Supplementary materials for "Procedural sensitivities of effect sizes for single-case designs with directly observed behavioral outcome measures"

James E. Pustejovsky

December 6, 2017

Contents

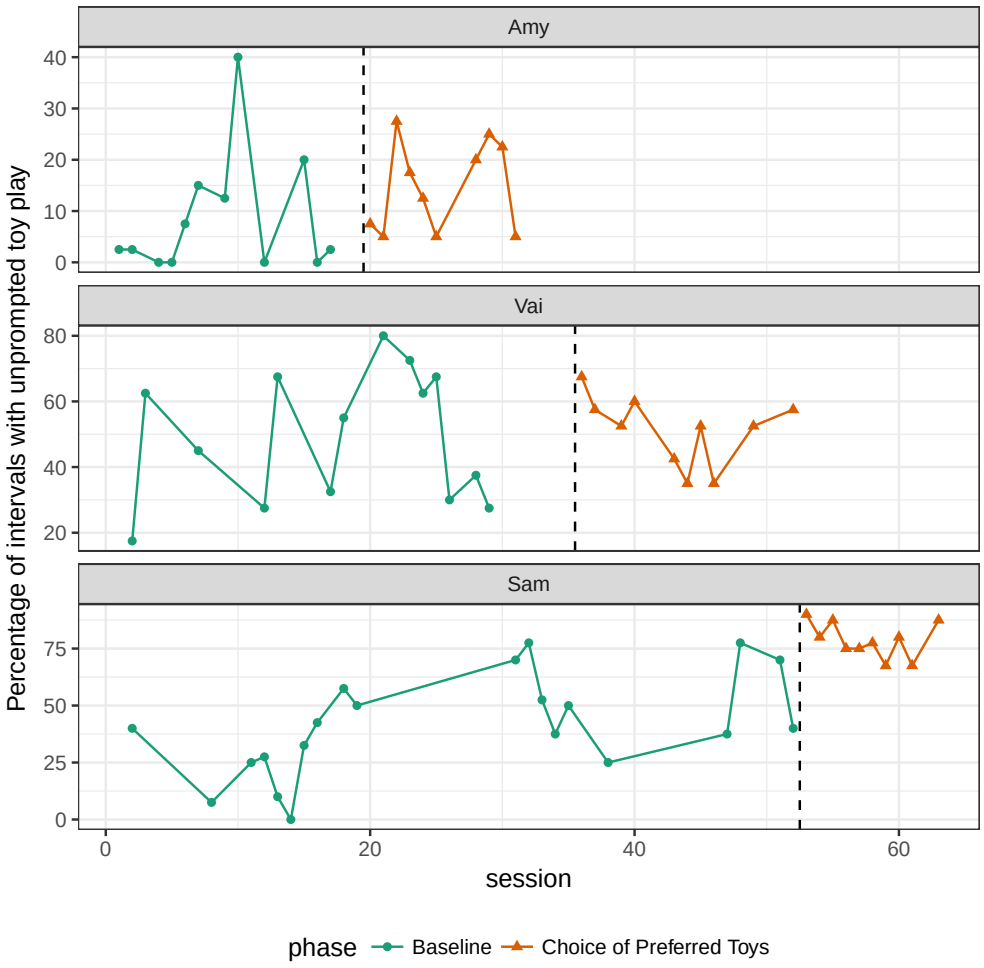

Figure S1: Data on rates of unprompted toy play from DiCarlo, Reid, & Stricklen (2003).

S1 Empirical example of SCD effect size calculations

This section provides numerical examples of each of the effect size indices described in the main text, using real empirical data from a single-case study. DiCarlo, Reid, and Stricklin (2003) used an across-participant multiple probe design to evaluate the effects of providing choice among preferred toys on the rates of play behavior for three young children with disabilities. Rates of unprompted toy play were assessed using a 15 s partial interval recording system over 10 min observation sessions. The baseline phases included between 12 and 20 observation sessions, while the treatment phases consisted of 10 observation sessions for each participant. Figure S1 depicts the study design and outcome data for each of the three participants. For readers wishing to compute the effect size indices for themselves, Table S1 reports the values of the raw outcome data from each phase, for each of the three participants.

Table S2 reports the six non-overlap measures (NOMs) for each of the three cases from DiCarlo et al. (2003). It is interesting to note that different NOMs yield different orderings of the magnitude

Table S1: Raw data and summary statistics from DiCarlo, Reid, & Stricklen (2003). Percentage of intervals with unprompted toy play during baseline and treatment phase observation sessions for each of three participants.

	Amy		Vai		Sam	
	Baseline	Treatment	Baseline	Treatment	Baseline	Treatment
1	2.5	7.5	17.5	67.5	40.0	90.0
2	2.5	5.0	62.5	57.5	7.5	80.0
3	0.0	27.5	45.0	52.5	25.0	87.5
4	0.0	17.5	27.5	60.0	27.5	75.0
5	7.5	12.5	67.5	42.5	10.0	75.0
6	15.0	5.0	32.5	35.0	0.0	77.5
7	12.5	20.0	55.0	52.5	32.5	67.5
8	40.0	25.0	80.0	35.0	42.5	80.0
9	0.0	22.5	72.5	52.5	57.5	67.5
10	20.0	5.0	62.5	57.5	50.0	87.5
11	0.0		67.5		70.0	
12	2.5		30.0		77.5	
13			37.5		52.5	
14			27.5		37.5	
15					50.0	
16					25.0	
17					37.5	
18					77.5	
19					70.0	
20					40.0	
Mean	8.54	14.75	48.93	51.25	41.50	78.75
SD	11.99	8.85	20.18	10.69	22.32	7.93
n	12	10	14	10	20	10

of effects across the three cases. Amy and Vai both have PND values of zero (the minimum possible), while Sam has a larger PND that would still be classified as a "questionable" effect by extant benchmarks. In contrast, Amy and Sam have maximal values for PEM, while Vai has a lower PEM of 70%. The PAND values for the three cases range from 62.5% to 86.7% and the IRD values range from 0.229 to 0.700. Following the benchmarks proposed by Parker, Vannest, and Brown (2009), the RIRD value for Vai would be classified as "questionable," that for Amy as "medium," and that for Sam as "large." However, these classifications are themselves questionable given that the minimum possible value and null value of RIRD depend on the number of observations in each phase, which varies across the three cases. NAP values range from 51.4%—quite close to the expected null value—for Vai to 93.5% for Sam. Following the benchmarks proposed by Parker and Vannest (2009), the NAP values yield similar classifications as for RIRD. The values for Tau are simply a linear transformation of NAP, and so share the same interpretation.

Table S3 reports estimates for the within-case standardized mean difference (SMD, denoted as d), the bias-corrected within-case SMD (g), the raw log-response ratio (LRR, denoted as R_1), and

Table S2: Non-overlap measures for DiCarlo, Reid, & Stricklen (2003) data.

case	PND	PEM	PAND	RIRD	NAP	Tau
Amy	0.0	100.0	77.3	0.542	75.4	0.508
Vai	0.0	70.0	62.5	0.229	51.4	0.029
Sam	50.0	100.0	86.7	0.700	93.5	0.870

Table S3: Parametric effect size estimates for DiCarlo, Reid, & Stricklen (2003) data.

case	d	g	R_1	R_2
Amy	0.518	0.485	0.546	0.482
Vai	0.115	0.109	0.046	0.042
Sam	1.669	1.605	0.641	0.634

the bias-corrected log response ratio (R_2) for each of the three cases from DiCarlo et al. (2003). All of these parametric effect sizes can be calculated directly from the summary statistics in Table S1. For these data, bias correction makes little difference for the SMD or the LRR because the phases are all fairly long (each containing 10 or more observations). Following the benchmarks proposed by Harrington and Velicer (2015), the SMD values for Amy and Vai would be classified as "small," while the SMD for Sam would be classified as "medium." The LRR values for Amy, Vai, and Sam correspond to percentage increases in play behavior of 73%, 5%, and 90%, respectively.

S2 SCD effect sizes that account for time trends

The review of SCD effect sizes in the main text was limited to indices that are appropriate for data without systematic time trends, or what I shall call "basic" effect sizes. However, extensions to many of the basic indices have been proposed that account for certain forms of time trend during the baseline phase. In this section, I briefly review these extensions, demonstrate their relationship to the basic effect sizes reviewed in the main text, and consider why the simulation findings regarding basic effect sizes are likely to generalize to the indices that account for time trends.

As in the main text, I assume that the outcome is defined so that decreases are therapeutically desirable. I consider an effect size estimate based on the data from a single baseline phase with m observations and a single treatment phase with n observations. Denote the outcome measurements during the baseline phase as $y_1^A, ..., y_m^A$ and the outcome measurements during the treatment phase as $y_1^B, ..., y_n^B$. Note that the order of the measurements now matters, and I will assume that observations are ordered temporally within each phase.

The main technique used to account for baseline time trends in most of the following effect size indices involves first estimating a linear trend from the baseline phase data. This trend is then projected throughout the treatment phase, and the residuals around the trend line are computed. Finally, these residuals are used to compute a non-overlap measure or parametric effect size index.

A model for the baseline phase data that includes a linear time trend is given by:

$$y_t^A = \alpha + \beta t + \epsilon_t, \tag{1}$$

for $t = 1, ..., m$. Different effect size indices use different estimators of the linear trend coefficient β, but such differences are of secondary interest here. Let $\hat{\beta}$ denote some (generic) estimate of β. Let $e_t^A = y_t^A - \hat{\beta}t$ for $t = 1, ..., m$ denote the residuals around the linear trend during the baseline phase. Let $e_t^B = y_t^B - \hat{\beta}(t + m)$ denote the residuals from projecting the same linear trend into the treatment phase.

S2.1 SCD effect sizes based on residuals

I now show that many of the extant effect sizes for SCDs that account for time trends are computed using the formula for a corresponding basic effect size, but substituting the baseline and treatment phase residuals in place of the raw outcomes.

Manolov and Solanas (2009) proposed the percentage of non-overlapping corrected data (PNCD), an extension of the percentage of non-overlapping data that accounts for baseline time trends. In this approach, β is estimated by the average difference between adjacent observations in the baseline phase:

$$\hat{\beta}_{diff} = \frac{1}{m-1} \sum_{i=2}^{m} (y_i - y_{i-1}).$$

The effect size index is then calculated as PND of the residuals:

$$\text{PNCD} = 100\% \times \frac{1}{n} \sum_{i=1}^{n} I\left(e_i^B < e_{(1)}^A\right), \tag{2}$$

where $e_{(1)}^A = \min\left\{e_1^A, ..., e_m^A\right\}$.

Wolery, Busick, Reichow, and Barton (2010) proposed the percentage exceeding the median trend (PEM-T) index, an extension to percentage exceeding the median (PEM) for handling baseline time trends. PEM-T uses the split-middle technique, or extended celeration line, to estimate a baseline time trend. Let $\tilde{y}_1^A$ denote the median of the first half of the baseline phase observations, $y_1^A, ..., y_{\lceil m/2 \rceil}^A$, and let $\tilde{y}_2^A$ denote the median of the second half of the baseline phase observations, $y_{\lfloor m/2 \rfloor}^A, ..., y_m^A$. The split-middle estimator of β is then $\hat{\beta}_{split} = 2\left(\tilde{y}_2^A - \tilde{y}_1^A\right)/m$. PEM-T is then calculated as the percentage of treatment phase observations that improve upon the median trend line. With an outcome where decrease is therapeutically desirable, the observation y_i^B represents an improvement over the median trend line if $e_i^B < \tilde{e}_A$, where $\tilde{e}_A = \text{median}\left(e_1^A, ..., e_m^A\right)$ is the median of the residuals from the baseline phase. Thus, PEM-T is calculated as

$$\text{PEM-T} = 100\% \times \frac{1}{n} \sum_{i=1}^{n} \left[I\left(e_i^B < \tilde{e}_A\right) + 0.5 I\left(e_i^B = \tilde{e}_A\right)\right]. \tag{3}$$

Just as with PEM, treatment phase observations that are exactly equal to the median trend are counted with a weight of one half.

To my knowledge, no extensions for handling time trends have been proposed for the PAND, RIRD, or NAP indices.

Tarlow (2017) proposed the baseline-corrected Tau index, an extension to the Tau index (Parker, Vannest, Davis, & Sauber, 2011) that corrects for linear baseline trend using a Theil-Sen estimator. Let $q_{ij} = (y_j^A - y_i^A)/(j-i)$ for $i = 1, ..., m-1$ and $j = i+1, ..., m$. The Theil-Sen slope estimator is defined as the median of the $m(m-1)/2$ values of q_{ij}, i.e., $\hat{\beta}_{TS} = \text{median}\{q_{12}, q_{13}, ..., q_{m-1,m}\}$. Adjusted residuals, calculated based on $\hat{\beta}_{TS}$, are then summarized using the Tau index, given by

$$\text{Tau}_{BC} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left[I\left(e_j^B < e_i^A\right) - I\left(e_j^B > e_i^A\right)\right]. \tag{4}$$

Tarlow (2017) proposed that in practice, the statistical significance of the baseline trend should be tested as a preliminary step. Following this approach, baseline-corrected Tau would be applied if the trend is statistically distinct from zero based on Kendall's rank correlation test, while the original, unadjusted form of Tau would be applied if the trend is not statistically significant.

Maggin, Swaminathan, et al. (2011) proposed an extension of the within-case SMD that accounts for time trends in the baseline and treatment phases, defined in terms of the parameters of a piecewise linear regression model. Let $y_t = y_t^A$ for $t = 1, ..., m$ and $y_t = y_{t-m}^B$ for $t = m+1, ..., m+n$. Let z_t be an indicator for observations in the treatment phase, so that $z_t = 0$ for $t = 1, ..., m$ and $z_t = 1$ for $t = m+1, ..., m+n$. Maggin, Swaminathan, et al. (2011) posit the following regression model:

$$y_t = \beta_0 + \beta_1 t + \beta_2 z_t + \beta_3 z_t \times (t - m - 1) + \epsilon_t,$$

where it is assumed that $\text{E}(\epsilon_t) = 0$ and $\text{Var}(\epsilon_t) = \sigma_e^2$. Under this model, the effect size is defined as the average shift in the regression line across points in the treatment phase, scaled by the standard deviation of the residuals:

$$\delta_{pw} = \left(\beta_2 + \beta_3\frac{n-1}{2}\right)/\sigma_e. \tag{5}$$

The numerator of this effect size is estimated by replacing the regression coefficients with corresponding generalized least-squares estimates, assuming that the errors follow a first-order autoregressive process. The denominator is estimated using the root mean squared residuals, yielding the estimate

$$d_{pw} = \left(\hat{\beta}_2 + \hat{\beta}_3\frac{n-1}{2}\right)/\hat{\sigma}_e.$$

Although less obvious than with the non-overlap indices, the Maggin, Swaminathan, et al. (2011) effect size estimate can also be expressed as a standardized mean difference of adjusted residuals. Note that the regression residuals from the baseline phase are equivalent to $e_t = (e_t^A - \hat{\beta}_0)$, $t = 1, ..., m$. For purposes of illustration, I will assume that σ_e is estimated using the residuals from the baseline phase only (in practice, Maggin proposed to use an estimate that is pooled across

phases), so

$$\hat{\sigma}_e^2 = \frac{1}{m-1}\sum_{t=1}^{m} e_t^2 = \frac{1}{m-1}\sum_{t=1}^{m}\left(e_t^A - \frac{1}{m}\sum_{i=1}^{m} e_i^A\right)^2.$$

Furthermore, because fitted regression lines pass through the mean of the data,

$$\bar{e}_A = \frac{1}{m}\sum_{t=1}^{m} e_t^A = \hat{\beta}_0 + \hat{\beta}_1\frac{t+1}{2} \qquad \text{and} \qquad \bar{e}_B = \frac{1}{n}\sum_{t=1}^{m} e_t^B = \hat{\beta}_0 + \hat{\beta}_1\frac{t+1}{2} + \hat{\beta}_2 + \hat{\beta}_3\frac{n-1}{2}.$$

It follows that

$$d_{pw} = \frac{\bar{e}_B - \bar{e}_A}{\sqrt{\frac{1}{m-1}\sum_{t=1}^{m}(e_t^A - \bar{e}_A)^2}}.$$

Thus, Maggin's effect size estimator is equivalent to the within-case standardized mean difference calculated from the adjusted residuals.

