Practice Innovations
2016, Vol. 1, No. 4, 221–233

2377-889X/16/$12.00 http://dx.doi.org/10.1037/pri0000029

Outcome Assessment and Inference With the Percentage of Nonoverlapping Data (PND) Single-Case Statistic

Kevin R. Tarlow
Texas A&M University

Andrew Penland
Western Carolina University

Single-case experimental designs allow practitioners to conduct clinical outcomes research without the large samples and substantial resources required by randomized clinical trials. Single-case designs have been used to conduct outcomes research for many decades; however, the statistical measurement of treatment effect sizes remains an unresolved issue. The percentage of nonoverlapping data (PND) is one widely used statistic for effect size measurement of single-case experimental designs. Despite its limitations, PND is useful because it is easy to calculate and interpret. However, null hypothesis significance testing (i.e., the use of *p* values) is not currently feasible with PND because it has an unknown sampling distribution. A method to calculate *p* values for PND is introduced and discussed. An online calculator and statistical computing code are also made available to single-case investigators who wish to calculate *p* values for their data. Calculating PND and its associated *p* values may provide practitioners with valuable insights about their treatment outcomes when PND is used appropriately and its statistical assumptions are not violated.

Keywords: clinical outcomes assessment, single-case, single-subject, percentage of nonoverlapping data

Clinical practitioners across a range of mental health fields face pressure to demonstrate evidence of treatment efficacy but are often not well served by basic outcomes research (Goldfried & Wolfe, 1996; Kazdin, 2008; Messer, 2004). Single-case research designs offer independent practitioners one way of validating treatment effects without the large samples and resources required by randomized clinical trials (Morgan & Morgan, 2001). A growing number of practitioners and applied researchers have implemented single-case experimental designs in numerous fields, including behavior therapy, special education, school psychology, speech-language pathology, disability treatment, sport psychology, psychotherapy, physical therapy, occupational therapy, and other allied health fields (Smith, 2012). One obstacle to wider implementation of these designs is the lack of accepted statistical methods for evaluating single-case data and measuring treatment effects (Shadish, 2014a). An array of quantitative methods is available to single-case investigators, with statistics varying in complexity, accessibility, and flexibility (Campbell, 2004; Gast, 2010; Gorman & Allison, 1996; Kratochwill & Levin, 1992; Parker et al., 2005; Parker, Vannest, & Davis, 2011; Shadish, 2014b).

There are many distinct single-case experimental designs; however, the interrupted time series "AB" design is the most commonly used (Shadish & Sullivan, 2011; Smith, 2012). In an AB single-case design, the investigator first records a series of measurements from the patient prior to treatment. This baseline phase is then used to evaluate a second phase of observations that occur during or after treatment. When a stable (i.e., flat, predictable) baseline is attained, the baseline and treatment phases may be compared to determine if the patient's outcomes after treatment differed from what was expected given the baseline phase.

This article was published Online First September 26, 2016.

Kevin R. Tarlow, Department of Educational Psychology, Texas A&M University; Andrew Penland, Department of Mathematics and Computer Science, Western Carolina University.

Correspondence concerning this article should be addressed to Kevin R. Tarlow, Department of Educational Psychology, Texas A&M University, 4225 TAMU, College Station, TX 77843-4225. E-mail: krtarlow@gmail.com

For example, consider the hypothetical AB design data in Figure 1, which illustrates a patient's depression scores as measured by the second edition of the Beck Depression Inventory (BDI-II; Beck, Steer, & Brown, 1996). The hypothetical BDI-II data for Figure 1 are {25, 23, 26, 19 / 20, 19, 16, 9, 12}. A cursory visual inspection of the data suggests treatment could account for the decrease in depression score. Further statistical analysis would help the investigator determine the size of treatment effect.

