Supplementary materials for "Operational sensitivities of the non-overlap effect sizes for single-case designs"

James E. Pustejovsky

August 5, 2015

Contents

1 Mathematical definitions of the non-overlap measures

This sections provides precise mathematical definitions of each of the NOM effect sizes and further discussion of their statistical properties. As in the main text, the definitions are based on the assumption that an increase in the outcome is desirable; for an outcome where a decrease is desirable, the formulas would be evaluated after multiplying each of the outcome measurements by -1. Let m denote the number of observations in the baseline phase and n denote the number of observations in the treatment phase. Let $Y_1^A, ..., Y_m^A$ denote the outcome measurements during the baseline phase and $Y_1^B, ..., Y_n^B$ denote the outcome measurements during the treatment phase. Let $I(A)$ denote the indicator function, which is equal to one when the criterion A is true and equal to zero when A is false.

1.1 Percentage of non-overlapping data

PND is calculated as

$$\text{PND} = 100\% \times \frac{1}{n} \sum_{i=1}^{n} I\left(Y_i^B > Y_{(m)}^A\right), \tag{1}$$

where $Y^A_{(m)} = \max\left\{Y^A_1, ..., Y^A_m\right\}$. Assuming that the observations in the treatment phase are identically distributed, the expected magnitude of PND is equal to

$$\mathrm{E}\left(\mathrm{PND}\right) = 100\% \times \Pr\left(Y^B_1 > Y^A_{(m)}\right).$$

The expected magnitude of PND decreases as the length of the baseline phase increases because the maximum value of the baseline phase $\left(Y^A_{(m)}\right)$ will tend to be larger when m is larger.

1.2 Percentage exceeding the median

PEM is calculated as

$$\mathrm{PEM} = 100\% \times \frac{1}{n}\sum_{i=1}^{n}\left[I\left(Y^B_i > M^A\right) + 0.5I\left(Y^B_i = M^A\right)\right], \tag{2}$$

where $M^A = \mathrm{median}\left\{Y^A_1, ..., Y^A_m\right\}$. Assuming that the observations in the treatment phase are identically distributed, the expected magnitude of PEM is equal to

$$\mathrm{E}\left(\mathrm{PEM}\right) = 100\% \times \Pr\left(Y^B_1 > M^A\right).$$

Unlike PND, the expected magnitude of PEM is not much influenced by the number of observations in the baseline phase (and is not at all influenced by the number of observations in the treatment phase) because the distribution of the median baseline phase observation does not depend strongly on m. If the observations are mutually independent and identically distributed within each phase, then the expected magnitude of PEM is 50%.

1.3 Percentage of all non-overlapping data

PAND is rather difficult to express mathematically. Let $Y^A_{(1)}, Y^A_{(2)}, ..., Y^A_{(m)}$ denote the values of the baseline phase data, sorted in increasing order, and let $Y^B_{(1)}, Y^B_{(2)}, ..., Y^B_{(n)}$ denote the values of the sorted treatment phase data. In general, PAND can be calculated as

$$\mathrm{PAND} = 100\% \times \frac{1}{m+n}\max\left\{(i+j)\, I\left(Y^A_{(i)} < Y^B_{(n+1-j)}\right)\right\}, \tag{3}$$

where the maximum is taken over the values $1 \leq i \leq m$ and $1 \leq j \leq n$.

The logical range of PAND is more complicated than that of PND or PEM. If there is complete separation between phases, so that $Y^A_{(m)} < Y^B_{(1)}$, then PAND will reach the maximum possible value of 100%. The minimum possible value of PAND is not zero, but rather $\max\{m, n\}/(m+n)$, i.e., the number of observations in the longer of the two phases, divided by the total number of observations. This can be seen by considering a case where the minimum of the baseline phase observations is larger than the maximum of the treatment phase observations, so that $Y^A_{(1)} > Y^B_{(n)}$. In order to obtain no overlap, one must either remove all baseline phase observations or all treatment phase

observations; thus, the minimum number of observations that must be removed is equal to the number of observations in the shorter of the two phases, and the number of observations remaining is equal to the number of observations in the longer phase.

Parker and colleagues (Parker, Hagan-Burke, & Vannest, 2007; Parker, Vannest, & Davis, 2011) stated that when treatment has no effect on the outcome, the expected magnitude of PAND is 50%. However, this is incorrect because the minimum possible value of PAND is greater than or equal to 50% and the expected magnitude must be strictly larger than the minimum possible value. Even under the simplifying assumption that all of the observations are independent and identically distributed, the expected magnitude of PAND depends on the distribution of the order statistics in each phase and is difficult to derive analytically.

1.4 Robust improvement rate difference

RIRD is calculated as follows. Let x be the minimum number of observations that must be removed from either phase so that the maximum of the remaining baseline phase observations is less than the minimum of the remaining treatment phase observations. Mathematically, $x = (m+n)(1 - \text{PAND}/100\%)$ where PAND is calculated as in Equation (3). IRD is then calculated as

$$\text{RIRD} = \frac{n - x/2}{n} - \frac{x/2}{m}. \tag{4}$$

Straight-forward algebraic manipulations lead to the fact that IRD (robust phi) is a linear re-scaling of PAND, where

$$\text{RIRD} = \frac{1}{2mn}\left[(m+n)^2 \frac{\text{PAND}}{100\%} - m^2 - n^2\right].$$

The logical range of RIRD can be derived using the algebraic relationship between it and PAND. The maximum possible value of RIRD is 1, which occurs when there is complete separation between phases and PAND equals 100%. The minimum possible value of RIRD occurs when $Y^A_{(1)} > Y^B_{(n)}$, in which case

$$\text{RIRD} = \frac{1}{2} - \min\left\{\frac{m}{n}, \frac{n}{m}\right\}.$$

Even when treatment has no effect on the outcome, the expected magnitude of RIRD is difficult to analyze mathematically.

1.5 Non-overlap of all pairs

NAP is calculated as

$$\text{NAP} = 100\% \times \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left[I\left(Y^B_j > Y^A_i\right) + 0.5I\left(Y^B_j = Y^A_i\right)\right]. \tag{5}$$

Parker and Vannest (2009) noted that the numerator of NAP corresponds to the U statistic from Wilcoxons rank-sum test. If the observations within each phase are identically distributed (but allowing that the distribution in the baseline phase may differ from the distribution in the treatment

phase), then the expected magnitude of NAP is

$$\mathrm{E}\left(\mathrm{NAP}\right) = 100\% \times \left[\Pr\left(Y_1^A < Y_1^B\right) + \frac{1}{2}\Pr\left(Y_1^A = Y_1^B\right)\right].$$

If the observations in both phases are independent and identically distributed, then the expected magnitude of NAP is 50%.