S2.2 Remarks

Given the close relationship between the basic effect size indices and the above extensions for handling time trends, findings from the simulation study of basic effect sizes are likely to generalize to the extent that the adjusted residuals are affected by procedural factors in the same way that the raw outcomes are affected. It is plausible that this will be the case under many data-generating processes. For instance, consider a design with a very long baseline phase, so that the baseline trend can be estimated precisely, and further suppose that the time trend is very slight. The adjusted residuals $e_t^A = y_t^A - \hat{\beta}t$ and $e_t^B = y_t^B - \hat{\beta}_{t+m}$ will then be nearly identical to the raw outcomes. Procedural factors that influence the variability of the raw outcomes will thus also influence the adjusted residuals, and so effect sizes calculated from those residuals will retain the same procedural sensitivities as the corresponding basic indices.

The main difference between an effect size index calculated from adjusted residuals and one calculated from raw outcomes is that the residuals will have some additional variability due to the fact that $\hat{\beta}$ must be estimated from the data. The degree of additional variability will depend on the precision of $\hat{\beta}$ and consequently will be influenced by m, the number of observations in the baseline phase. Thus, effect size indices calculated from adjusted residuals are likely to be influenced to some extent by the number of observations in the baseline phase. In short, it is reasonable to expect that the effect size indices will have *additional* procedural sensitivities, as well as inheriting the procedural sensitivities of the corresponding basic effect size, as identified in the simulation study.

S2.3 Tau-U

Parker, Vannest, Davis, and Sauber (2011) proposed Tau-U, a family of several effect size indices that extend the Tau index to account for time trends in the baseline phase, treatment phase, or both. Although several different versions of Tau-U have been described, only the form that involves adjustment for baseline time trends appears to be in common use. I therefore confine the remaining

discussion to that version of the index.

Unlike the effect size indices described in Section S2.1, Tau-U is not a residualized version of the Tau index. Instead, Tau-U extends Tau using an additive correction term for baseline time trend. Let r_A denote Kendall's rank correlation between the outcomes and measurement occasions in the baseline phase, calculated as

$$r_A = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \left[I\left(y_j^A < y_i^A\right) - I\left(Y_j^A > y_i^A\right) \right].$$

In this context, r_A measures the degree of monotonic trend (rather than linear trend) in the baseline phase observations. The baseline-corrected Tau-U index is then calculated as

$$\text{Tau-U} = \text{Tau} - \frac{(m-1)}{2n} r_A.$$

Thus, Tau-U differs from Tau by an additive term that depends on the number of observations in the baseline and treatment phases and on the degree of monotonic trend in the baseline phase. Given this additive relationship, it is reasonable to infer that Tau-U will be procedurally sensitive to the duration of observation sessions and to the recording system used to measure outcomes, in the same way that Tau is sensitive to these factors. Furthermore—and unlike the Tau index—it is also reasonable to infer that the magnitude of Tau-U will be influenced by the number of observations in the baseline and treatment phases because the additive correction factor depends on both m and n. This prediction is also borne out in simulations conducted by Tarlow (2017), which demonstrated that, under simple parametric models with time trends, the magnitude of Tau-U changes depending on the number of observations in each phase.

S3 Calculating effect sizes for SCDs with multiple pairs of phases

The effect sizes reviewed in the main text were all defined in terms of a single pair of phases (i.e., a baseline phase and a treatment phase). However, SCDs involving multiple phase contrasts are also common in practice. For instance, ABAB designs involve four phases: an initial baseline phase (A1), an initial treatment phase (B1), a return-to-baseline phase (A2) and a treatment re-introduction phase (B2). Other treatment reversal designs may involve further return-to-baseline and treatment re-introduction phases (i.e., A3, B3, etc.). In this section, I briefly review approaches for calculating effect sizes from such designs and explain why these approaches will share the same procedural sensitivities as effect sizes calculated from a single pair of phases.

In the context of systematic reviews and meta-analyses of SCDs, researchers have taken a variety of approaches to calculating effect size indices for SCDs involving multiple phase contrasts. Based on a systematic review of quantitative syntheses of SCDs, Maggin, O'Keeffe, and Johnson (2011) identified three common strategies for handling designs with multiple pairs of phases. First, some researchers select a single pair of phases that best represents the functional relationship of interest.

For example, Heyvaert, Saenen, Campbell, Maes, and Onghena (2014) computed summary effect sizes from the initial baseline and final treatment phase, while Heath, Ganz, Parker, Burke, and Ninci (2015) used the initial baseline phase and initial treatment phase only, arguing that this phase contrast was most directly comparable to the baseline and treatment phases of a multiple baseline design. Second, some researchers pool the data across phases that share the same treatment condition (e.g., White, Rusch, Kazdin, & Hartmann, 1989). For instance, in an ABAB design, this would entail pooling the initial baseline phase (A1) with the return-to-baseline phase (A2) and pooling the initial treatment phase (B1) with the treatment re-introduction phase (B2), thereby simplifying the design to a single phase contrast. Third and finally, some researchers calculate separate effect sizes for each phase contrast within a design and then average the results. For instance, in an ABAB design, the analyst might calculate effect size indices for the A1-B1 contrast, B1-A2 contrast, and A2-B2 contrast, then average these together into a single summary effect size for the case (e.g., Maggin, Chafouleas, Goddard, & Johnson, 2011).

The first of these approaches involves calculating effect sizes from a single pair of phases, while the second approach involves treating the design as if it included only a single pair of phases. Thus, the findings of the simulation study in the main text would apply directly. With the third approach, it is reasonable to infer that effect sizes which are procedurally sensitive for a single phase-pair will remain so when averaged together because procedural sensitivity is a function of the expected value of the effect size index. Thus, averaging together two or more effect sizes that share a common expected value will not change that value.

To see why this is the case, consider an example of two SCD studies, each of which uses a treatment reversal design with multiple pairs of phases. Suppose that the studies are exact replications except that all phases in the first study include 5 observation sessions, while all phases in the second study include 12 observation sessions. The two studies are exact replications, and so any differences in the expected value of an effect size index represent procedural sensitivity. Let T_{pq} denote the effect size index calculated from phase comparison p in study q, for $p = 1, ..., P$ and $q = 1, 2$. If one focused only on the initial phase pair within each study, the extent of procedural sensitivity with respect to the number of observations per phase would be $\mathrm{E}(T_{12} - T_{11}) = \mathrm{E}(T_{12}) - \mathrm{E}(T_{11})$. If one instead averaged across all P phase comparisons within the design, the extent of procedural sensitivity would be:

$$\mathrm{E}\left[\frac{1}{P}\sum_{p=1}^{P} T_{p2} - \frac{1}{P}\sum_{p=1}^{P} T_{p1}\right] = \frac{1}{P}\sum_{p=1}^{P} \mathrm{E}(T_{p2} - T_{p1}) = \mathrm{E}(T_{12} - T_{11}).$$

Thus, the degree of procedural sensitivity is unchanged by aggregating across multiple phase contrasts. Similar arguments hold for other procedural factors, such as observation session length and observation recording system.

S4 Additional simulation results for state behaviors

This section reports some additional results of the state behavior simulation, the design of which is described in the main text. Figure S2 displays examples of simulated SCDs for each combination of prevalence, and incidence, and change in behavior. These examples were generated using continuous recording with 10 min observation sessions and 10 sessions in each phase. Readers accustomed to visual analysis of SCDs may wish to judge for themselves whether the simulated data resemble the data from real SCDs.

The remainder of this section reports results pertaining to the robust improvement rate difference (RIRD), percentage of all non-overlapping data (PAND), percentage exceeding the median (PEM), standardized mean difference (SMD), and log response ratio (LRR).

S4.1 RIRD

Figure S3 depicts the expected value of RIRD as a function of the number of observations in the baseline phase and in the treatment phase, for the subset of results where continuous recording is used for 10 min sessions and where incidence is once per minute. The top row of the figure indicates that the expected value of RIRD varies between 0.17 and 0.41 when the treatment has no effect. For larger degrees of change between phases, RIRD becomes somewhat less sensitive to the number of observations in each phase; for instance, when treatment produces a 50% change in behavior and prevalence is 50%, the expectation of RIRD ranges from 0.78 to 0.86.

S4.2 PAND

Figure S4 depicts the expected value of PAND as a function of the number of observations in the baseline phase and in the treatment phase, for the subset of results where continuous recording is used for 10 min sessions and where incidence is once per minute; it is constructed in the same way as Figure 3 in the main text. Although Parker and colleagues (Parker, Hagan-Burke, & Vannest, 2007; Parker, Vannest, & Davis, 2011) suggested that 50% is the expected value of PAND when treatment has no effect on the outcome, the top row of the figure indicates that this is not the case. Instead, the expected value varies between 59% and 81% when the treatment has no effect. In contrast to RIRD, PAND tends to be smaller when the number of observations in the baseline phase is equal to the number in the treatment phase. The middle and bottom rows of the figure indicate that PAND becomes less sensitive to sample size when the treatment produces larger effects, though this appears to be mainly because it approaches the ceiling level of 100%.

Figure S5 depicts the expected value of PAND for varying session lengths and recording systems, based on the subset of results where treatment leads to a 50% reduction in behavior and both phases include 10 observation sessions; it is constructed in the same way as Figure 4 in the main text. The degree to which the magnitude of PAND is influenced by the observation session length and the recording system closely parallels the results for RIRD. As with RIRD, the degree of sensitivity depends on the magnitude of the change from baseline to intervention phase. When treatment

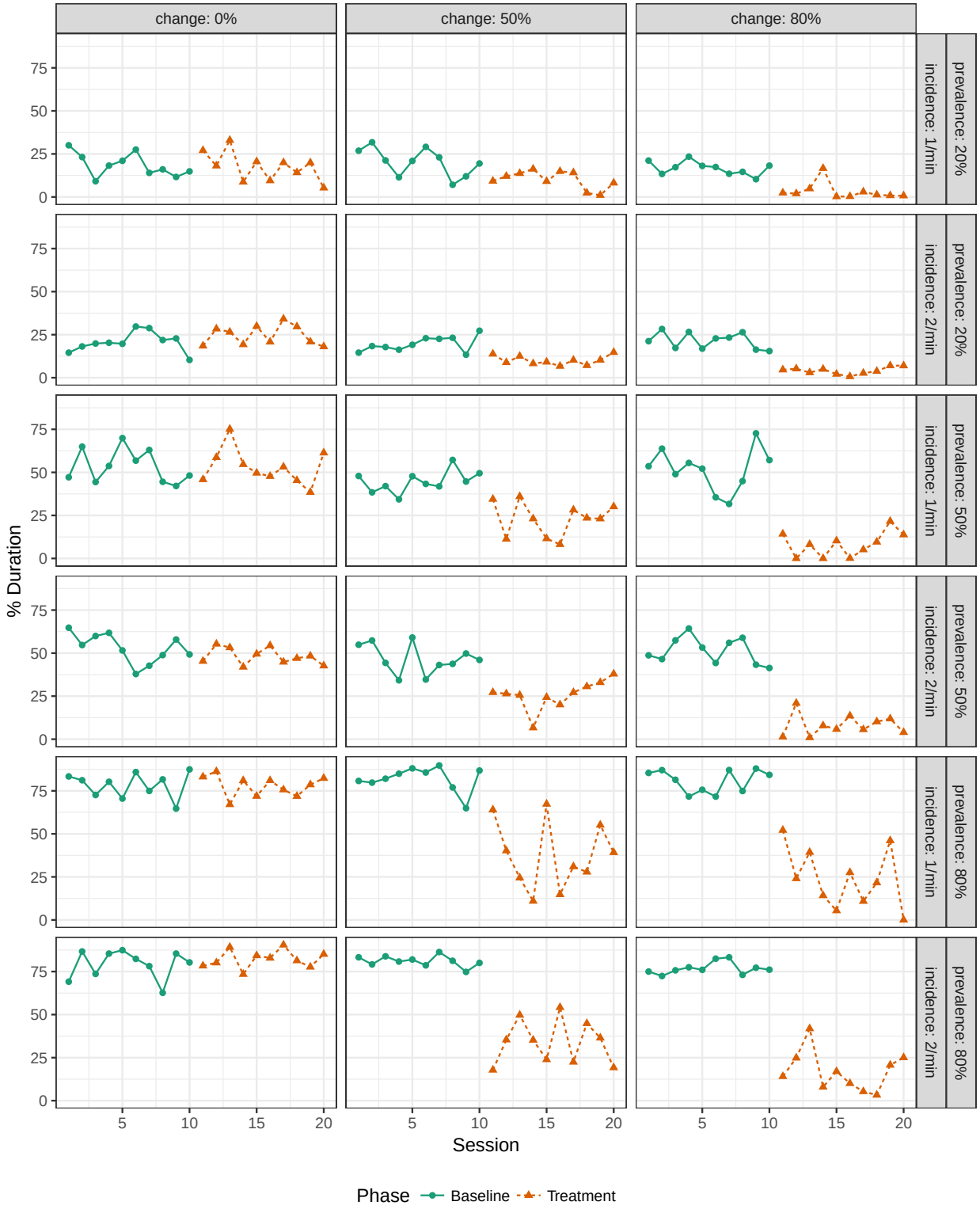

Figure S2: Simulated SCDs based on the alternating renewal process model, using continuous recording for 10 min sessions, for varying levels of prevalence, incidence, and change in behavior.

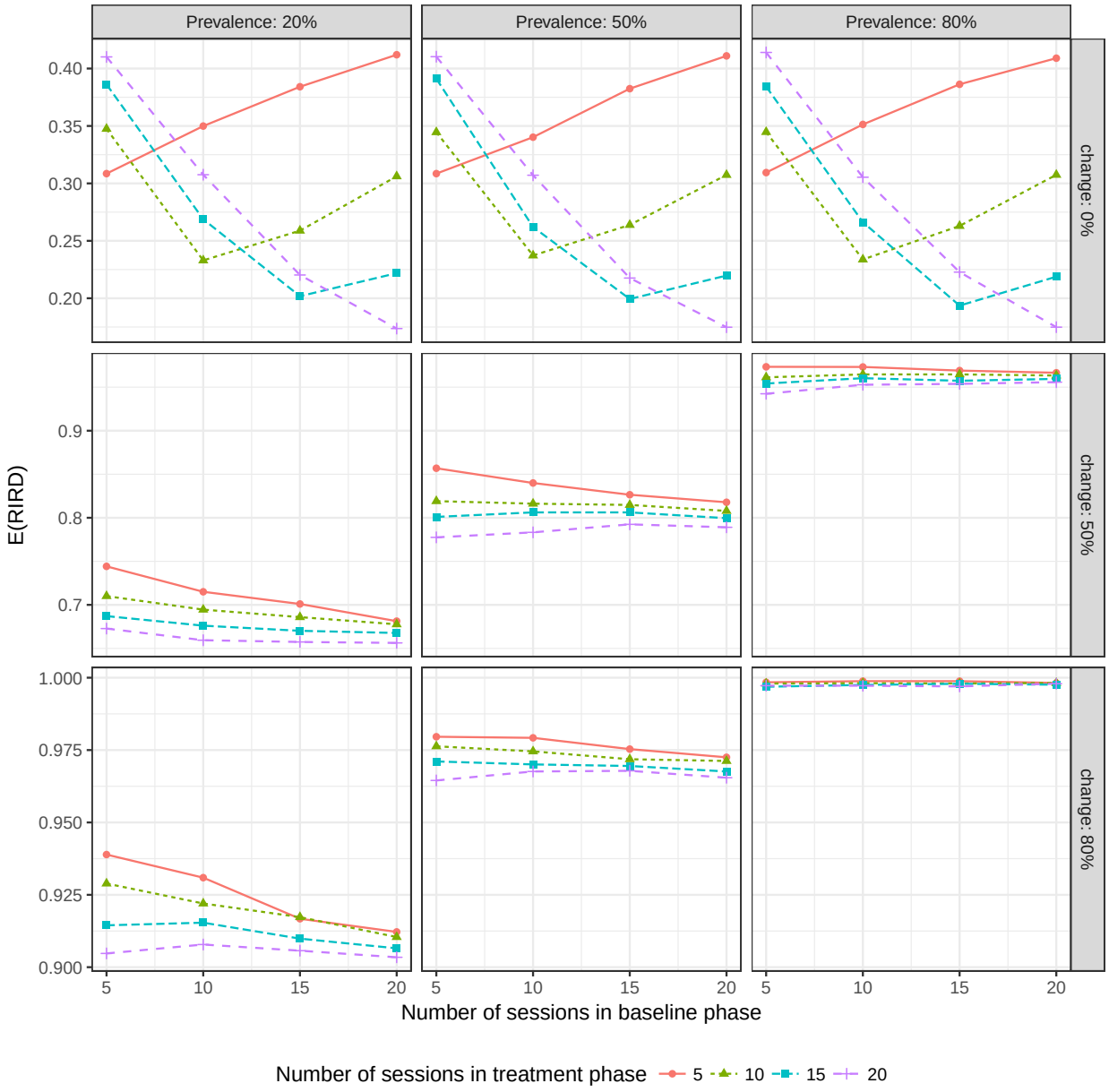

Figure S3: Expected value of RIRD based on continuous recording data with 10 min observation sessions, when incidence is 1/min, for varying numbers of sessions in the baseline and treatment phases.