The percentage of nonoverlapping data (PND; Scruggs, Mastropieri, & Casto, 1987) is among the most widely used effect size statistics for single-case experimental designs (Beretvas & Chung, 2008; Maggin, O'Keeffe, & Johnson, 2011; Parker et al., 2011; Schlosser, Lee, & Wendt, 2008). PND has endured debate about its limitations (Allison & Gorman, 1993; Ma, 2006; Salzberg, Strain, & Baer, 1987; Wolery, Busick, Reichow, & Barton, 2010) and merits (Manolov & Solanas, 2008, 2009; Schlosser et al., 2008; Scruggs & Mastropieri, 1994b, 1998, 2001). Despite criticism, it remains a popular measure for single-case research. PND's utility and longevity are due in part to its easy calculation and straightforward interpretation. It also correlates well with visual analysis under certain circumstances (Ma, 2006; Scruggs & Mastropieri, 1994a; Wolery et al., 2010). Compared to other popular single-case statistics, PND is relatively simple and accessible to investigators interested in empirically evaluating their data.

For an AB single-case design, PND is calculated by dividing the number of "nonoverlapping" treatment phase scores by the total number of scores in the treatment phase—thus, PND is a "percentage of nonoverlap." Nonoverlapping treatment phase scores are the data points that exceed the most extreme score in the baseline phase. Put more concretely, when treatment is expected to cause an increase in scores, nonoverlapping treatment phase scores will be greater than the maximum score in the baseline phase; when treatment is expected to cause a decrease in scores, nonverlapping treatment phase scores will be less than the minimum score in the baseline phase. PND has a range from 0% to 100%, with greater nonoverlap indicating a greater treatment effect size. To illustrate the calculation of PND, for the interrupted time series in Figure 1, PND = 3/5 = 60% (three out of five treatment phase scores are less than the minimum score in the baseline—recall that treatment is expected to decrease depression scores, so the minimum baseline score is used). According to Scruggs and Mastropieri (1998, p. 224), treatments with PND effect size scores below 50 are interpreted as "ineffective," scores 50 to 70 are interpreted as "questionable," scores of 70 to 90 are inter-

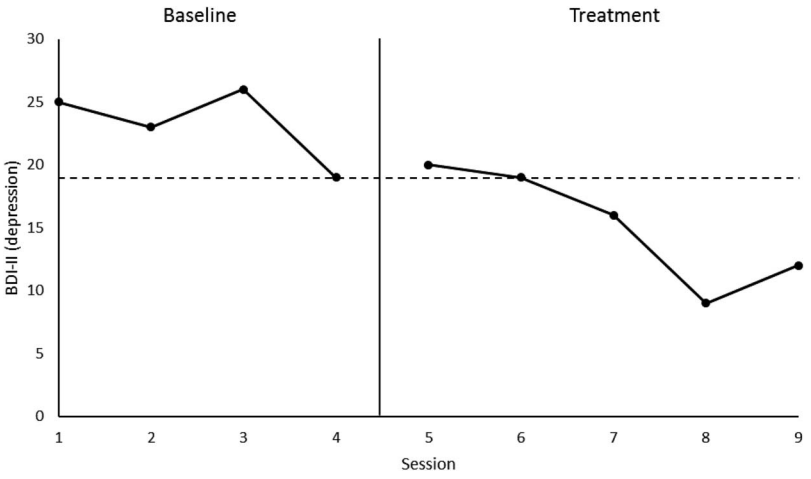

Figure 1. Hypothetical Beck Depression Inventory, 2nd ed. (BDI-II) data of a patient treated for depression. The dashed line indicates the minimum score in the baseline phase (BDI-II = 19).

preted as "effective," and scores above 90 are interpreted as "very effective."