1.6 Tau

As noted in the main text, in the absence of time trends in the data, the Tau statistic is a linear re-scaling of NAP. Specifically,

$$\mathrm{Tau} = 2 \times \frac{\mathrm{NAP}}{100\%} - 1. \tag{6}$$

If the observations in both phases are independent and identically distributed, then the expected magnitude of Tau is 0.

2 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. The additional results pertain to the percentage of all non-overlapping data (PAND).

Figure S1 depicts the expected magnitude 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 5 min sessions and where incidence is once per minute; it is constructed in the same way as Figure 4 in the main text. Although Parker and colleagues (Parker et al., 2007, 2011) suggested that 50% is the expected magnitude 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 magnitude varies between 59% and 72% 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 S2 depicts the expected magnitude of PAND for varying session lengths and recording systems, based on the subset of results where prevalence is 50%, incidence is once per minute, and the baseline phase includes 10 observation sessions; it is constructed in the same way as Figure 5 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 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

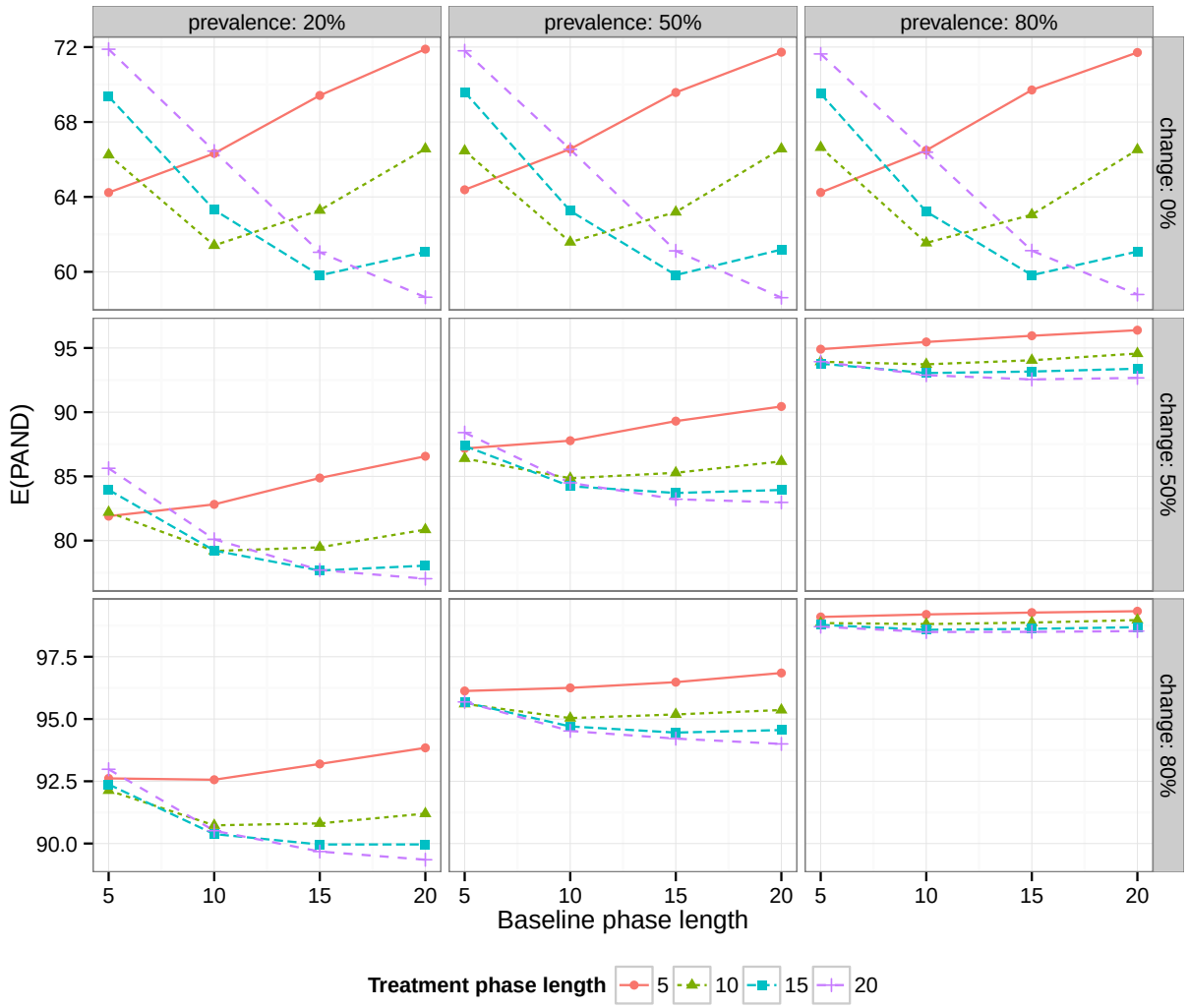

Figure S1: Expected magnitude of PAND based on continuous recording data for 5 min sessions, when incidence is once per minute, for varying baseline and treatment phase lengths.

magnitude of PAND becomes sensitive to the length of the observation session, with longer sessions leading to higher values for PAND. t 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 magnitude 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.