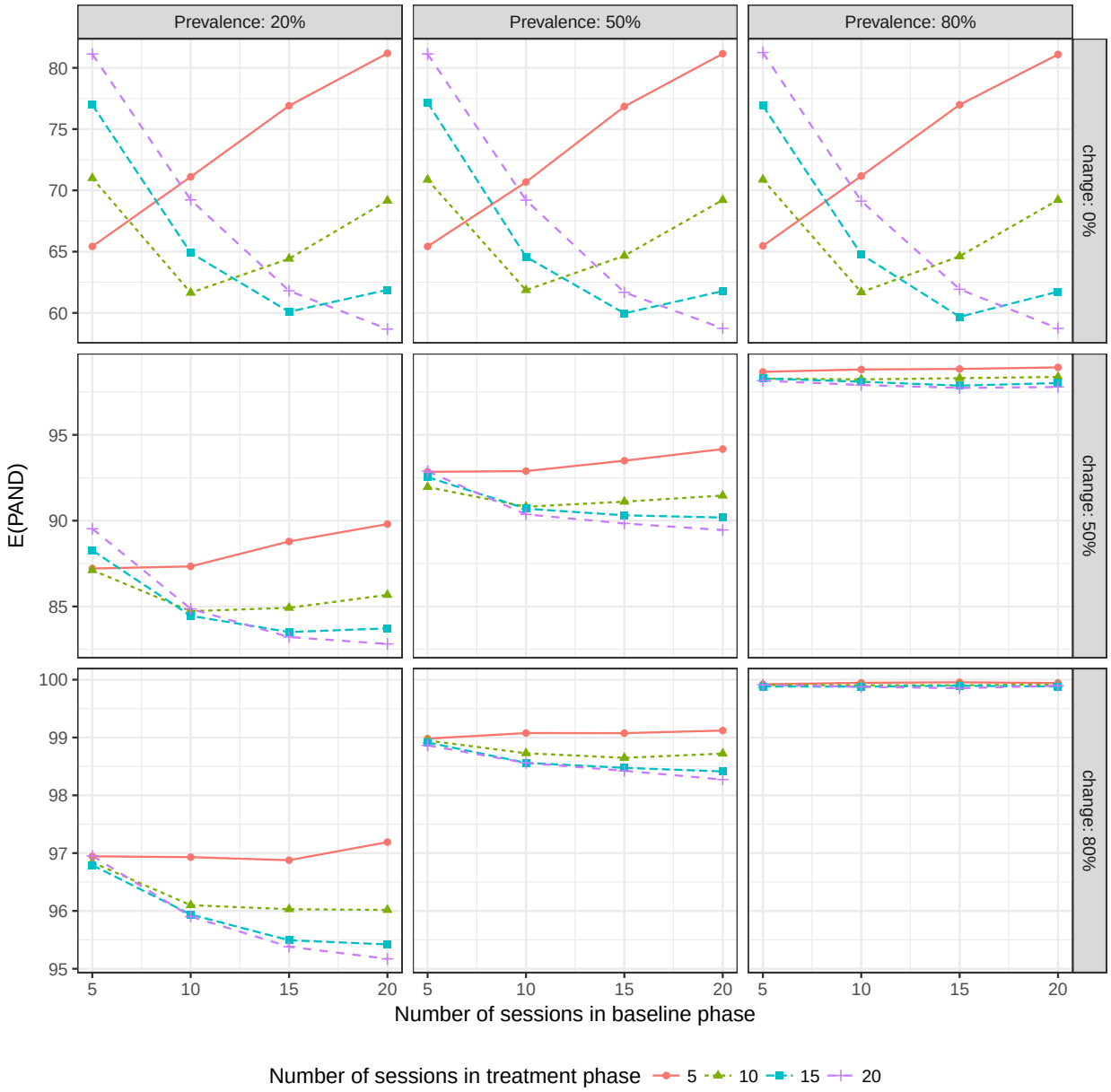

Figure S4: Expected value of PAND based on continuous recording data with 10 min observation sessions, when incidence is 1/min, for varying numbers of sessions in the baseline and treatment phases.

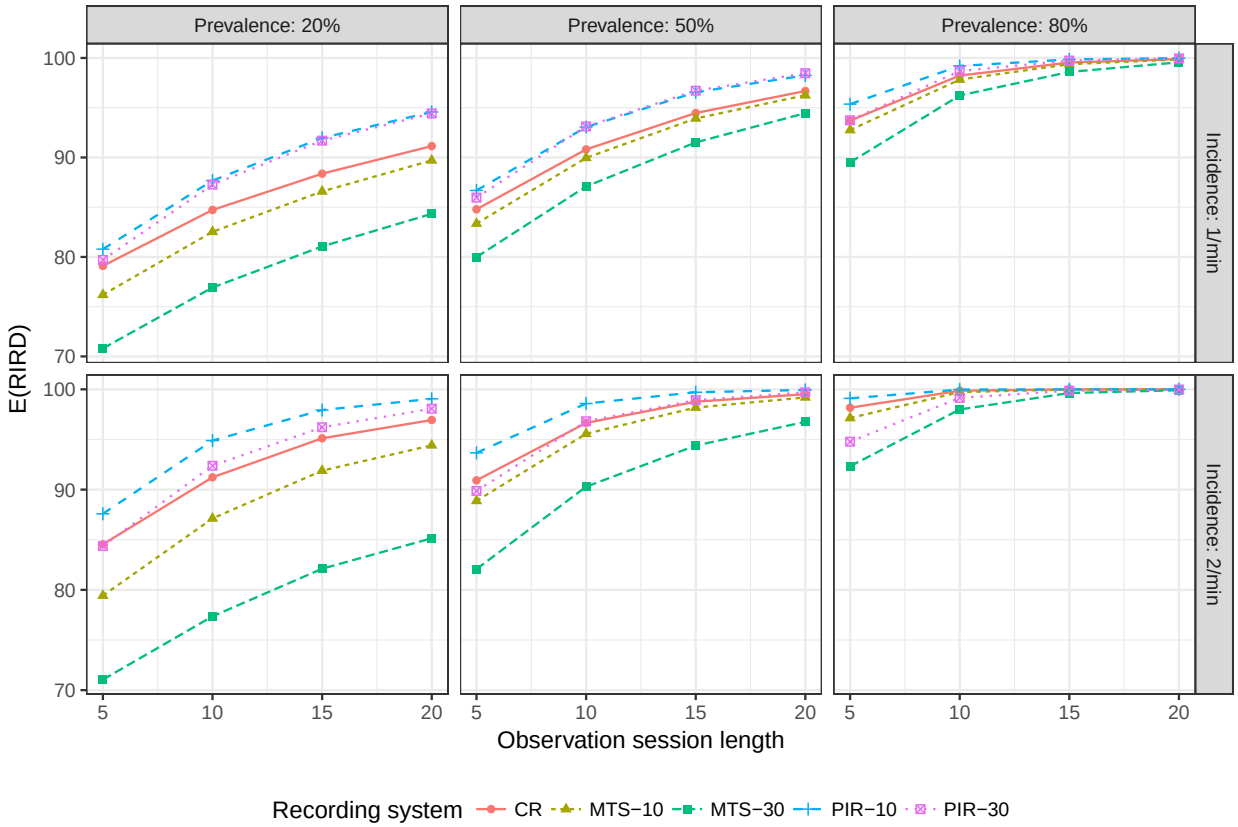

Figure S5: Expected value of PAND for varying session lengths and recording systems, when treatment leads to a 50% change and both phases include 10 sessions. CR = continuous recording; MTS = momentary time sampling; PIR = partial interval recording.

has no effect, PAND is largely unaffected by the length of the observation session. In contrast, when the treatment reduces the prevalence of the behavior by 50%, the expected value of PAND becomes sensitive to the length of the observation session, with longer sessions leading to higher values for PAND. At both 0% and 50% change due to treatment, PAND is also at least somewhat affected by what recording system is used (e.g., continuous recording produces slightly larger values of PAND than MTS with 10 s or 30 s intervals). Finally, when treatment leads to an 80% reduction in the prevalence of the behavior, the expected value of PAND approaches the ceiling level of 100% regardless of the session length. Taken together, the simulation results demonstrate that PAND is sensitive to the number of observations in the baseline and treatment phases, sensitive to observation session length, and at least somewhat sensitive to recording system.

S4.3 PEM

As noted in the main text, the expected value of PEM depends only very weakly on the number of observations in either phase. Furthermore, if treatment has no effect on the behavior then the expected value of PEM is always exactly 50%, regardless of the length of the observation sessions

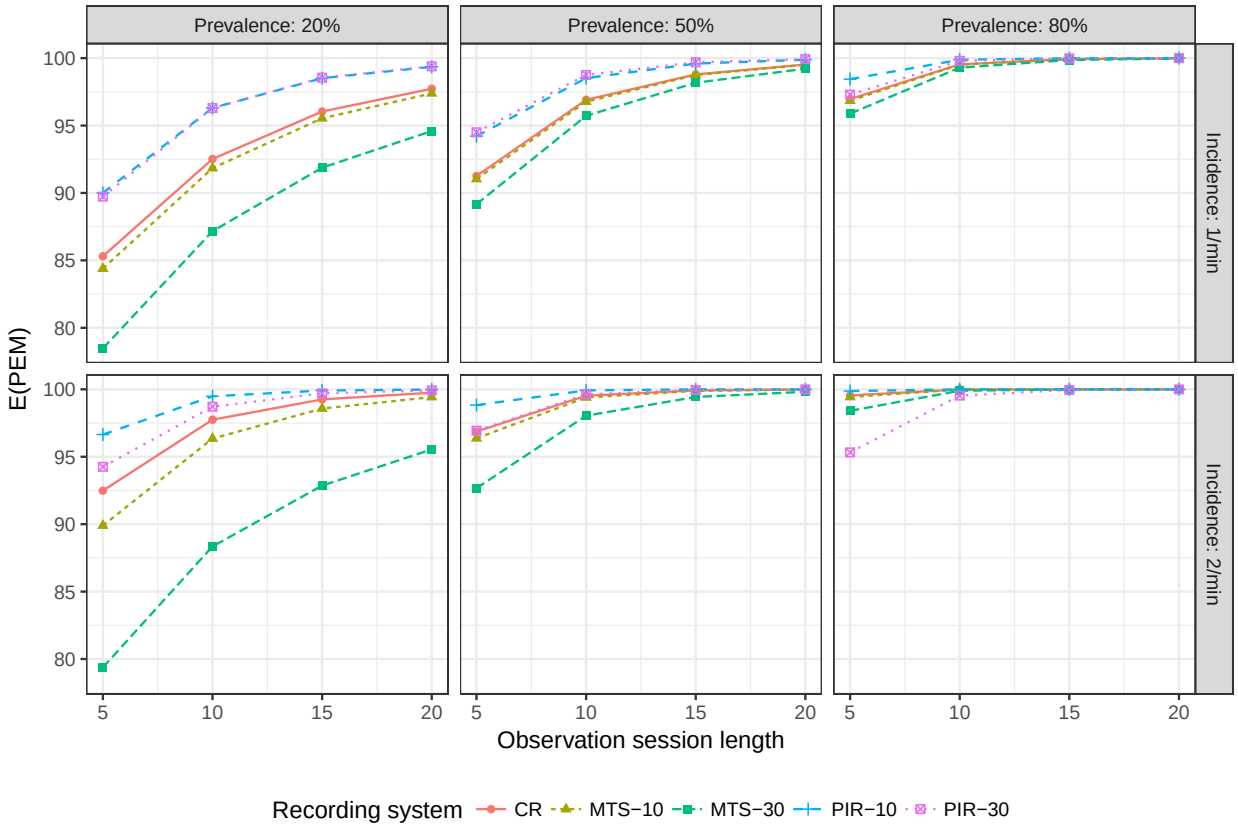

Figure S6: Expected value of PEM for various recording systems and session lengths, when treatment leads to a 50% change in behavior. CR = continuous recording; MTS = momentary time sampling; PIR = partial interval recording.

or of the recording system used to collect outcome data.

To illustrate the sensitivity of PEM to variation in session length and recording system, Figure S6 plots its expected value when treatment leads to a 50% decrease in behavior, for varying session lengths and recording systems; each panel displays results for a different combination of prevalence and incidence during baseline. It can be seen that, for some types of behavior, the magnitude of PEM is sensitive to the length of the observation session and to which recording procedure is used. For instance, for a behavior with baseline prevalence of 20% and baseline incidence of twice per minute, measuring the behavior using 30 s MTS for 5 min sessions would lead to an expected value of 79%, whereas measuring the same behavior using 10 s PIR for 5 min sessions would lead to an expected value of 97%.

Like NAP, the extent to which PEM is sensitive to session length and recording procedure depends on the characteristics of the behavior. Specifically, PEM is less sensitive to session length and recording procedure when the behavior has higher levels of baseline prevalence or baseline incidence. However, this reduced sensitivity appears to be largely due to the fact that PEM is at or near the ceiling level of 100% for all session lengths and recording systems. Thus, for changes in

behavior in the range to which PEM is sensitive, the expected value of this statistic is sensitive to the choice of observation session length and recording system.

S4.4 SMD

As explained in the main text, the expected value of the basic SMD estimator d has a small-sample bias that makes its sensitive to the number of sessions in the baseline phase. Figure S7 depicts the expected value of d when treatment leads to a 50% reduction in behavior and the outcome is measured using 10 min observation sessions. It can be seen that the magnitude of d changes as the number of sessions in baseline increases and that the degree of sensitivity depends on the baseline prevalence of the behavior. Figure S8 depicts the expected value of the bias-corrected SMD estimator g, and is constructed in parallel to Figure S7. For most recording systems, the number of sessions in the baseline phase has very little effect on the magnitude of g, although its magnitude is strongly influenced by the number of baseline sessions when the outcome is measured using 20s or 30 s PIR. This sensitivity is an artifact of the fact that PIR systems overestimate the true prevalence of the behavior, and are thus susceptible to ceiling effects that distort the sampling distribution of g.

S4.5 LRR

As explained in the main text, the expected value of the LRR moment estimator R_1 has a small-sample bias that makes its sensitive to the length of the baseline and treatment phases. Figure S9 illustrates this sensitivity by depicting the expected value of R_1 when incidence is once per minute and the behavior is measured using 30 s MTS for 5 min sessions. It can be seen that the magnitude of R_1 is influenced by the number of observations in each phase when treatment produces non-null changes in behavior. For example, when prevalence is 20% and treatment leads to a 50% decrease, the expected value of R_1 ranges from -0.8 to -0.67; with an 80% decrease, it ranges from -1.8 to -1.63. The degree of sensitivity is lessened when longer session lengths or more precise recording systems (e.g., 10 s MTS or CR) are used.

Figure S10 depicts the expected value of the bias-corrected LRR estimator R_2, and is constructed the same way as Figure S9. Compared to R_1, the degree to which R_2 is sensitive to the length of the baseline or treatment phase is greatly reduced. For example, when prevalence is 20% and treatment leads to a 50% decrease, the expected value of R_1 ranges from -0.7 to -0.67; with an 80% decrease, it ranges from -1.62 to -1.53. Just as with R_1, 30 S MTS represents a worst-case scenario for R_2; its sensitivity to the number of observations in each phase is further reduced when longer session lengths or more precise recording systems are used.