Statistics like PND are used to estimate unobserved parameters, for example, the "true" effect of treatment on outcome. The sampling distribution of a statistic, based on both the statistic and the number of observations (n), describes the likelihood of observing a certain statistical value (e.g., PND = 60%) under the null hypothesis (e.g., no true effect of treatment). It is assumed that all effect sizes could be observed even under the "no effect" null hypothesis because of chance variability in performance and measurement—although large effect sizes are less likely to be observed under the null hypothesis than small ones, assuming equal sample sizes. Investigators can determine if their observed statistical results would be likely or unlikely under the assumption of no treatment effect using p values, which are derived from sampling distributions. If the probability of observing an effect size under the null hypothesis is less than 5% ($p < .05$), investigators often choose to reject the null hypothesis (other critical values of p are frequently used, including .10, .01, .001, etc.). Investigators who fail to reject the null hypothesis would conclude that their observed results could plausibly occur by chance, even without a true treatment effect. Null hypothesis significance testing is a widely used but controversial procedure in research, often due to widespread misunderstanding of its essential inferential logic (Cohen, 1994; Meehl, 1978; Rozeboom, 1960; Thompson, 1999). When reported, p values should be accompanied by the corresponding effect size measure (Wilkinson, 1999).

A limitation of PND is its lack of a known sampling distribution. Because its sampling distribution is unknown, it is difficult to perform null hypothesis significance testing (Beretvas & Chung, 2008). PND p values can be estimated with Monte Carlo simulation techniques (Manolov & Solanas, 2013), though few applied researchers may be able or willing to use those computationally advanced methods. Without a defined sampling distribution, PND is a descriptive statistic whereas other effect size measures are inferential (Allison & Gorman, 1994). Statistical development of nonoverlap methods like PND will improve their usefulness (Shadish, 2014a). This article describes a method of calculating p values for PND by converting observed scores to rank scores. An online calculator also accompanies this manuscript where users can calculate p values for PND (Tarlow & Penland, 2016).

A Method to Calculate p Values for PND

PND can be calculated from either observed (raw) scores or rank scores. Converting raw scores to rank scores (where ranks are assigned across phases, not within phases) does not affect the PND measurement. PND will always yield identical results for raw scores or rank scores. This makes logical sense, given that identifying the nonoverlapping scores in the experimental treatment phase depends not on their exact values but on their ordinal positions relative to the maximum (or minimum) score in the baseline phase. Returning to the example data in Figure 1, the raw scores {25, 23, 26, 19 / 20, 19, 16, 9, 12} and the corresponding rank scores {8, 7, 9, 4.5/6, 4.5, 3, 1, 2} both yield an effect size of PND = 60%. Several popular single-case statistics use rank scores to calculate effect size, including the percentage of data exceeding the median (Ma, 2006), nonoverlap of all pairs (Parker & Vannest, 2009), and Tau-U (Parker, Vannest, Davis, & Sauber, 2011).

In a time series with n total rank scores, and assuming that scores may be tied, there are n^n possible time series permutations. Assuming independent observations (an assumption discussed below in greater detail), every possible time series is equally likely under the null hypothesis of no treatment effect, with each permutation of n scores having a probability of $(1/n^n)$. However, although every n^n sequence of rank scores is equally probable, not all PND values are equally probable. The probabilities of each PND value for a given n comprise the statistic's sampling distribution and may be used to calculate p values for null hypothesis significance testing.

The formula below gives the exact probability of observing an equal or greater PND value under the null hypothesis described above

$$p_{PND}(n_A, n_B, k) = \frac{\sum_{i=k}^{n_B} \sum_{M=1}^{n} \sum_{r=1}^{n_A} \left[C(n_A, r)(M-1)^{(n_A - r)} C(n_B, i)(n-M)^i M^{(n_B - i)} \right]}{n^n}$$

where $n = n_A + n_B$ and $C(x, y)$ represents the binomial coefficient $\frac{y!}{x!(y-x)!}$; n_A is the number of

scores in the baseline phase, n_B is the number of scores in the treatment phase, and k is the number of scores in the treatment phase that exceed the most extreme score in the baseline phase, *M,* so that PND = k/n_B. The argument for this formula is presented in Appendix A.