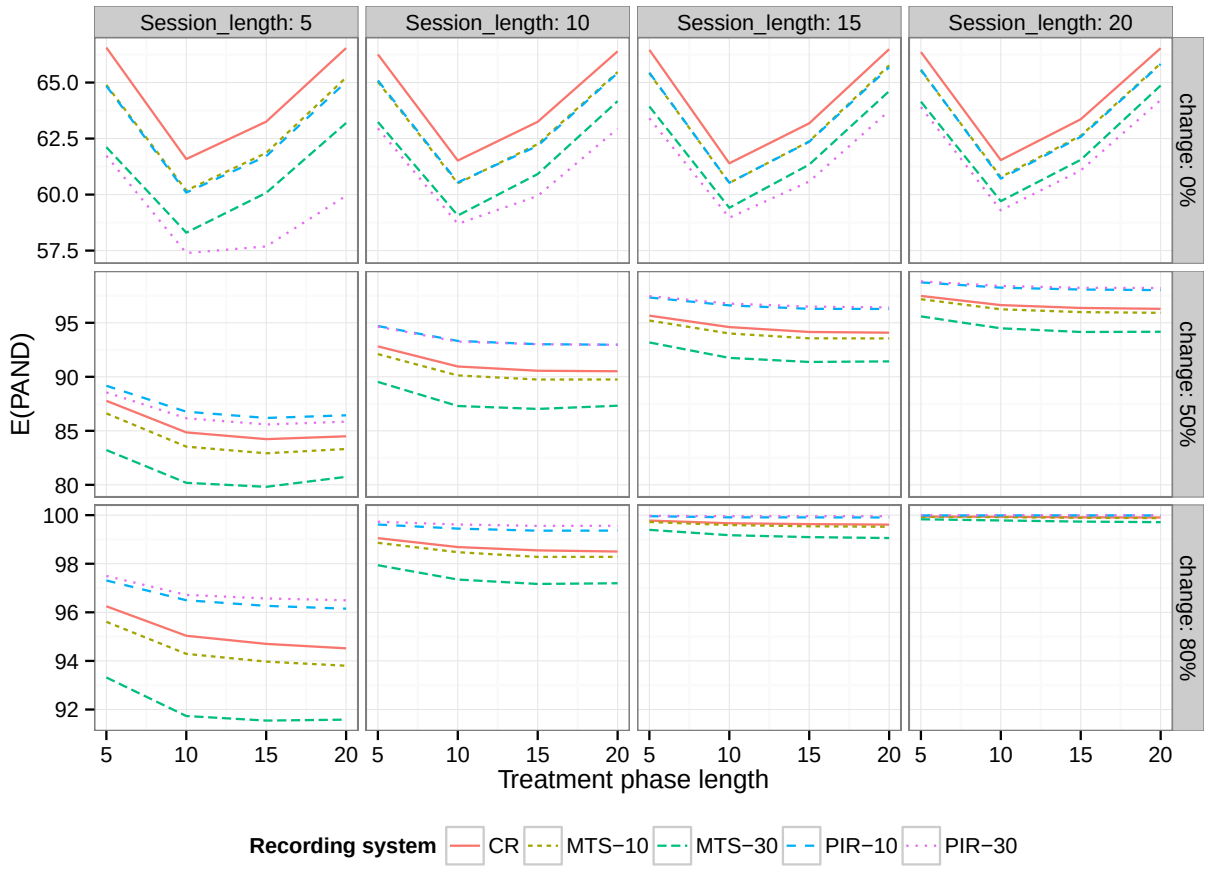

Figure S2: Expected magnitude of PAND based on 10 sessions in the baseline phase, when prevalence is 50% and incidence is once per minute, for varying session lengths and recording systems.

3 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 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 for each of the non-overlap measures.

Table S1: 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
Baseline phase length	5, 10, 15, 20
Treatment phase length	5, 10, 15, 20

3.1 Simulation Design

Table S1 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 S3 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 phase lengths. 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 baseline and treatment phase lengths were set to 5, 10, 15, or 20 sessions.

For each combination of factor levels, 10,000 simulated AB designs were generated and the PND, PAND, RIRD, PEM, and NAP statistics were calculated based on each simulated study. The simulated values of each non-overlap measure were averaged across replications in order to estimate its expected magnitude. The computer code that implements the simulation and full numerical results are available in the supplementary materials that accompany this document, and are described further in Section 4.

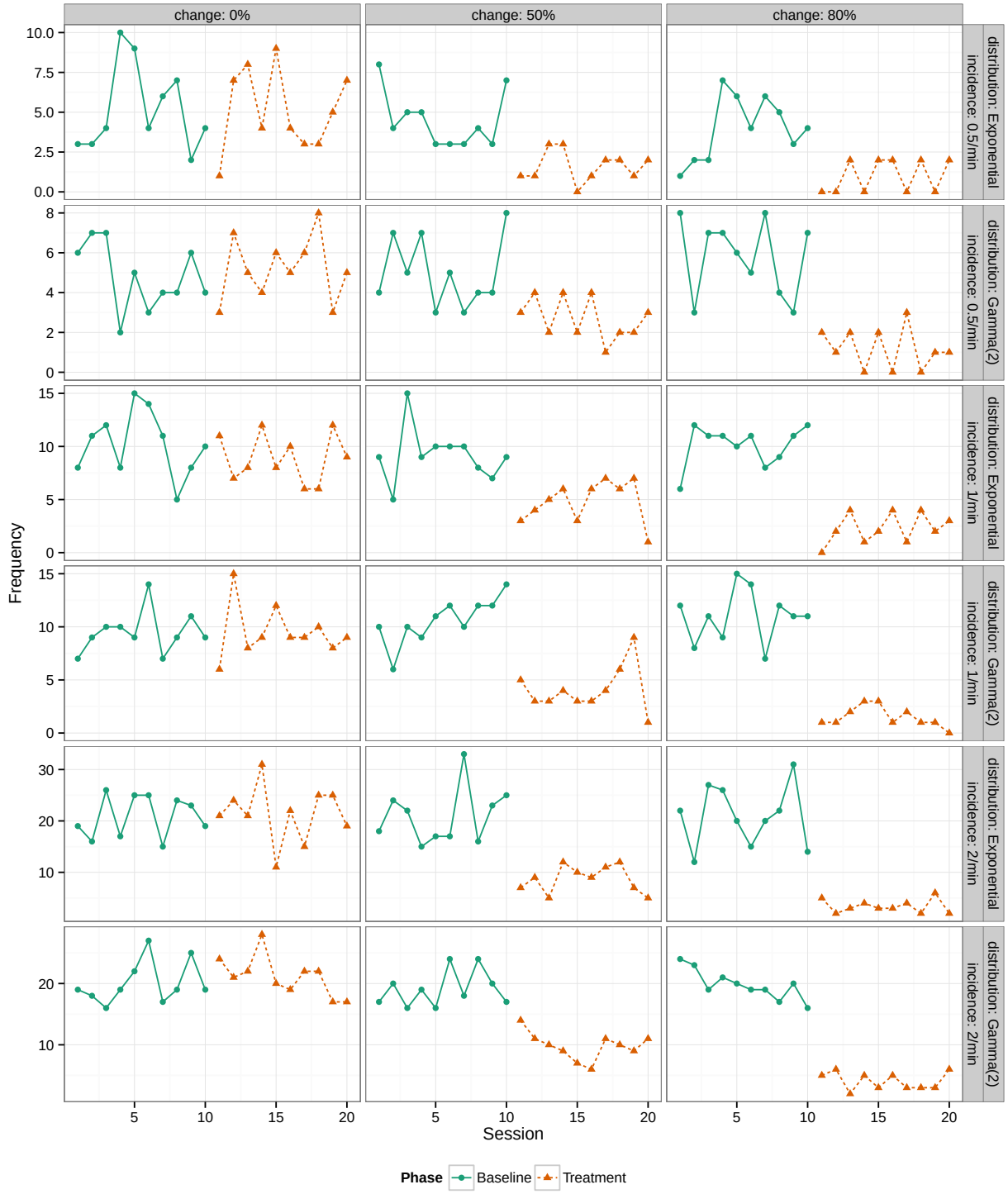

Figure S3: Simulated SCDs based on the alternating renewal process model, using frequency counting for 10 min observation sessions, for varying levels of incidence, inter-response time distribution, and change in behavior