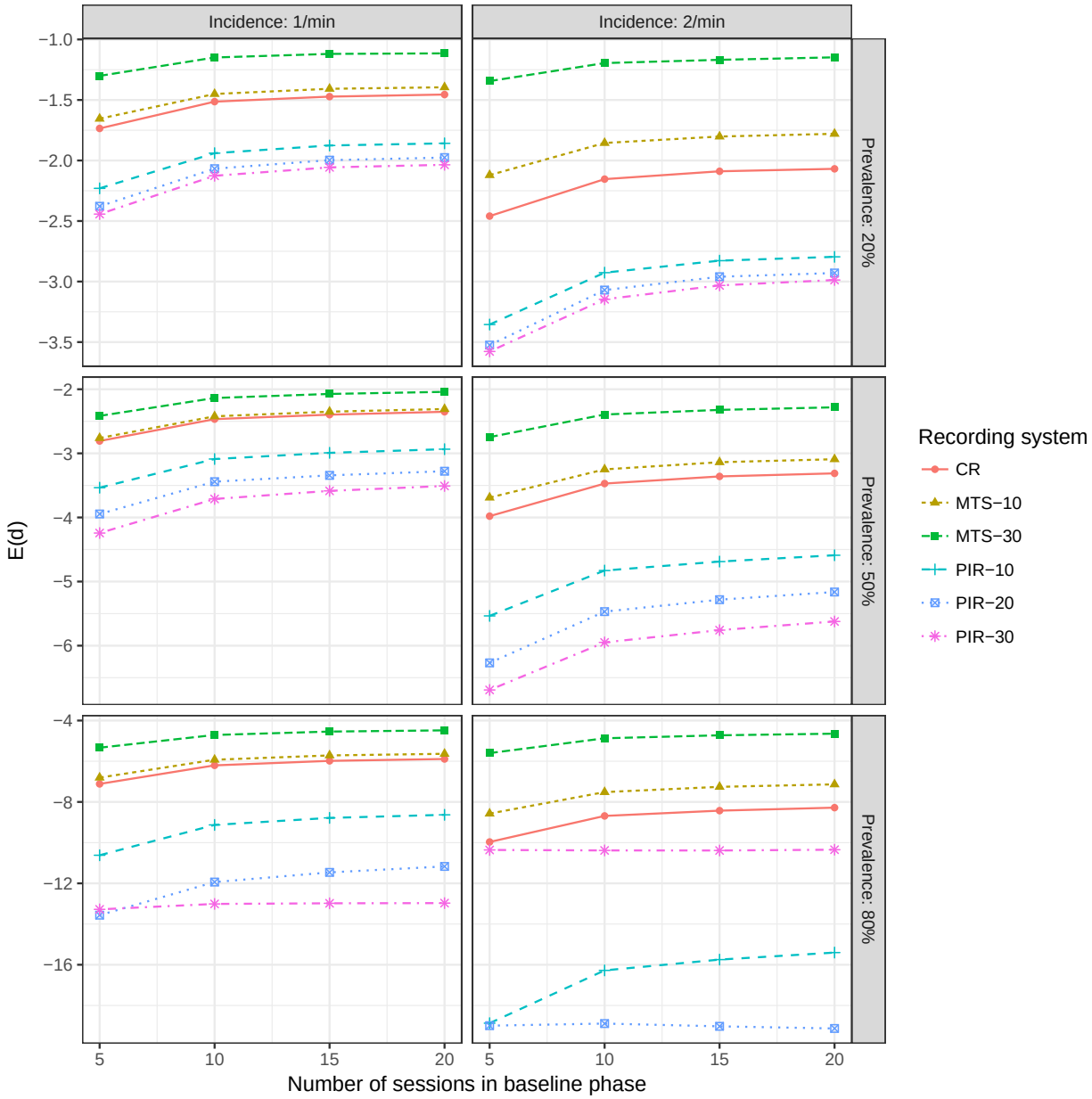

Figure S7: Expected value of d when treatment leads to a 50% change and sessions are 10 min, for varying numbers of sessions in the baseline phase and varying recording systems. CR = continuous recording; MTS = momentary time sampling; PIR = partial interval recording.

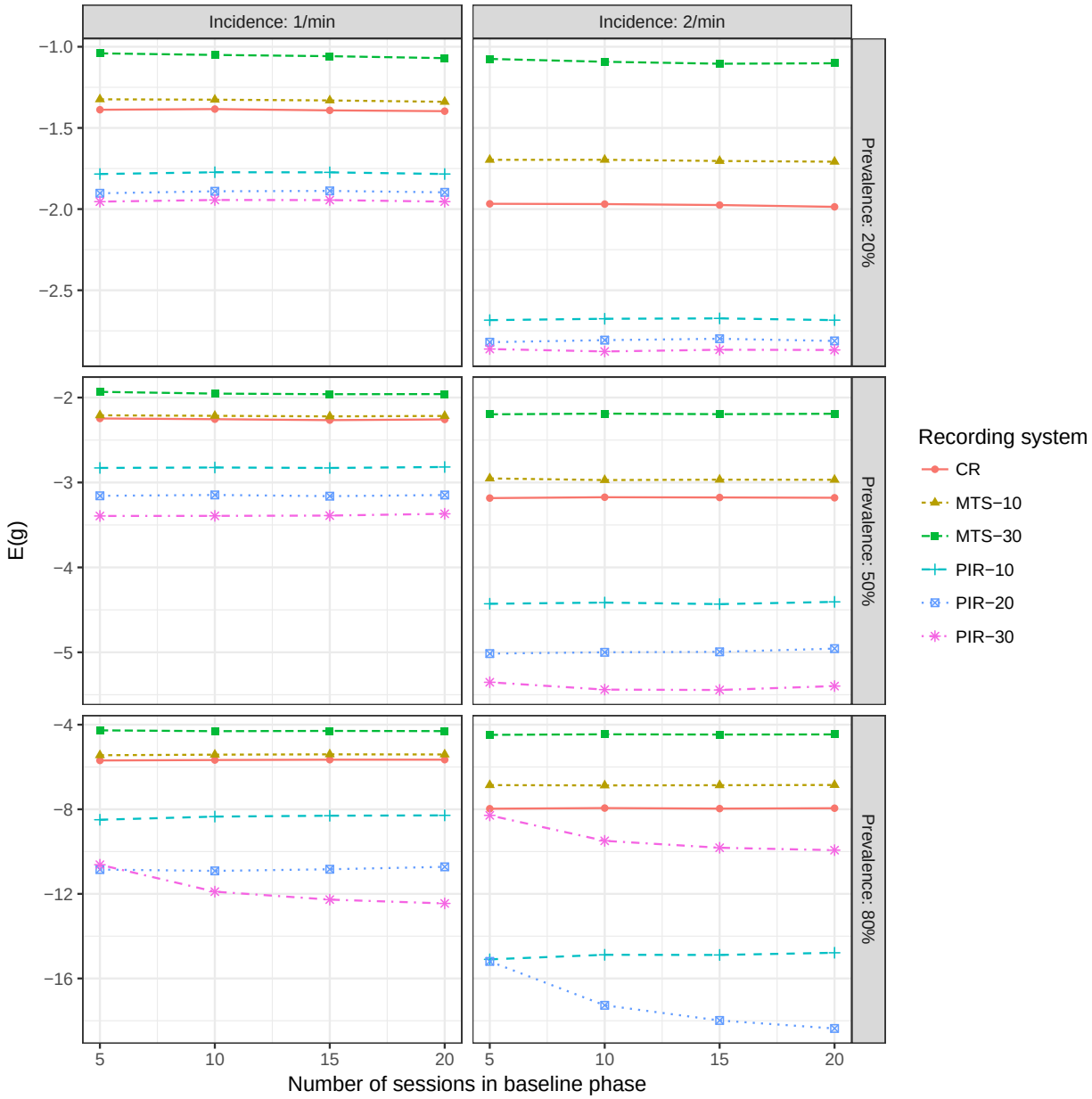

Figure S8: Expected value of g when treatment leads to a 50% change and sessions are 10 min, for varying numbers of sessions in the baseline phase and varying recording systems. CR = continuous recording; MTS = momentary time sampling; PIR = partial interval recording.

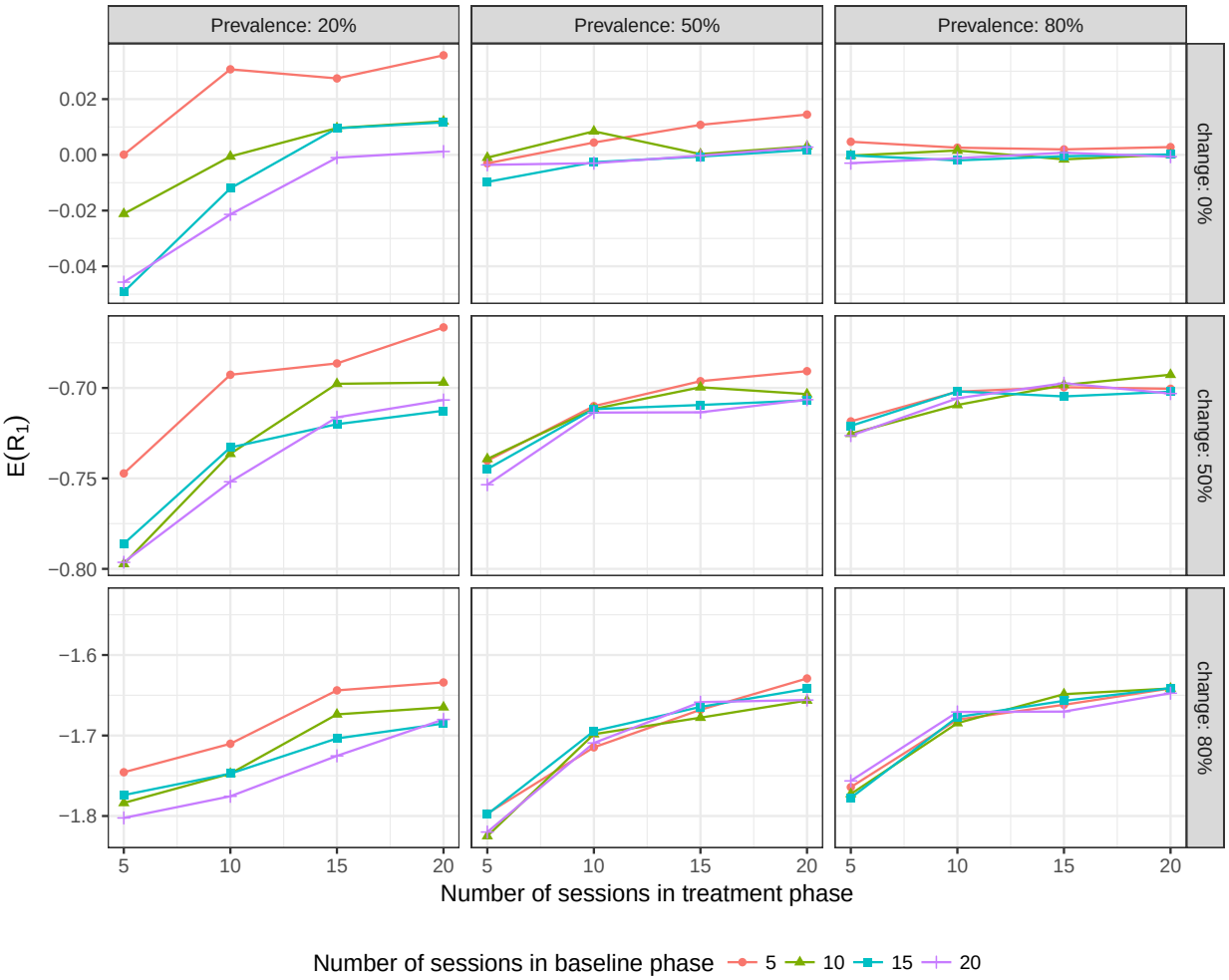

Figure S9: Expected value of R_1 when incidence is 1/min and the behavior is measured using 30 s MTS for 5 min sessions, for varying numbers of sessions in the baseline and treatment phases.

S5 Event behavior simulation

In addition to the simulations of state behavior, which are reported in the main text, I conducted a separate simulation to examine the operational sensitivities of the non-overlap measures and parametric effect sizes when the outcomes are based on observations of an event behavior. Event behavior streams can be simulated using the alternating renewal process model by setting all event durations to a number close to or exactly equal to zero, so that the features of the behavior are determined entirely by the inter-response time distribution. The most direct procedure for measuring an event behavior is to use frequency counting, which involves simply counting the number of occurrences of the behavior over the course of the observation session. However, partial interval recording (PIR) is also sometimes used to measure event behaviors, despite the fact that doing so can lead to distortions in the apparent magnitude of treatment effects (Pustejovsky & Swan, 2015). The remainder of this section describes the design of the simulation and the results

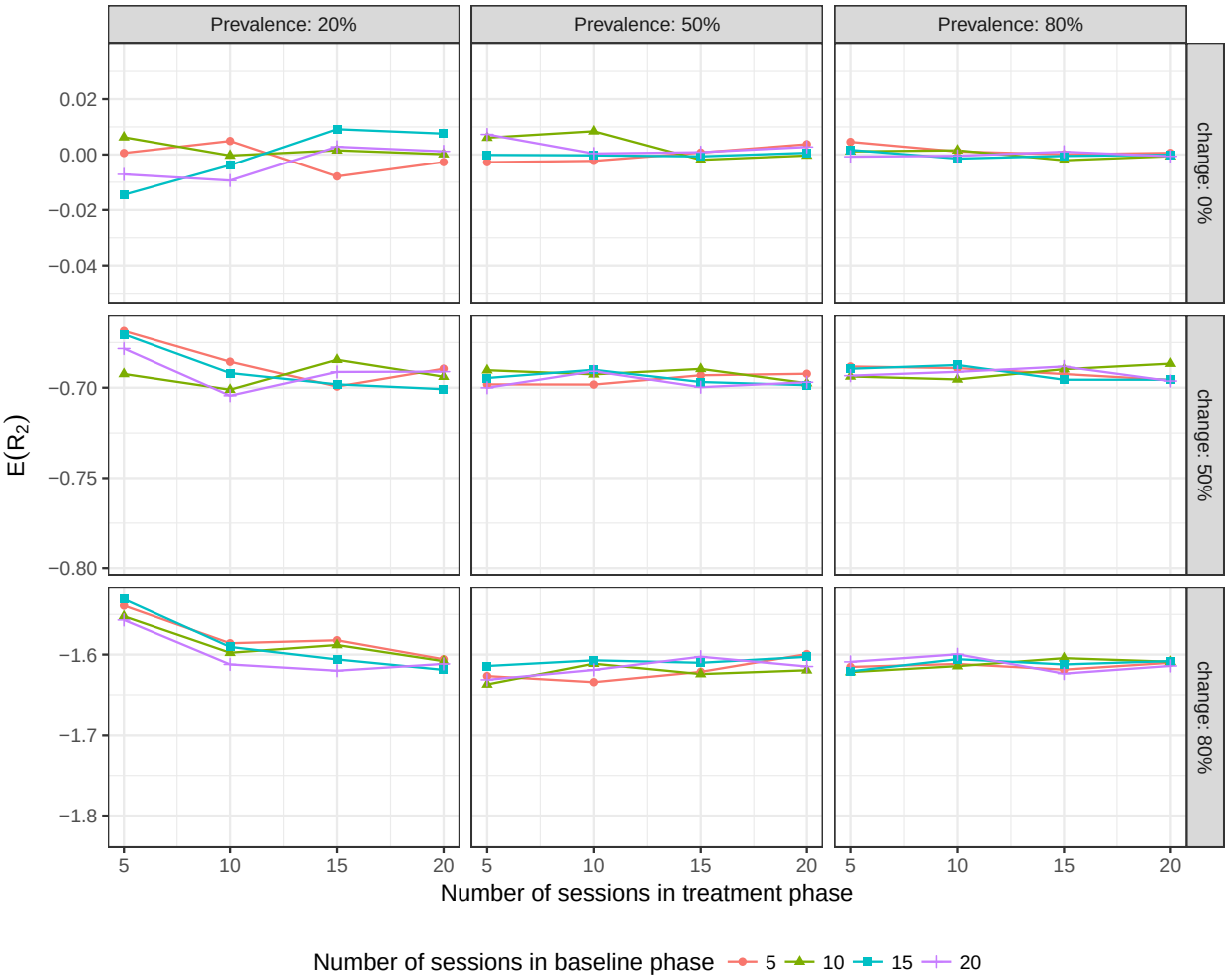

Figure S10: Expected value of R_2 when incidence is 1/min and the behavior is measured using 30 s MTS for 5 min sessions, for varying numbers of sessions in the baseline and treatment phases.

Table S4: Event behavior simulation design

Parameter	Levels
Incidence (per min)	$\frac{1}{2}$, 1, 2
Distribution	exponential, gamma(2)
Change (% decrease)	0%, 50%, 80%
Recording system	Frequency counting, PIR (10, 20, 30 s)
Session length (min)	5, 10, 15, 20
Sessions in the baseline phase	5, 10, 15, 20
Sessions in the treatment phase	5, 10, 15, 20

for each of the non-overlap measures.

