \mu_j/\mu_{j-1}\}$ and $\lambda_j = \mu_j - \phi_j\mu_{j-1}$ for $j = 2, ..., n$. We then simulated Y_1 from a Poisson distribution with mean μ_1. For the remaining observations, $j = 2, ..., n$, we simulated Z_j from a Poisson distribution with mean λ_j, simulated X_j from a Binomial distribution with Y_{j-1} trials and probability ϕ_j, and calculated $Y_j = X_j + Z_j$. The resulting observations are each Poisson-distributed (marginally) and serially correlated when $\phi > 0$. So long as the proportionate change in means between two sequential observations is less than $\phi - 1$, the correlation between sequential observations will be $\text{cor}(\epsilon_j, \epsilon_{j+1}) = \phi$ and $\text{cor}(\epsilon_j, \epsilon_{j+k}) = \phi^k$. If there is a large reductions in the mean from one occasion to the next (i.e., $\mu_j/\mu_{j-1} < \phi$), then the correlations will not exactly follow the first-order auto-regressive structure, but will still be serially dependent. This is acceptable because our interest is in the robustness of the gradual effects model estimates in the presence of un-modeled serial dependence, rather than in estimating a model with a specific serial dependence structure.