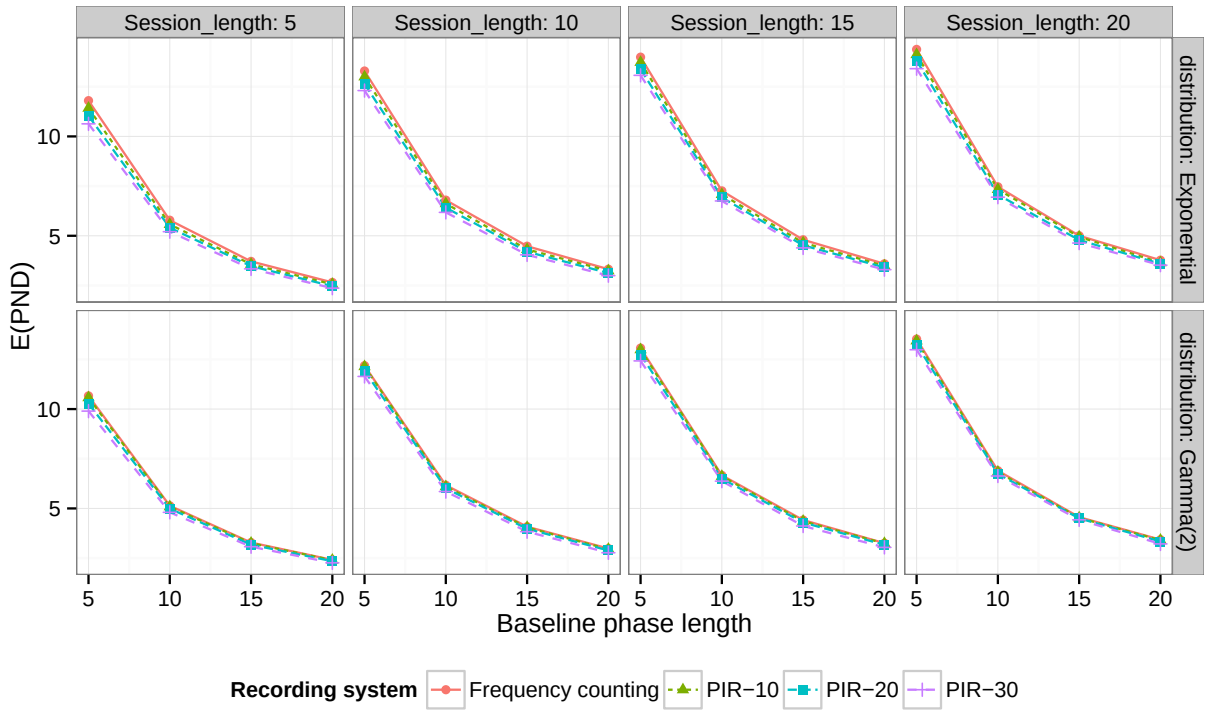

Figure S4: Expected magnitude of PND when incidence is 1/min and treatment has no effect, for varying baseline phase lengths and varying recording systems

3.2 Results

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.

3.2.1 Percentage of non-overlapping data

Figure S4 plots the expected magnitude of PND when incidence is once per minute and treatment has no effect, for varying baseline lengths, recording systems, session lengths, and inter-response time distributions. As with state behavior, the expected magnitude of PND depends on the number of observations in the baseline phase. Its magnitude is affected by session length and recording procedure only slightly, 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 S5 plots the expected magnitude of PND for a 50% change due to treatment, where inter-response times are gamma(2)-distributed. Across all levels of incidence, PND is highly sensitive to the baseline phase length and to the length of the observation session, though it appears to be less affected by

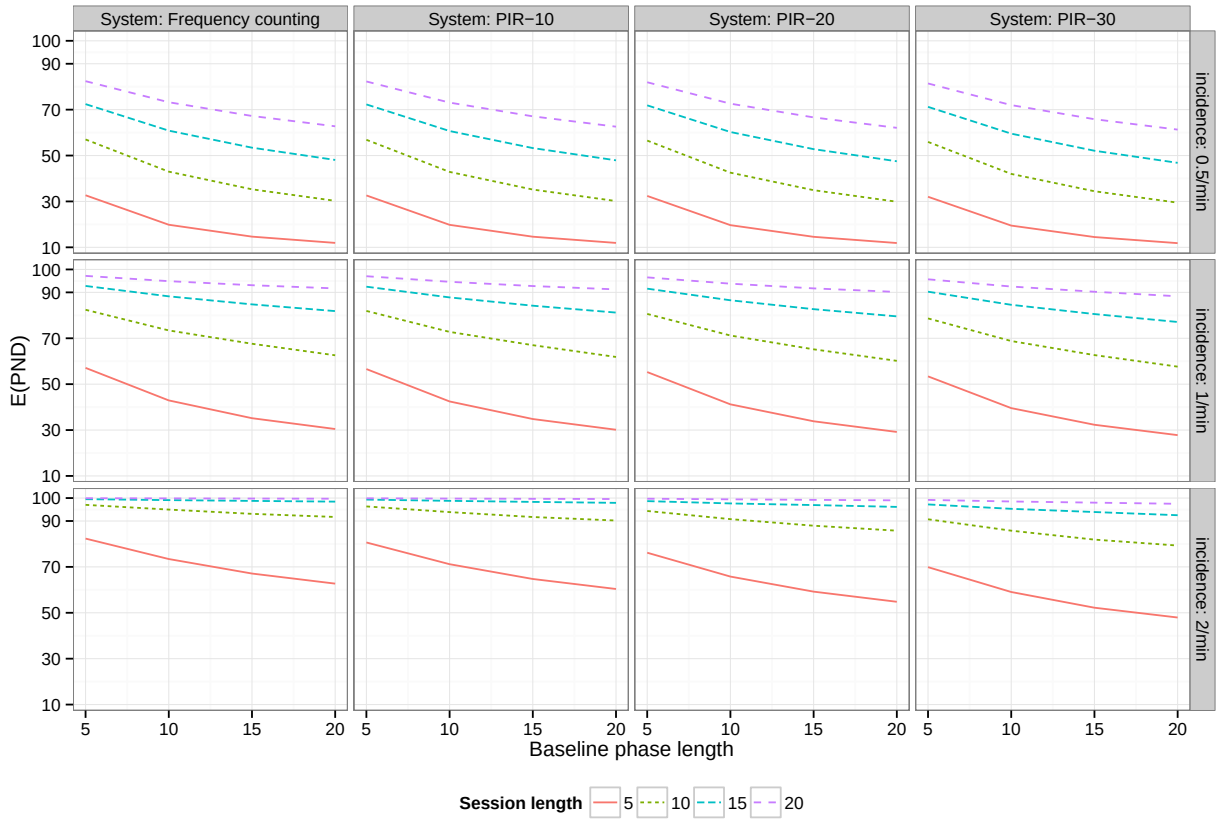

Figure S5: Expected magnitude of PND when inter-response times are gamma(2)-distributed and treatment leads to a 50% change, for varying session lengths and baseline phase lengths

the choice of recording system.