S5.1 Simulation Design

Table S4 summarizes the design of the simulation study, which used a $3 \times 2 \times 3 \times 4 \times 4 \times 4 \times 4$ full factorial design. Three of the parameters determined the characteristics of the simulated behavior streams. First, the incidence of the behavior was set to $\frac{1}{2}$ (i.e., once per two minutes), one, or two times per minute. Second, inter-response times were assumed to follow either an exponential distribution or a gamma distribution with shape 2; the former distribution leads to frequency counts that are more variable around the average level (with variance equal to the mean), whereas the latter distribution leads to counts that are less variable. Third, treatment was assumed to lead to a 0%, 50%, or 80% reduction in the incidence of the behavior. Finally, all episode durations were set equal to zero in order to create event behavior streams. In order to illustrate the implications of these choices regarding parameter values and assumptions, Figure S11 displays examples of SCDs simulated based on each combination of incidence, distribution, and change in behavior. The observations were generated using frequency counting for 10 min sessions, with 10 sessions in each phase.

The event behavior simulations varied the same procedural factors as in the state behavior simulations, including recording system, session length, and the number of sessions in each phase. Because continuous recording and momentary time sampling are not appropriate for event behaviors, the simulations were limited to frequency counting and partial interval recording with 10, 20, or 30 s intervals. In keeping with the state behavior simulations, session length was set to 5, 10, 15, or 20 min and the numer of sessions in the baseline and treatment phases were set to 5, 10, 15, or 20 sessions.

For each combination of factor levels, 10,000 simulated AB phase-pairs were generated and the PND, PAND, RIRD, PEM, NAP, SMD, and LRR statistics were calculated based on each simulated dataset. The simulated values of each measure were averaged across replications in order to estimate its expected value. The computer code that implements the simulation and full numerical results are available in the supplementary materials that accompany this document; they are described further in Section S6.

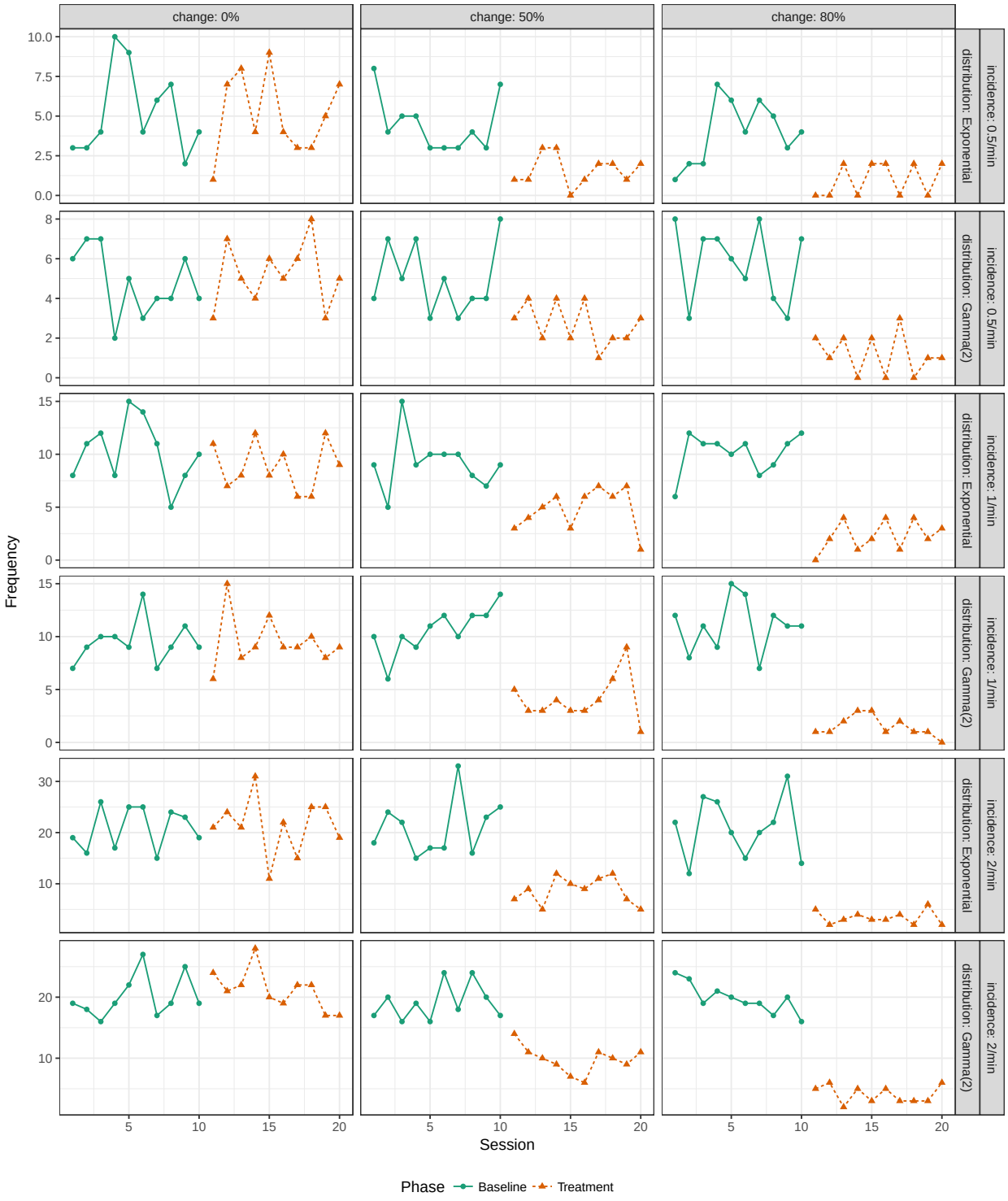

Figure S11: Simulated SCDs based on the alternating renewal process model, using frequency counting for 10 min observation sessions, for varying levels of incidence, inter-response time distribution, and change in behavior

S5.2 Results for non-overlap measures

The presentation of results follows the same organization as in the main text. As in the main text, some of the following figures present results for selected subsets of the conditions; in these cases, the results that are presented are generally consistent with the other simulation conditions, except when otherwise noted.

Figure S12 depicts the conditional ranges for each of the NOMs with respect to the four procedural factors in the simulation: number of sessions in the baseline phase, number of sessions in the treatment phase, observation session length, and recording system. It is constructed to parallel Figure 2 in the main text. Each column corresponds to a procedural factor; each row to one of the effect size measures. Within each panel, the full distribution of conditional ranges is represented using a violin plot, where width corresponds to relative frequency of a given value for the conditional range; the horizontal bars within the violin plot correspond to the quintiles of the distribution. Results are discussed separately for each of the NOMs.

S5.2.1 Percentage of non-overlapping data

Results in the top row of Figure S12 indicate that PND is sensitive to the number of sessions in the baseline phase and to observation session length, somewhat sensitive to recording system (particularly for 50% reductions in behavior), and not sensitive to the number of sessions in the treatment phase. Just as with state behavior, the expected value of PND depends on the number of observations in the baseline phase. When treatment has no effect, PND is only slightly affected by session length and recording procedure, due to the non-zero probability of exact ties between the measurements.

In the conditions where treatment produces beneficial effects, PND remains sensitive to baseline length and becomes considerably more sensitive to length of the observation session. Figure S13 plots the expected value of PND for a 50% change due to treatment, when outcomes are measured using frequency counting. Across all levels of incidence, PND is highly sensitive to the number of observations in the baseline phase and to the length of the observation session.

S5.2.2 Percentage of all non-overlapping data

Results for PAND are presented in the second row of Figure S12. Notably, PAND is sensitive to the number of sessions in both the baseline and treatment phase when treatment has no effect, although the degree of sensitivity decreases when treatment produces more beneficial effects. For further detail, Figure S14 depicts the expected value of PAND as a function of the number of observations in the baseline phase and in the treatment phase, for the subset of results where frequency counting is used for 10 min sessions and where inter-response times are exponentially distributed (results for gamma(2)-distributed inter-response times are very similar). The influence of sample size on the magnitude of PAND is quite similar to what was observed in the state behavior simulations (e.g., Figure S4). Just as in those simulations, when treatment has no effect,

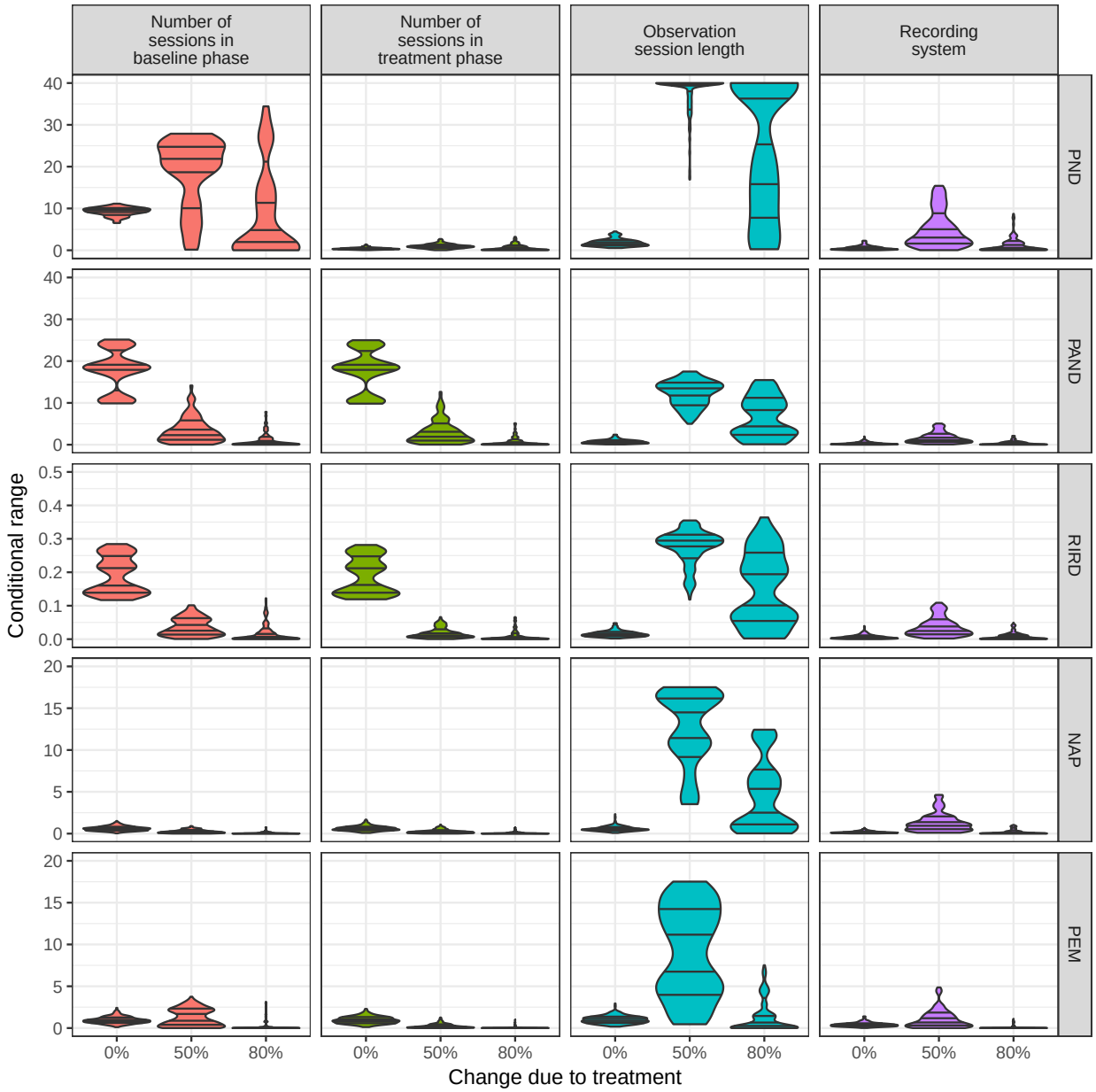

Figure S12: Conditional range distributions of the non-overlap effect size measures for each procedural factor, by percentage change from baseline to treatment. For clarity of illustration, the conditional range of PND is truncated at 40.

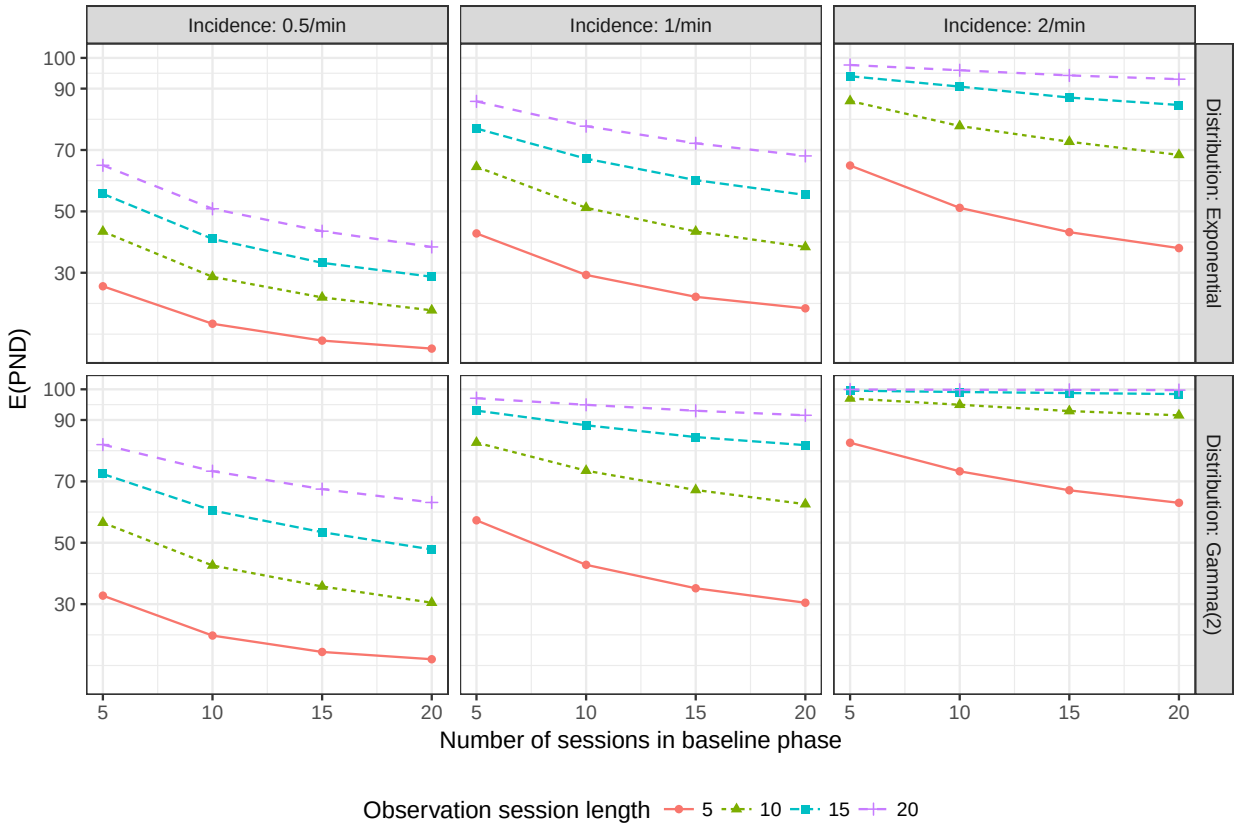

Figure S13: Expected value of PND when inter-response times are gamma(2)-distributed and treatment leads to a 50% change, for varying session lengths and vary numbers of observations in the baseline phase.

the magnitude of PAND is sensitive to the number of observations in the baseline and treatment phases. For example, when incidence is once per minute and treatment has no effect, the expected value of PAND varies between 57% and 81%. Again, PAND tends to be smaller when the number of observations in the baseline phase is equal to the number in the treatment phase. When treatment leads to larger effects, PAND becomes less sensitive to sample size; the decrease in sensitivity is more apparent for higher baseline incidence, though this appears to be due largely to ceiling effects.

The magnitude of PAND is also strongly affected by session length. Figure S15 depicts the expected value of PAND for varying session lengths and recording systems, based on the subset of results where treatment leads to a 50% reduction in behavior and both baseline and treatment phases include 10 observation sessions. Across all levels of incidence, the expected value of PAND is strongly influenced by session length. As in the state behavior simulation, it appears that PAND is only slightly sensitive to the choice of recording system, with the largest degree of sensitivity occurring for higher baseline incidence and shorter session lengths. Overall, the results indicate that PAND is sensitive to the number of observations in the baseline and treatment phases and to observation session length, but less sensitive to the recording system.