This method is illustrated by considering a hypothetical AB single-case design with three observed scores; although it is inadvisable to interpret such a short time series, the example is useful for heuristic purposes. Table 1 presents all possible permutations of rank scores for a time series of $n_A = 1$ and $n_B = 2$, along with those series' PND values. For a series with three total scores (in this case, one baseline phase score and two treatment phase scores), there are $n^n = 27$ possible rank score sequences, all of which are listed. Under the null hypothesis, these permutations represent all equally likely sequences of n rank scores, with each row of Table 1 occurring at a rate of 1/27. It is clear that for a time series where $n_A = 1$ and $n_B = 2$, PND may equal 0%, 50%, or 100%. A PND value of 0% occurs in 14/27 or 52% of the cases; a value of 50% occurs in 8/27 or 30% of the cases, and a value of 100% occurs in 5/27 or 18% of the cases. Together, these probabilities form the PND sampling distribution for $n_A = 1$ and $n_B = 2$. The formula above yields the cumulative probability of a PND value that is equal or greater than the observed PND value. For example, the p value of PND = 50% is the sum of all probabilities where PND ≥ 50%, or (8 + 5)/27 = .48. For the $p_{PND}(n_A,n_B,k)$ formula above, $p_{PND}(1,2,1) = .48$. Alternately, Table 2 presents all possible sequences of rank scores for a time series with $n = 3$, but with $n_A = 2$ and $n_B = 1$. For the null hypothesis in Table 2, PND may equal 0%, with $p = (22 + 5)/27 = 1.00$, or 100%, with $p = 5/27 = .18$.

Table 1
PND for All Possible Permutations of Ranks in an AB Design With $N_A = 1$ and $N_B = 2$

Case	Phase A	Phase B		PND
1	1	1	1	0
2	1	1	2	50
3	1	1	3	50
4	1	2	1	50
5	1	2	2	100
6	1	2	3	100
7	1	3	1	50
8	1	3	2	100
9	1	3	3	100
10	2	1	1	0
11	2	1	2	0
12	2	1	3	50
13	2	2	1	0
14	2	2	2	0
15	2	2	3	50
16	2	3	1	50
17	2	3	2	50
18	2	3	3	100
19	3	1	1	0
20	3	1	2	0
21	3	1	3	0
22	3	2	1	0
23	3	2	2	0
24	3	2	3	0
25	3	3	1	0
26	3	3	2	0
27	3	3	3	0

Note. Percentage of nonoverlapping data (PND) is calculated for treatment effects in the positive (increasing) direction.

Applications

Web Application

A free web application developed concurrently with this paper uses the method presented above to calculate p values for PND. The application is available for use at http://www.ktarlow.com/stats/pnd (Tarlow & Penland, 2016). The use of this calculator is straightforward: the user simply inputs the n_A, n_B, and k values for a time series, and the program calculates PND and p, as illustrated in Figure 2.

R Software Script

A software script for the free R statistical analysis software (R Core Team, 2014) was also developed to perform the same function as the online p value calculator. The script is available in Appendix B. Use of this script is straightforward. First, the user loads the script by copying and pasting the syntax into the R console. The user can then execute the function

```
pnd(nA, nB, k)
```

replacing the arguments n_A, n_B, and k with the number of scores in the baseline phase, the number of scores in the treatment phase, and the number of treatment phase scores that

Table 2
PND for All Possible Permutations of Ranks in an AB Design With $N_A = 2$ and $N_B = 1$

Case	Phase A		Phase B	PND
1	1	1	1	0
2	1	1	2	100
3	1	1	3	100
4	1	2	1	0
5	1	2	2	0
6	1	2	3	100
7	1	3	1	0
8	1	3	2	0
9	1	3	3	0
10	2	1	1	0
11	2	1	2	0
12	2	1	3	100
13	2	2	1	0
14	2	2	2	0
15	2	2	3	100
16	2	3	1	0
17	2	3	2	0
18	2	3	3	0
19	3	1	1	0
20	3	1	2	0
21	3	1	3	0
22	3	2	1	0
23	3	2	2	0
24	3	2	3	0
25	3	3	1	0
26	3	3	2	0
27	3	3	3	0

Note. Percentage of nonoverlapping data (PND) is calculated for treatment effects in the positive (increasing) direction.

exceed the most extreme score in the baseline phase, respectively.