3.2.2 Percentage of all non-overlapping data

Figure S6 depicts the expected magnitude 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 gamma(2)-distributed (results for exponentially distributed inter-response times are very similar). The influence of phase lengths on the magnitude of PAND is quite similar to what was observed in the state behavior simulations (e.g., Figure S1). Just as in those simulations, when treatment has no effect, the magnitude of PAND is sensitive to the baseline and treatment phase lengths. For example, when incidence is once per minute and treatment has no effect, the expected magnitude of PAND varies between 56% and 68%. 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 phase lengths; the decrease in sensitivity is more apparent for higher baseline incidence, though this appears to be due largely to ceiling effects.

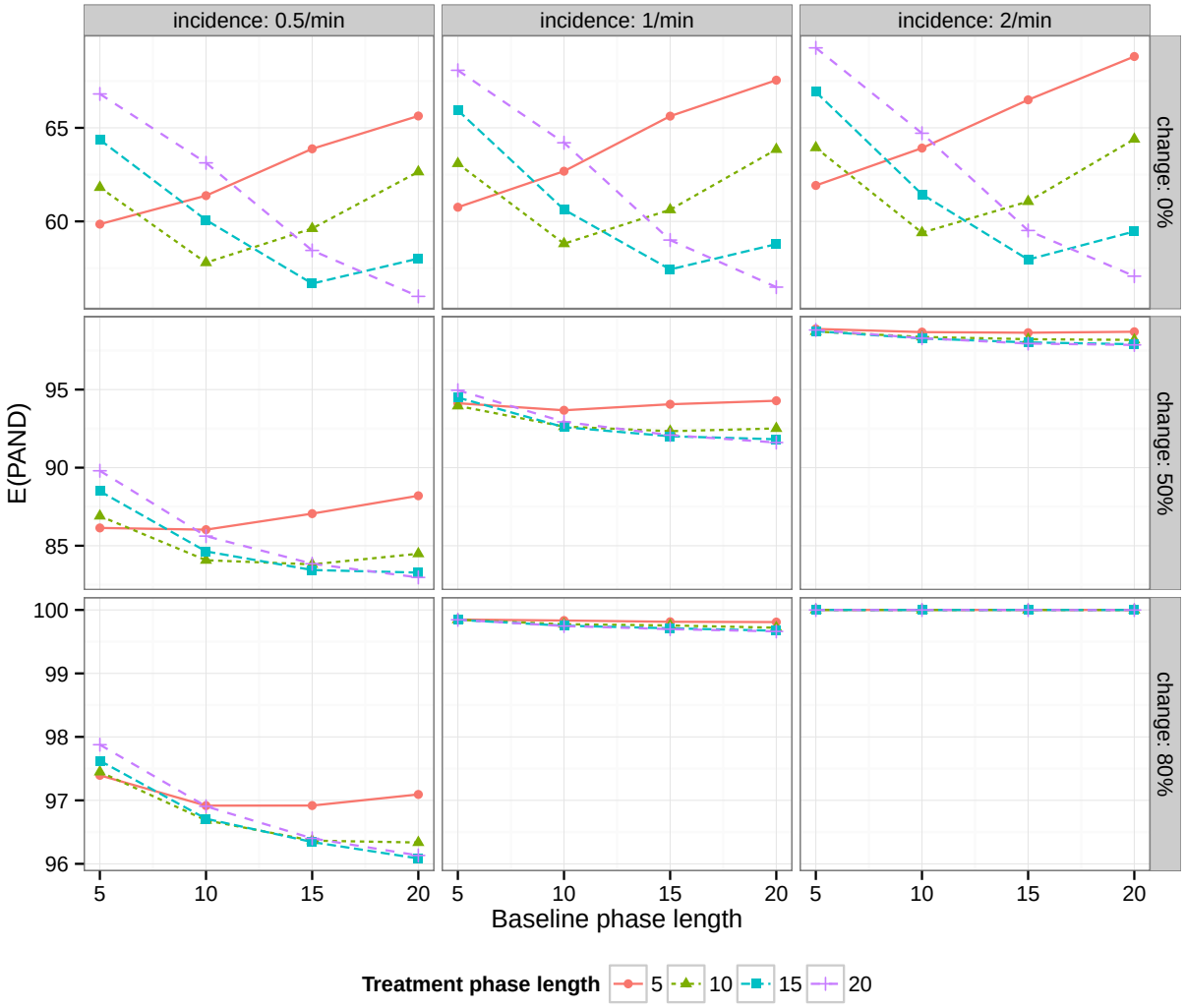

Figure S6: Expected magnitude of PAND based on frequency counting data with 10 min sessions, when the inter-response time distribution is gamma(2), for varying baseline and treatment phase lengths.