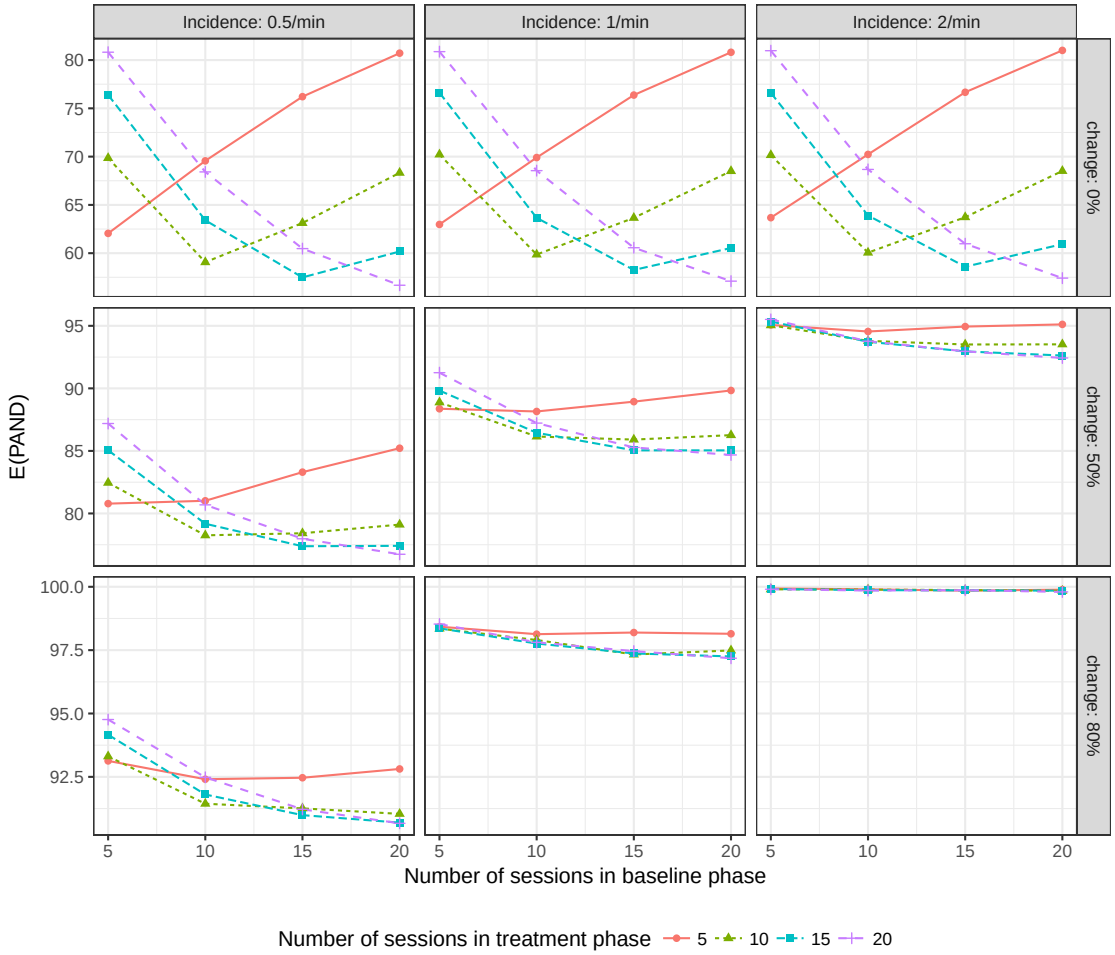

Figure S14: Expected value of PAND based on frequency counting data with 10 min sessions, when the inter-response time distribution is exponential, for varying numbers of sessions in the baseline and treatment phases.

S5.2.3 Robust improvement rate difference

Results for RIRD are presented in the second row of Figure S12. Just as in the state behavior simulation, RIRD is sensitive to the same procedural factors as PAND. Figure S16 depicts the expected value of RIRD as a function of the number of observations in the baseline phase and in the treatment phase, for the subset of results where frequency counting is used for 10 min sessions and where inter-response times are exponentially distributed; it is constructed in the same way as Figure S14 for PAND. The top row of the figure indicates that the expected value of RIRD varies considerably when the treatment has no effect; for instance, when incidence is once per two minutes, the expected value ranges from 0.13 to 0.4. For larger degrees of change between phases, RIRD becomes somewhat less sensitive to the number of observations in each phase; for instance, when incidence is once per two minutes and treatment produces a 50% change in behavior, RIRD ranges from 0.53 to 0.62. In contrast to PAND, the magnitude of RIRD tends to be larger when

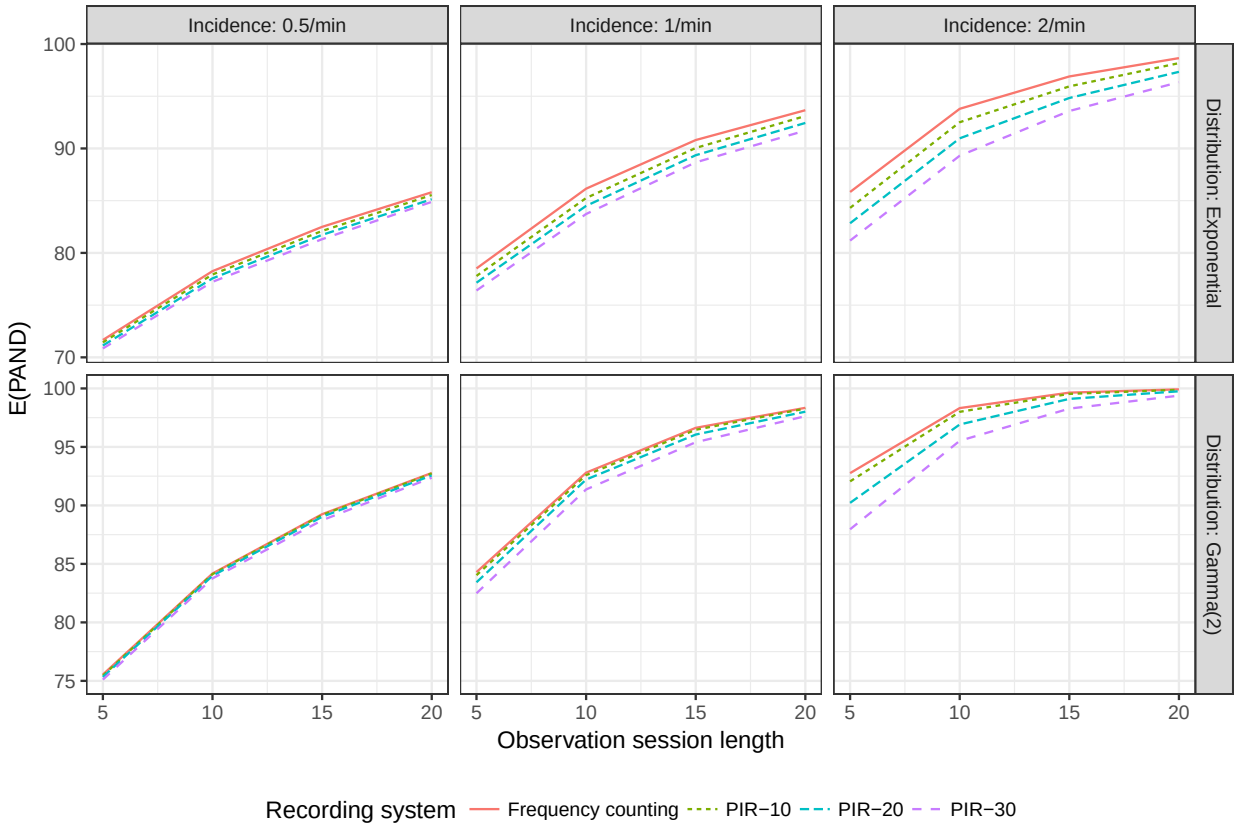

Figure S15: Expected value of PAND for varying session lengths and recording systems, when treatment leads to a 50% change and both phases include 10 sessions. PIR = partial interval recording.

the baseline and treatment phases include an equal number of sessions.

Figure S17 depicts the expected value of RIRD for varying session lengths and recording systems, based on the subset of results where treatment leads to a 50% reduction in behavior and both baseline and treatment phases include 10 observation sessions. The results closely parallel those for PAND (e.g., Figure S15). Like PAND, the expected value of RIRD is highly sensitive to session length. Also like PAND, RIRD appears to be only moderately sensitive to the choice of recording procedure. For example, treatment produces a 50% decrease from a baseline incidence of twice per minute and sessions last 5 min, the expected value of RIRD varies from 0.76 to 0.86. Also, the degree of sensitivity tends to be reduced when session lengths are longer.

S5.2.4 Percentage exceeding the median

Results for PEM are presented in the last row of Figure S12. Just as in the state behavior simulation, PEM is not sensitive to the number of sessions in the baseline or treatment phase. More specifically, when treatment has no effect on the behavior, the expected value of PEM is always exactly 50%, across all levels of the other factors. Furthermore, when treatment led to an 80% decrease in

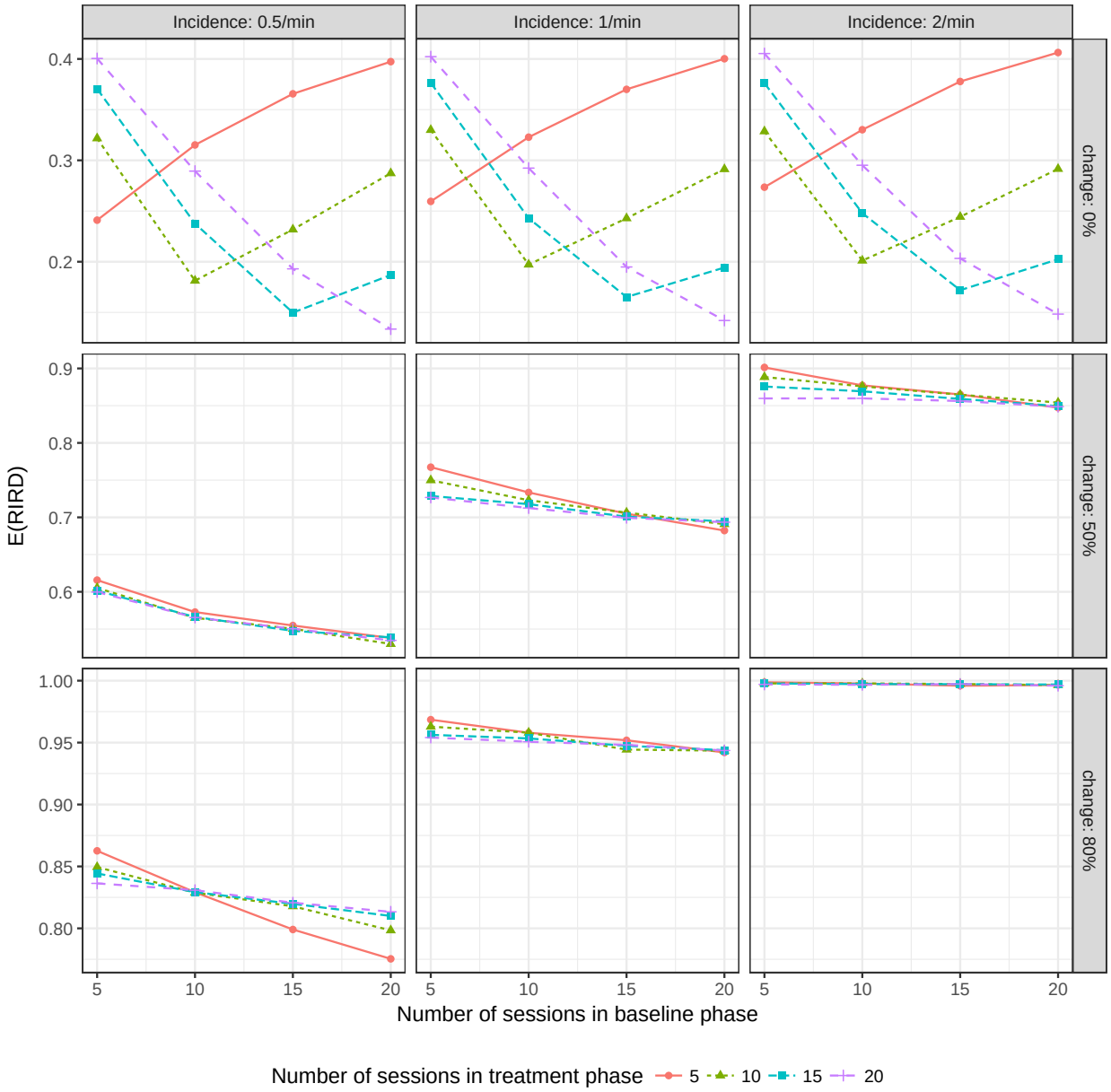

Figure S16: Expected value of RIRD based on frequency counting data with 10 min sessions, when the inter-response time distribution is exponential, for varying numbers of sessions in the baseline and treatment phases.

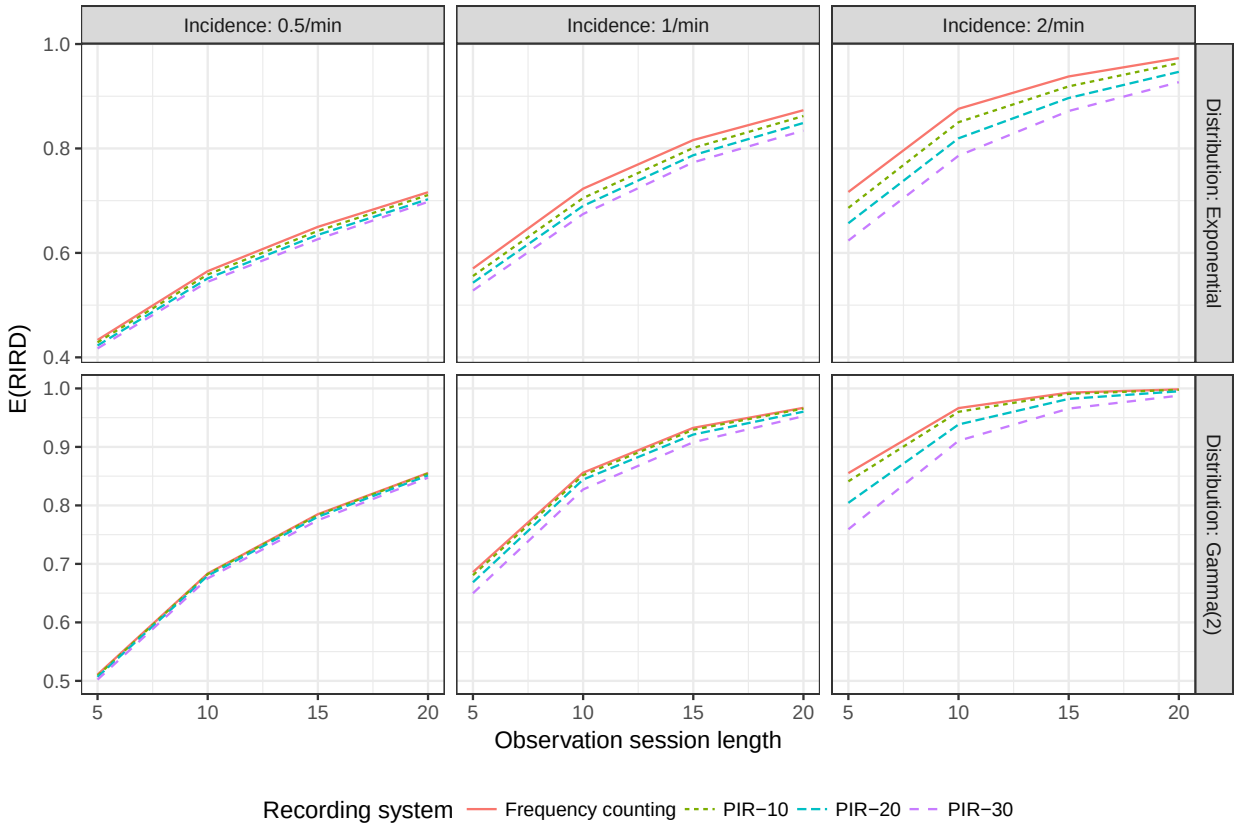

Figure S17: Expected value of RIRD for varying session lengths and recording systems, when treatment leads to a 50% change and both phases include 10 sessions. PIR = partial interval recording.

incidence of the behavior, the expected value of PEM is at or near the ceiling level of 100% across all levels of the other factors.