Both R script and Web-based calculator yield exact p values of PND for small or large time series. These applications perform well even for series with as many as 100 scores, where other methods such as Monte Carlo simulation would become cumbersome, and brute force enumeration of all rank score permutations (as in Tables 1 and 2) is not feasible.

Assumptions and Limitations

PND's most significant limitation is its inability to model baseline trend or autocorrelated data. The presence of upward or downward trend has the potential to distort PND effect size measurements and, consequently, p values. PND also demonstrates a ceiling effect, discriminating poorly among data sets with large treatment effects; for example, the time series {1, 1, 1, 1, 1/2, 2, 2, 2, 2} and {1, 1, 1, 1, 1/10, 10, 10, 10, 10} both yield PND = 100% (their corresponding rank scores are identical), though the second time series demonstrates a much larger treatment effect. PND is also sensitive to outliers in the baseline phase, which, when they occur in the expected direction of the treatment effect, will attenuate the effect size result. A less discussed but equally important limitation is PND's reported tendency to decrease as the number of baseline observations increases (Allison & Gorman, 1994). Several of PND's most salient limitations are discussed below in their relationship to null hypothesis significance testing. When a single-case investigator has a good experimental design and awareness of the measurement's statistical limitations, PND and its associated p values are a reasonable choice for effect size measurement. The use of p values can in fact counteract some of PND's limitations. Three guidelines for "best practice" with PND and its associated p values are also presented below.

Baseline Trend

As was pointed out by its originators (Scruggs et al., 1987, pp. 28–29), PND assumes there is no trend (e.g., slope) in single-case data. The method of calculating p values presented above makes the same statistical assumption, as do most single-case statistics based on rank scores or data nonoverlap. If baseline trend occurs in the expected direction of treatment effect, PND will be distorted upward and its p value will be distorted downward, leading the investigator to incorrectly conclude a larger, more statistically significant effect than actually present (i.e., Type I error). For example, the hypothetical BDI-II data in Figure 3 clearly exhibits a preexisting baseline trend that distorts the effect measured by PND (PND = 90%, $p <$ .001). It cannot be assumed this large statistically significant "effect" is a result of treatment but rather the result of the patient's improvement, which appears independent of treatment.

Like PND, many other single-case statistics fail to model trend (e.g., Parker et al., 2011; Wolery et al., 2010). Ideally, this limitation is managed by obtaining a stable baseline. When a stable baseline is obtained, even with fluctuations around a mean level of performance, "the

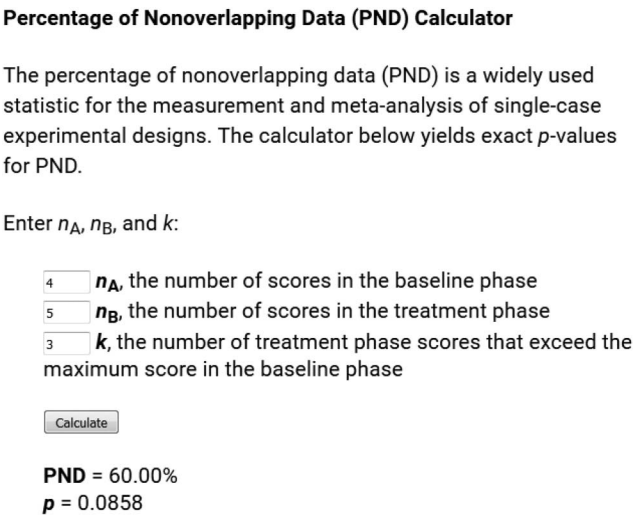

Figure 2. Screenshot of percentage of nonoverlapping data (PND) calculator available at http://www.ktarlow.com/stats/pnd (Tarlow & Penland, 2016).

absence of trend provides a clear basis for evaluating intervention effects" (Kazdin, 1982, p. 107). In many research settings, establishing a stable rate of performance in the baseline is not feasible, and in those cases trend should be statistically controlled or another more appropriate effect size measure should be used. However, achievement of flat baseline data is quite common in clinical research areas such as school psychology and behavior therapy. In many instances, investigators identify patients for treatment specifically because some target behavior is not changing (i.e., baseline responding is stable); the goal of treatment is often to introduce a positive behavior where none existed, or to eliminate a persistent unwanted behavior. Nonoverlap methods like PND are well suited for designs such as these.