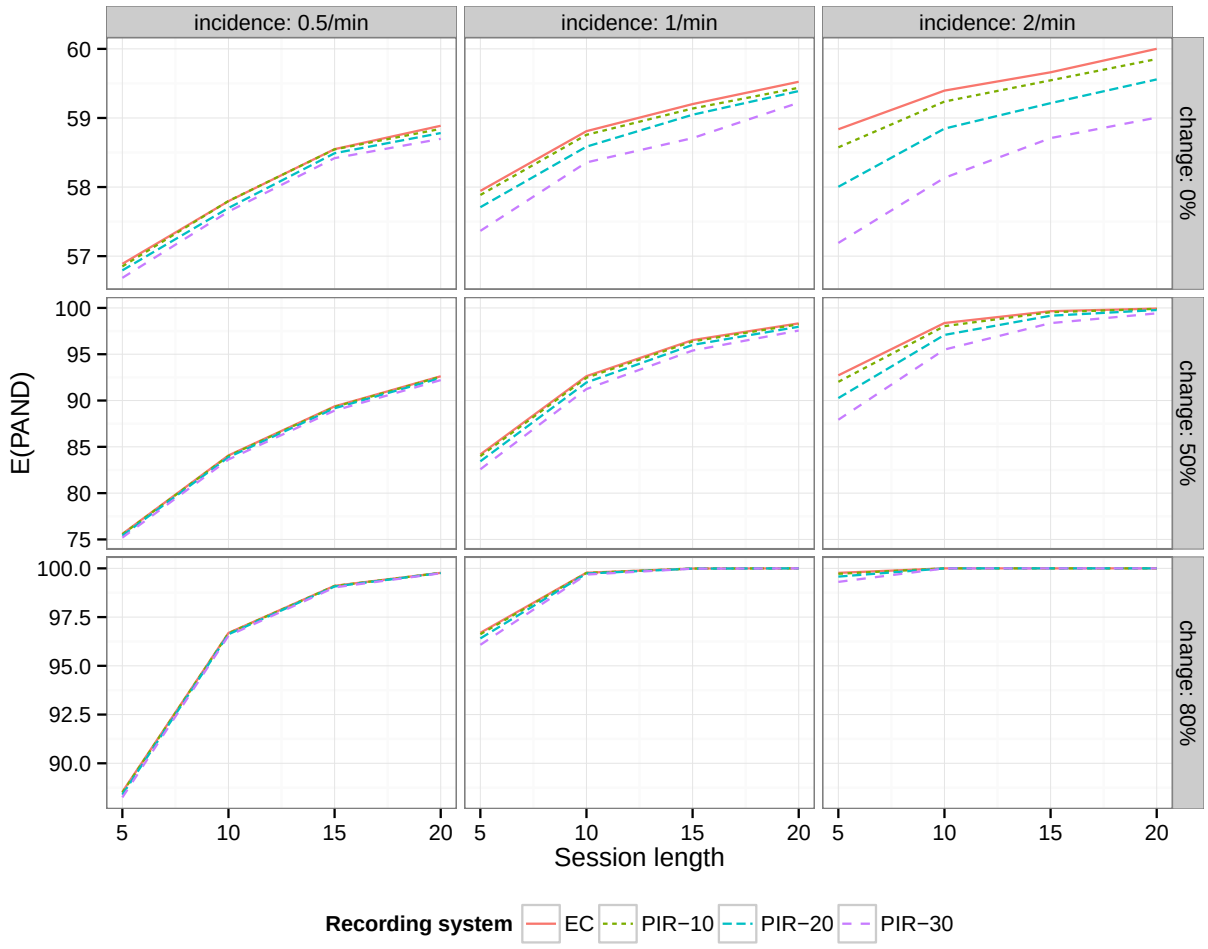

Figure S7: Expected magnitude of PAND when inter-response times are gamma(2)-distributed, based on 10 sessions in the baseline phase and 10 sessions in the treatment phase, for varying session lengths and recording systems.

Figure S7 depicts the expected magnitude of PAND for varying session lengths and recording systems, based on the subset of results where inter-response times follow a gamma(2) distribution and both baseline and treatment phases include 10 observation sessions. Across all levels of incidence and change in behavior, the expected magnitude 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.

3.2.3 Robust improvement rate difference

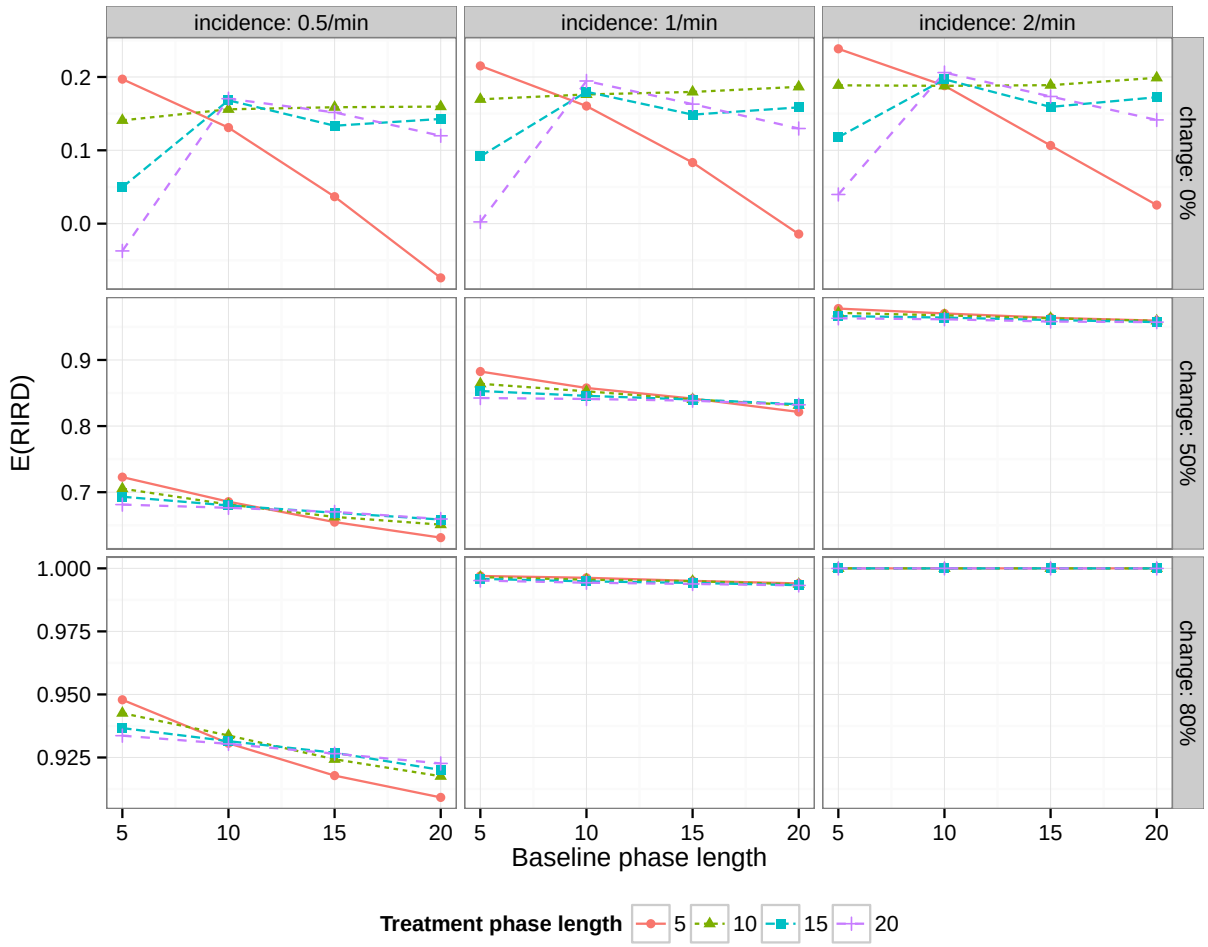

Figure S8: Expected magnitude of RIRD based on frequency counting data with 10 min sessions, when the inter-response time distribution is gamma(2), for varying baseline and treatment phase lengths.

Figure S8 depicts the expected magnitude 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 gamma(2)-distributed; it is constructed in the same way as Figure S6 for PAND. The top row of the figure indicates that the expected magnitude of RIRD varies considerably when the treatment has no effect; for instance, when incidence is once per two minutes, the expected magnitude ranges from -0.07 to 0.2. 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 0.63 to 0.72. In contrast to PAND, the magnitude of RIRD tends to be larger when the baseline and treatment phase lengths are closer to equal.