When treatment leads to a 50% decrease in incidence of the behavior, PEM becomes sensitive to session length. Figure S18 plots the expected value of PEM for a 50% change in behavior. The results are averaged across levels of the number of observations in baseline and treatment because these factors affect PEM only very slightly. For lower-incidence behaviors, it can be seen that the magnitude of PEM is quite sensitive to the length of the observation session. For example, for a behavior with baseline incidence of 0.5 events per minute and exponential inter-response times, measured by frequency counting, the expected value of PEM ranges from 80% for 5 min sessions to 96% for 20 min sessions. Notably, unlike its behavior in the state behavior simulations, PEM appears to be only slightly affected by the choice of recording system, and only when session lengths are short.

S5.2.5 Non-overlap of all pairs

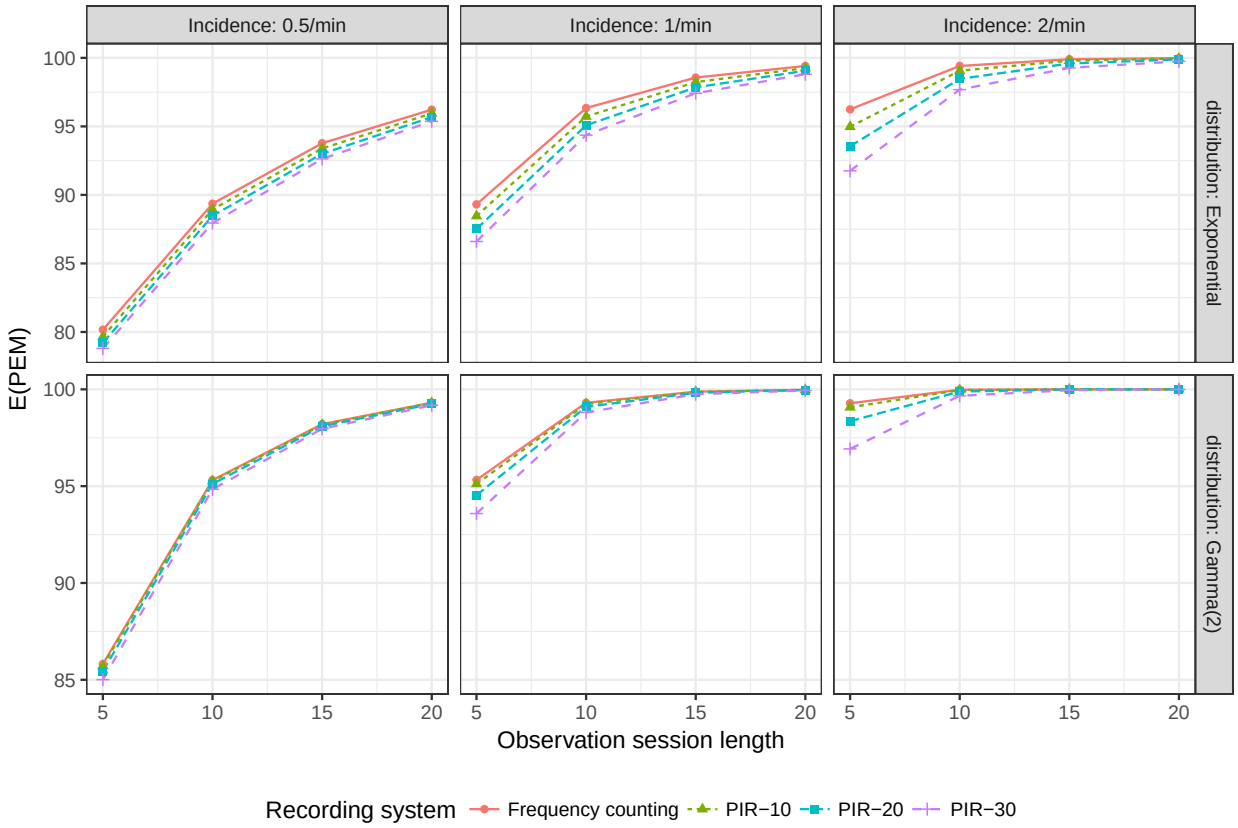

Figure S18: Expected value of PEM for various recording systems and session lengths, when treatment leads to a 50% change in behavior. PIR = partial interval recording.

Results for NAP are presented in the fourth row of Figure S12. Generally, the behavior of NAP is very similar to that of PEM. Like PEM, the expected value of NAP is exactly equal to 50% when treatment has no effect on the outcome and its expected value is not affected by the number of sessions in the baseline phase or treatment phase. Also like PEM, the expected value of NAP was at or near the ceiling level when treatment led to an 80% decrease in incidence of the behavior. Figure S19 plots the expected value of NAP when treatment leads to a 50% decrease in behavior, for varying session lengths and recording systems. Just as in the state behavior simulations, NAP is highly sensitive to observation session length for certain combinations of behavioral characteristics (particularly for behaviors with lower incidence).

S5.3 Results for parametric measures

Figure S20 depicts the conditional ranges for each of the parametric effect size measures, including the basic SMD (d), the bias-corrected SMD (g), the basic LRR (R_1), and the bias-corrected LRR (R_2), with respect to each of the four procedural factors. Its construction parallels that of Figure 6 from the main text.

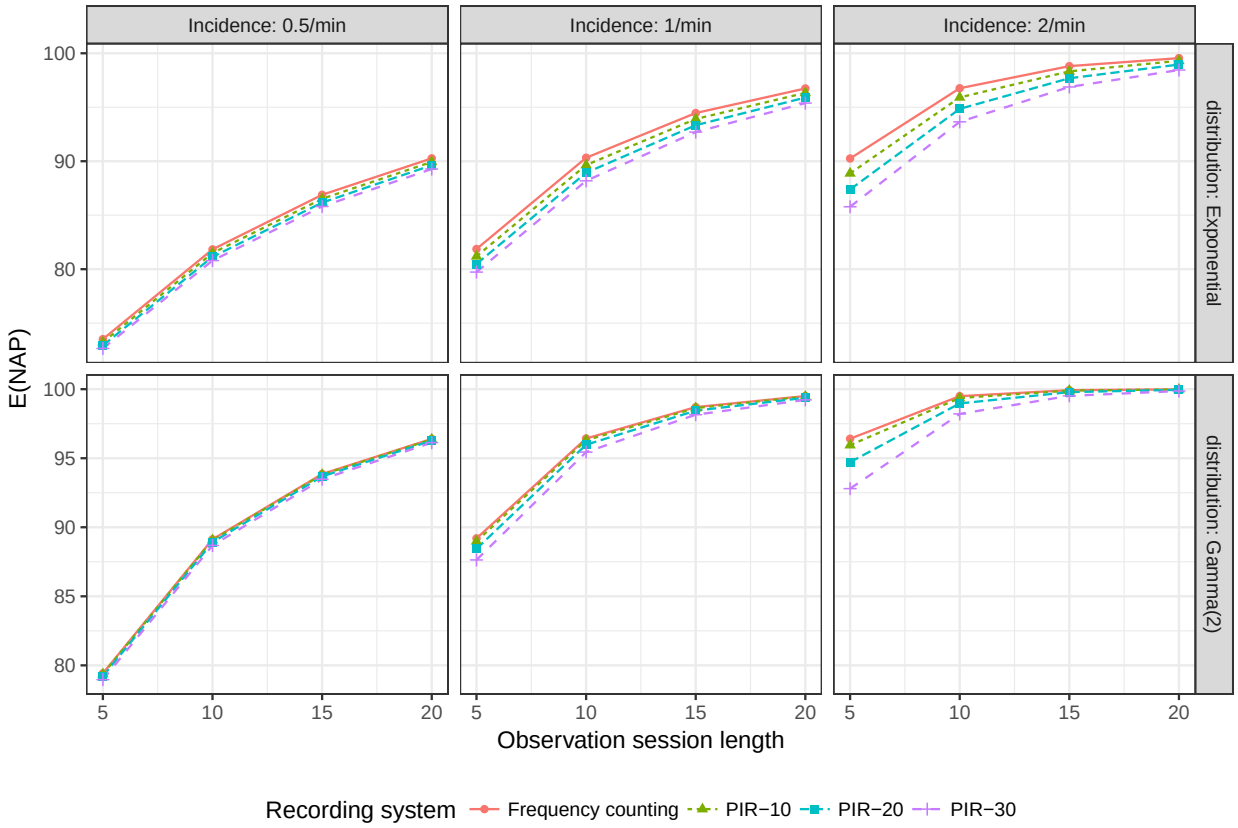

Figure S19: Expected value of NAP for various recording systems and session lengths, when treatment leads to a 50% change in behavior.

S5.3.1 SMD

Results for the basic (d) and bias-corrected (g) SMD estimators are presented in the first and second rows of Figure S20. The behavior of the basic and bias-corrected SMD estimators is very similar to their behavior for state behaviors. Figure S21 depicts the expected value of d (top panel) and g (bottom panel) when treatment leads to a 50% reduction in behavior and the outcome is measured using event counting. It can be seen that the expected value of d changes as number of baseline sessions increases, sometimes to an extent that exceeds 0.5 SD, while the expected value of g is practically unaffected by the number of observations in either phase. However, both estimators are strongly influenced by observation session length. For both estimators, the pattern of results is very similar when the outcome is measured using a PIR system. Unlike the findings from the state behavior simulation, neither estimator is sensitive to the choice of recording procedure for measuring event behavior.

S5.3.2 LRR

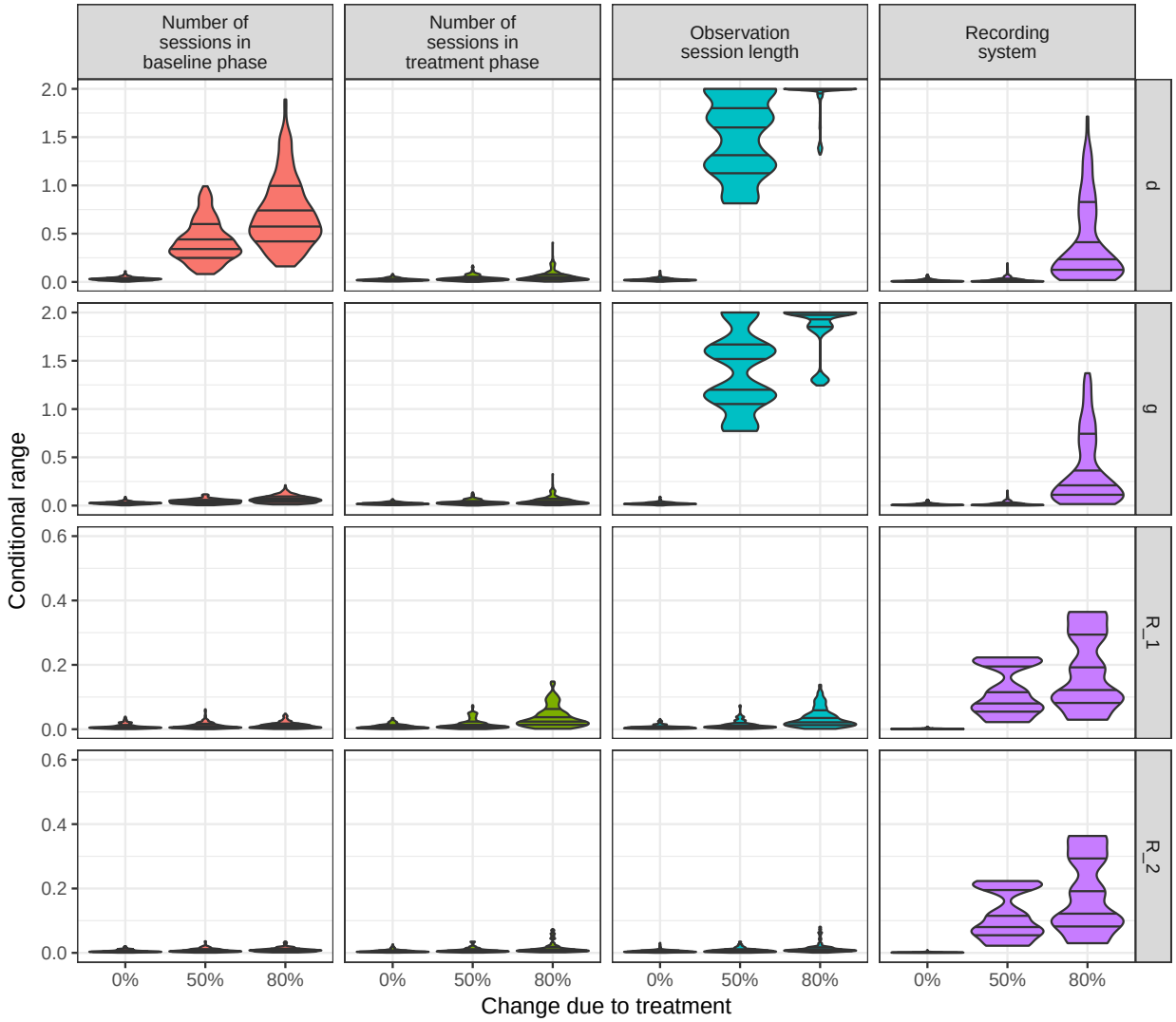

Figure S20: Conditional range distributions of the parametric effect size measures for each procedural factor, by percentage change from baseline to treatment.

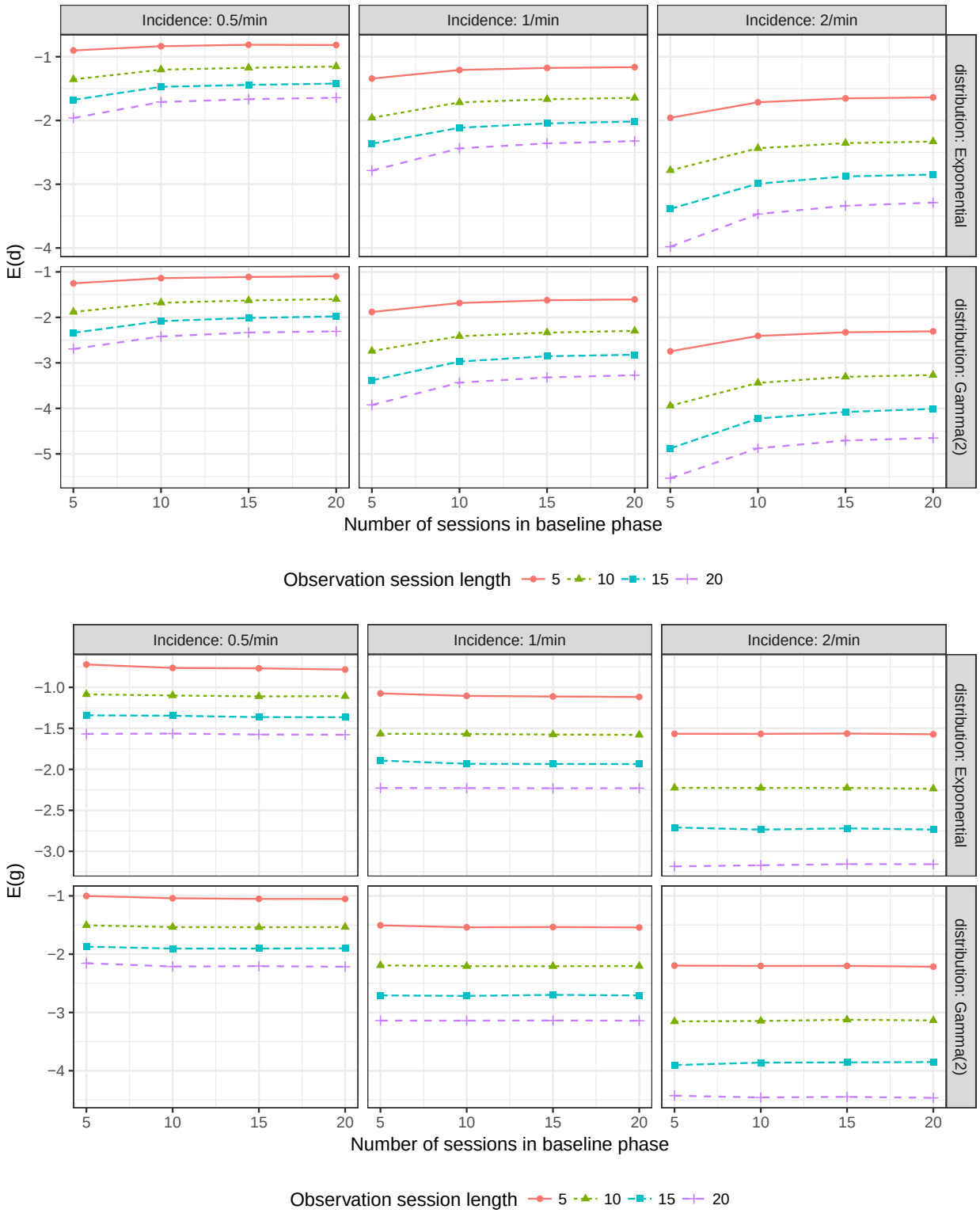

Figure S21: Expected value of SMD estimators when treatment leads to a 50% change and the outcome is measured using event counting, for varying numbers of sessions in the baseline and treatment phases.