Figure S9 depicts the expected magnitude of RIRD for varying session lengths and recording systems, based on the subset of results where inter-response times follow a gamma(2) distribution

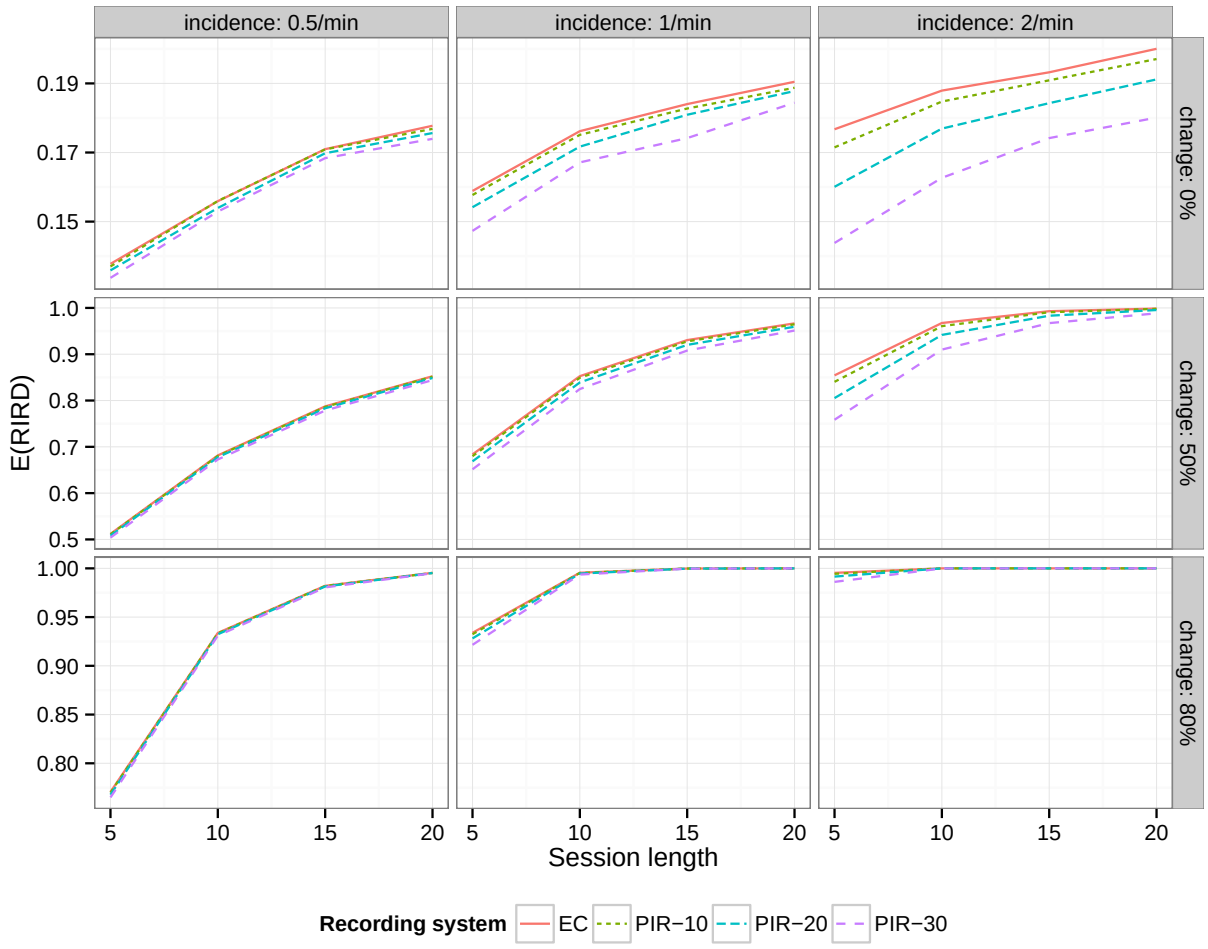

Figure S9: Expected magnitude of RIRD when inter-response times are gamma(2)-distributed, based on 10 sessions in the baseline phase and 10 sessions in the treatment phase, for varying session lengths and recording systems.

and both baseline and treatment phases include 10 observation sessions. The results closely parallel those for PAND (e.g., Figure S7). Like PAND, the expected magnitude 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, when incidence is twice per minute, treatment has no effect on the behavior, and observation sessions last 5 min, the expected magnitude of RIRD varies from 0.14 to 0.18, a fairly narrow range. If treatment produces a 50% decrease from a baseline incidence of twice per minute and sessions last 5 min, the range widens slightly, from 0.76 to 0.85. Also, the degree of sensitivity tends to be reduced when session lengths are longer.

3.2.4 Percentage exceeding the median

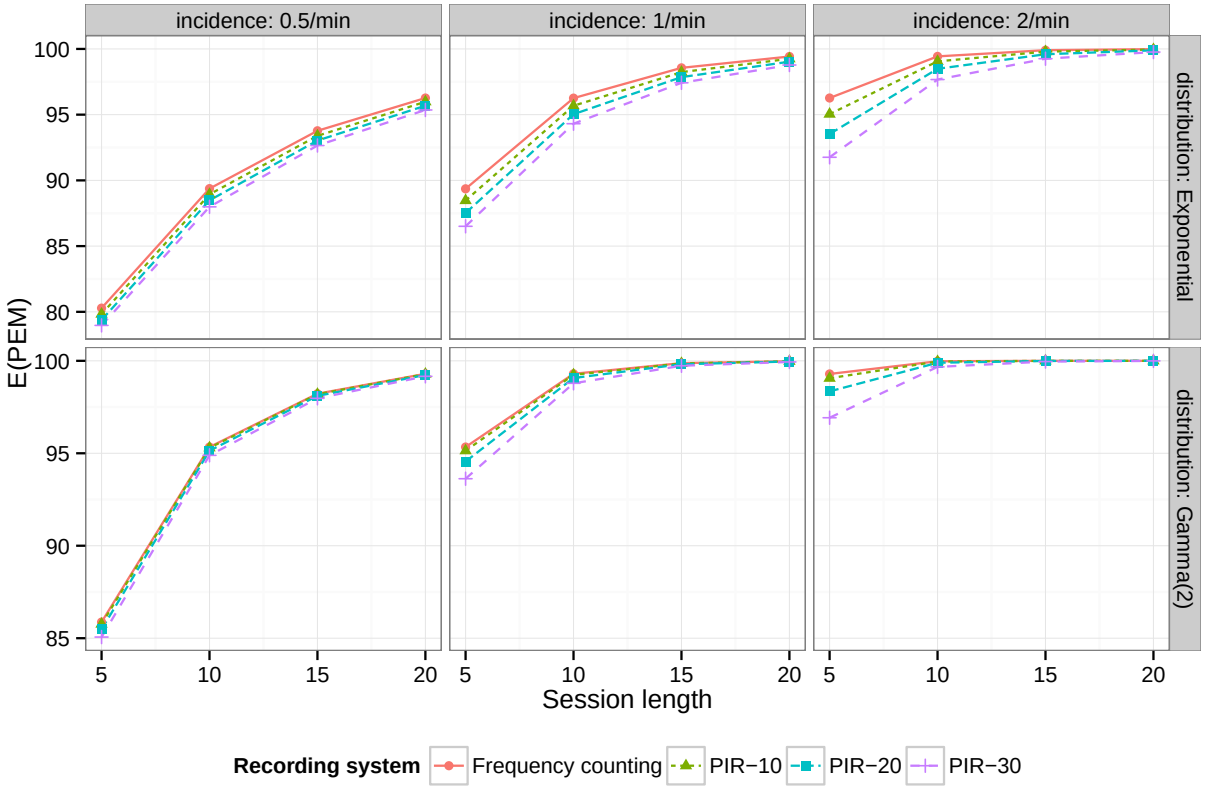

Figure S10: Expected magnitude of PEM for various recording systems and session lengths, when treatment leads to a 50% change in behavior

When treatment has no effect on the behavior, the expected magnitude of PEM is always exactly 50%, across all levels of the other factors. Furthermore, when treatment led to an 80% decrease in incidence of the behavior, the expected magnitude of PEM was at or near the ceiling level of 100% across all levels of the other factors. Figure S10 plots the expected magnitude of PEM when treatment leads to a 50% decrease in incidence of the behavior. The results are averaged across levels of the baseline and treatment phase lengths, 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 magnitude 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.

3.2.5 Non-overlap of all pairs

The behavior of NAP is very similar to that of PEM. Like PEM, the expected magnitude of NAP is exactly equal to 50% when treatment has no effect on the outcome and its expected magnitude is

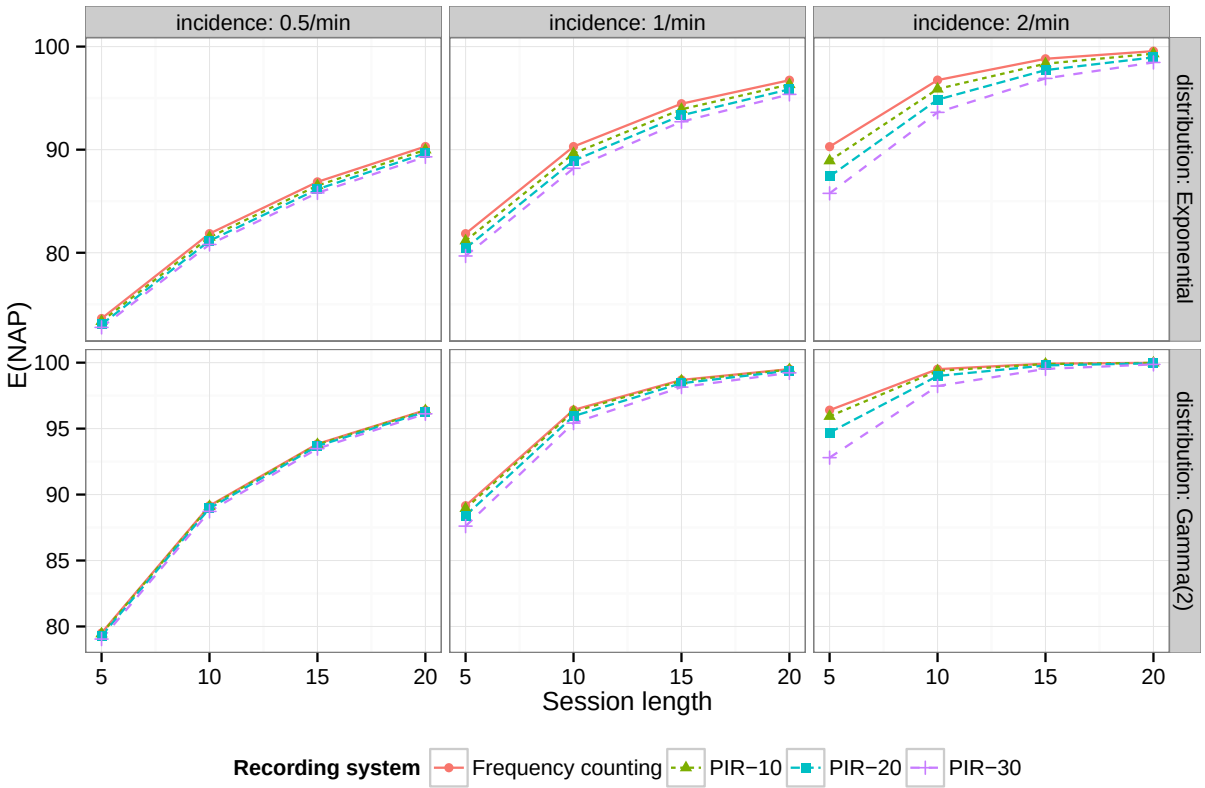

Figure S11: Expected magnitude of NAP for various recording systems and session lengths, when treatment leads to a 50% change in behavior

not affected by the number of sessions in the baseline phase or treatment phase. Also like PEM, the expected magnitude of NAP was at or near the ceiling level when treatment led to an 80% decrease in incidence of the behavior. Figure S11 plots the expected magnitude 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).

3.3 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 and NAP are insensitive to phase lengths but are affected by the length of observation sessions.

The main difference in findings between the event behavior and state behavior simulations concerns the degree to which the non-overlap measures are sensitive to the recording system. Across

all five measures, it appeared that the choice of recording system had 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.

4 Simulation replication materials

In addition to the material presented in this document, the supplementary materials include two additional files. The file `NOM_sims.R` contains the R script that was used to execute both the state behavior and event behavior simulations. The file `NonOverlapMeasures.csv` is a comma-separated value file containing the complete numerical results of the simulations. The file contains 14 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 magnitude of the percentage of non-overlapping data, estimated based on 10,000 simulated AB designs.

11. `PEM` - expected magnitude of the percentage exceeding the median, estimated based on 10,000 simulated AB designs.

12. `PAND` - expected magnitude of the percentage of all non-overlapping data, estimated based on 10,000 simulated AB designs.

13. `IRD` - expected magnitude of the improvement rate difference, estimated based on 10,000 simulated AB designs.

14. `NAP` - expected magnitude of the non-overlap of all pairs, estimated based on 10,000 simulated AB designs.

References

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., & 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

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