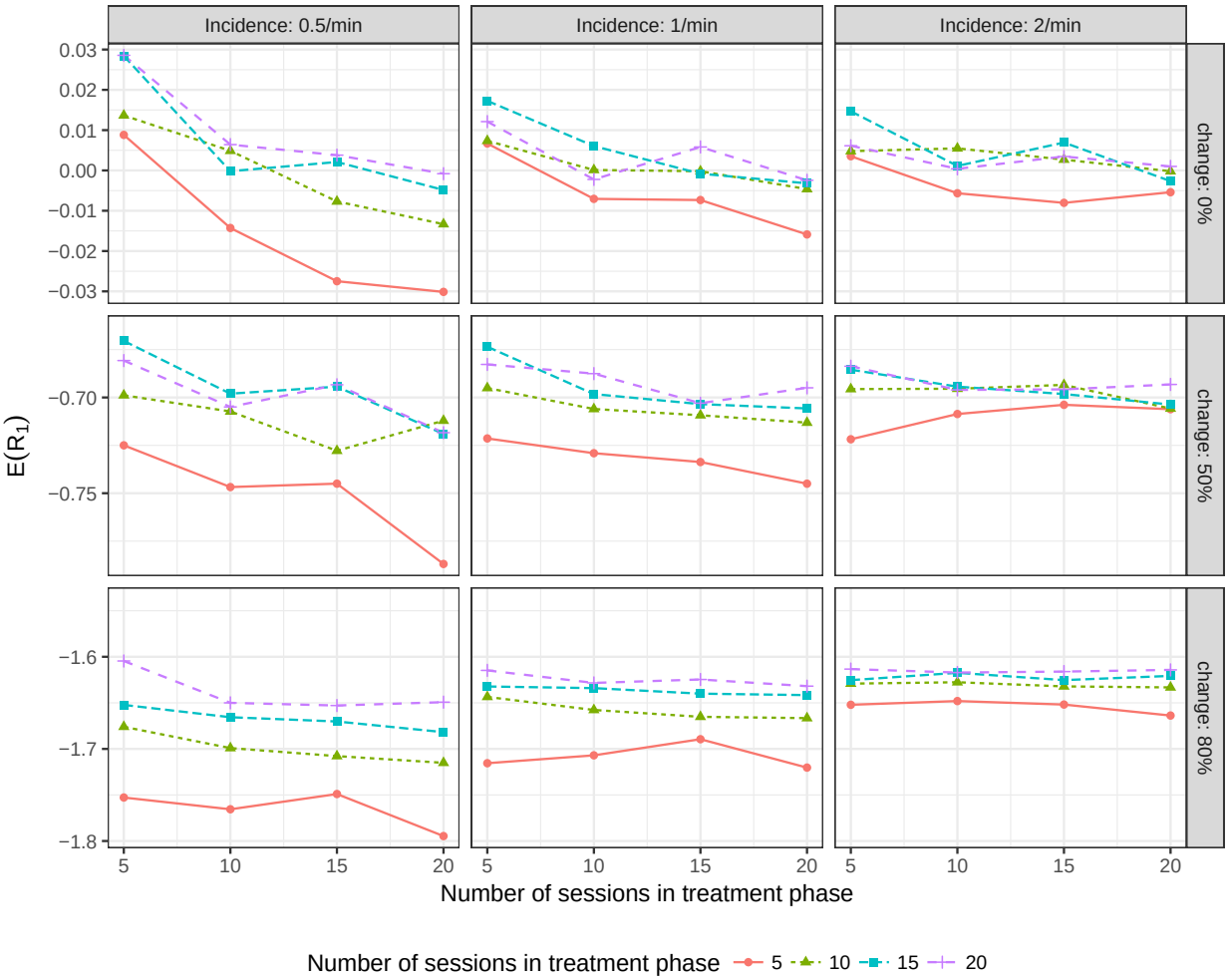

Figure S22: Expected value of the LRR moment estimator (R_1) for varying numbers of sessions in the baseline and treatment phases, when inter-reponse times are exponentially distributed and the behavior is measured with frequency counting for 5 min sessions.

Results for the basic (R_1) and bias-corrected (R_2) LRR estimators are presented in the third and fourth rows of Figure S20. The basic estimator is somewhat sensitive to the number of sessions in the treatment phase and to observation session length, particularly when treatment leads to larger reductions in behavior. Bias-correction removes this sensitivities, so that the magnitude of R_2 is unaffected by the number of sessions in either phase and unaffected by observation session length.

A more detailed view is given in Figures S22 and S23, which plot the expected value of the LRR estimators for varying numbers of sessions in each phase, when the outcome is measured using frequency counting. Both figures are based on the subset of results where the inter-response time distribution is exponential, and the behavior is measured for 5 min sessions; the LRR estimators are typically less sensitive to procedural factors when session length is longer. In Figure S22, it can be seen that the expected value of the LRR moment estimator R_1 has a small-sample bias that makes its sensitive to the length of the baseline and treatment phases, similar to its behavior with state

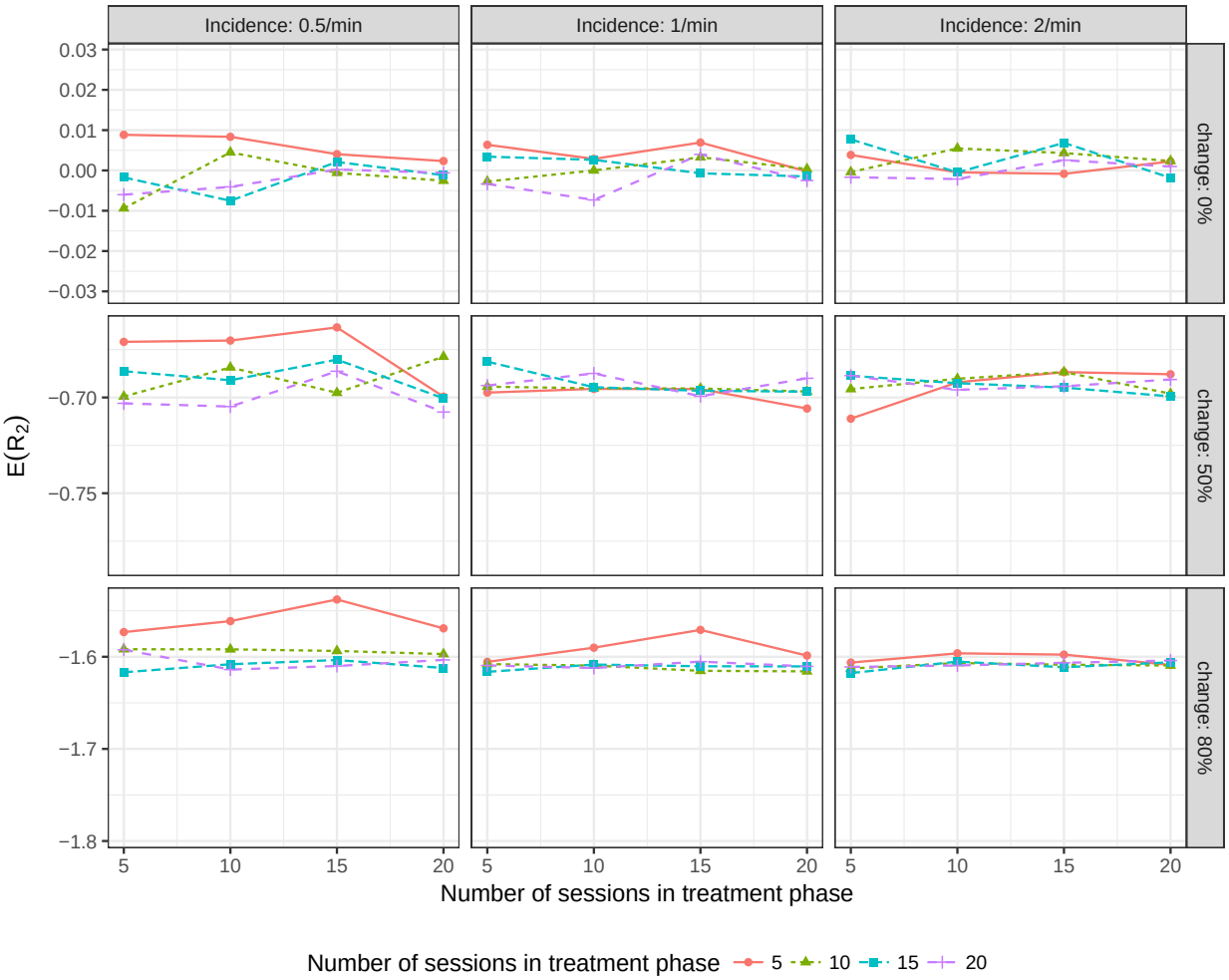

Figure S23: Expected value of bias-corrected LRR estimator (R_2) for varying numbers of sessions in the baseline and treatment phases, when inter-reponse times are exponentially distributed and the behavior is measured with frequency counting for 5 min sessions.

behaviors. In contrast, the expected value of the bias-corrected estimator R_2 is mostly unaffected by sample size (Figure S23). The estimator has residual bias only when the change in behavior is very large (80%) and the treatment phase includes only 5 session; outside of these conditions, R_2 is essentially unaffected by sample size.

S5.4 Discussion

The results of the the event behavior simulation are broadly consistent with the results of the state behavior simulation reported in the main text. Just as in the other simulations, PND, PAND, and RIRD are all sensitive to the number of sessions in the baseline phase, and PAND and RIRD are also sensitive to the number of sessions in the treatment phase. All three measures are sensitive to the length of observation sessions. As in the other simulations, PEM, NAP, and SMD are insensitive to the number of observations in either phase but are affected by the length of observation sessions,

while the bias-corrected LRR estimator is not sensitive to the number of observations in either phase or to observation session length.

The main difference in findings between the event behavior and state behavior simulations concerns the degree to which the measures are sensitive to the recording system. Across all seven effect sizes, it appeared that the choice of recording system has only slight or moderate effects on magnitude when used with measures of event behavior, whereas the choice of recording system had stronger effects on magnitude when used with measures of state behavior. However, the finding may be limited by the range of conditions examined in the event behavior simulation. Other work—also based on the alternating renewal process model—has identified other conditions under which the use of partial interval recording to measure event behavior can produce highly misleading inferences, such as concluding that treatment reduces the incidence of an undesirable behavior when in fact it increases it (Pustejovsky & Swan, 2015). Thus, the lack of sensitivity to recording system that was observed in the present study might not hold more broadly.

S6 Simulation replication materials

In addition to the material presented in this document, the supplementary materials include two additional files. The file `effect_size_sims.R` contains the R script that was used to execute both the state behavior and event behavior simulations. The file `EffectSizeMeasures.csv` is a comma-separated value file containing the complete numerical results of the simulations. The file contains 18 columns and 12672 rows of data. The content of the columns is as follows:

1. `behavior` - character string indicating whether the results are from the event behavior simulation or the state behavior simulation.

2. `prevalence` - percentage corresponding to the assumed prevalence of the behavior (equal to 0% for the event behavior simulation and 20%, 50%, or 80% for the state behavior simulation).

3. `incidence` - character string corresponding to the assumed incidence of the behavior.

4. `distribution` - character string corresponding to the distribution used to generate inter-response times (which was varied in the event behavior simulation but not in the state behavior simulation).

5. `change` - percentage corresponding to the assumed percentage reduction in behavior in the treatment phase. For the event behavior simulation, treatment was assumed to reduce the incidence of the behavior; for the state behavior simulation, treatment was assumed to reduce both prevalence and incidence equally.

6. `procedure` - character string indicating the recording system used to simulate outcome measurements. For the event behavior simulation, the recording system was `EC` for event counting or `PIR-x` for x s partial interval recording, with $x = 10$, 20, or 30. For the state behavior

simulation, the recording system was `CR` for continuous recording, `MTS-x` for x s momentary time sampling, or `PIR-x` for x s partial interval recording, in each case with $x = 10$, 20, or 30.

7. `Session_length` - integer corresponding to the length of the simulated observation session, in min
8. `Baseline_length` - integer corresponding to the number of sessions in the baseline phase.
9. `Treatment_length` - integer corresponding to the number of sessions in the treatment phase.
10. `PND` - expected value of the percentage of non-overlapping data, estimated based on 10,000 simulated AB designs.
11. `PEM` - expected value of the percentage exceeding the median, estimated based on 10,000 simulated AB designs.
12. `PAND` - expected value of the percentage of all non-overlapping data, estimated based on 10,000 simulated AB designs.
13. `IRD` - expected value of the improvement rate difference, estimated based on 10,000 simulated AB designs.
14. `NAP` - expected value of the non-overlap of all pairs, estimated based on 10,000 simulated AB designs.
15. `SMD.d` - expected value of the uncorrected SMD (d), estimated based on 10,000 simulated AB designs.
16. `SMD.g` - expected value of the bias-corrected SMD (g), estimated based on 10,000 simulated AB designs.
17. `LRR.R1` - expected value of the uncorrected LRR (R_1), estimated based on 10,000 simulated AB designs.
18. `LRR.R2` - expected value of the bias-corrected LRR (R_2), estimated based on 10,000 simulated AB designs.

References

DiCarlo, C. F., Reid, D. H., & Stricklin, S. B. (2003). Increasing toy play among toddlers with multiple disabilities in an inclusive classroom: A more-to-less, child-directed intervention continuum. *Research in Developmental Disabilities*, *24*(3), 195–209. doi: 10.1016/S0891-4222(03)00025-8

Harrington, M., & Velicer, W. F. (2015). Comparing visual and statistical analysis in single-case studies using published studies. *Multivariate Behavioral Research*, *50*(2), 162–183. doi: 10.1080/00273171.2014.973989

Heath, A. K., Ganz, J. B., Parker, R. I., Burke, M., & Ninci, J. (2015). A meta-analytic review of functional communication training across mode of communication, age, and disability. *Review Journal of Autism and Developmental Disorders*, *2*(2), 155–166. doi: 10.1007/s40489-014-0044-3

Heyvaert, M., Saenen, L., Campbell, J. M., Maes, B., & Onghena, P. (2014). Efficacy of behavioral interventions for reducing problem behavior in persons with autism: An updated quantitative synthesis of single-subject research. *Research in Developmental Disabilities*, *35*(10), 2463–2476. doi: 10.1016/j.ridd.2014.06.017

Maggin, D. M., Chafouleas, S. M., Goddard, K. M., & Johnson, A. H. (2011). A systematic evaluation of token economies as a classroom management tool for students with challenging behavior. *Journal of School Psychology*, *49*(5), 529–54. doi: 10.1016/j.jsp.2011.05.001

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

Manolov, R., & Solanas, A. (2009). Percentage of nonoverlapping corrected data. *Behavior Research Methods*, *41*(4), 1262–1271. doi: 10.3758/BRM.41.4.1262

Parker, R. I., Hagan-Burke, S., & Vannest, K. J. (2007). Percentage of all non-overlapping data (PAND): An alternative to PND. *The Journal of Special Education*, *40*(4), 194–204. doi: 10.1177/00224669070400040101

Parker, R. I., & Vannest, K. J. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy*, *40*(4), 357–67. doi: 10.1016/j.beth.2008.10.006

Parker, R. I., Vannest, K. J., & Brown, L. (2009). The improvement rate difference for single-case research. *Exceptional Children*, *75*(2), 135–150. doi: 10.1177/001440290907500201

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

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

Tarlow, K. R. (2017). An improved rank correlation effect size statistic for single-case designs: Baseline corrected Tau. *Behavior Modification*, *41*(4), 427–467. doi: 10.1177/0145445516676750

White, D. M., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P. (1989). Applications of meta

analysis in individual-subject research. *Behavioral Assessment*, *11*(3), 281–296.

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