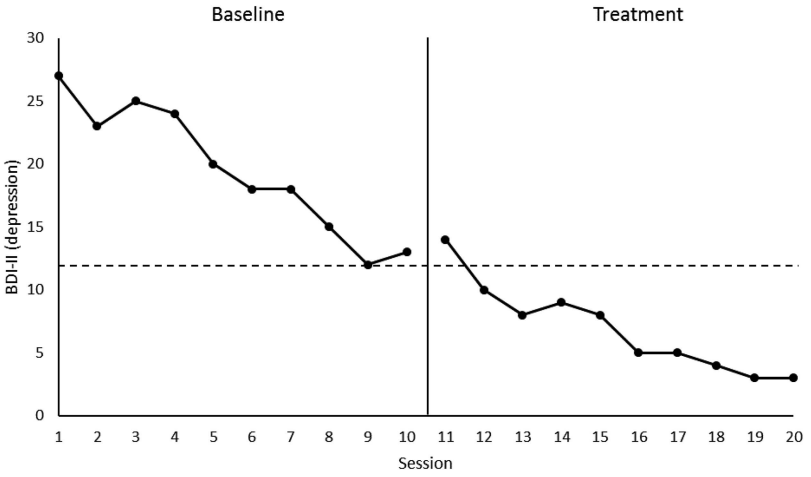

Figure 3. Hypothetical Beck Depression Inventory, 2nd ed. (BDI-II) data with improving baseline trend. Percentage of nonoverlapping data (PND) would be an inappropriate effect size measure as it yields a large value (PND = 90%) and small p value ($p < .001$) despite the visually apparent absence of treatment effect.

When investigators are unable to obtain a flat baseline, but are able to observe a predictable rate of change in the baseline phase, another option is to use some method of data correction that removes the influence of baseline trend from the entire time series. Manolov and Solanas (2009) introduced a baseline trend correction method for PND, the percentage of nonoverlapping corrected data (PNCD), which may be combined with the null hypothesis significance testing method presented above, provided some conditions are met. In the PNCD method, baseline trend is corrected via a two-step process: first, a baseline trend coefficient is estimated; second, both A and B phases are adjusted to account for the baseline trend. In the first step, the A phase data points are differenced as $n_{t+1} - n_t$, yielding n_{A-1} differenced data points. The mean of this differenced series equals the baseline trend coefficient, b. In the second step, a corrected AB series is calculated as $Y_{\text{Corrected}} = Y_t - (b \times T_t)$, where Y_t is the original data point at time t and T_t is the value of the time variable (x-axis value) at time t. PNCD (and its p value) is then calculated using the PND method, but with the corrected AB time series.

Regardless of the correction method, baseline trend correction should be used only when a sufficient number of baseline observations demonstrate a *stable* trend, as the random fluctuations of a few baseline scores can result in overcorrected or miscorrected data. Investigators may then use the corrected n_A, n_B, and k values to calculate a p value for the observed effect. This method of baseline trend correction is demonstrated in Figure 4, which shows PNCD applied to the time series from Figure 3. After scores are corrected for the stable, linear baseline trend, the time series no longer violates PND's assumptions, and results of effect size measurement and statistical significance testing are more appropriate (PNCD = 0%, p = 1.00).

When stable baselines are present in single-case data, PND is one reasonable measure of treatment effect, and p values obtained from those data will be accurate. This leads to the first of three recommendations for PND with p values, which is a restatement of Scruggs et al. (1987):

> 1. **Obtain a stable baseline;** if there is evidence of baseline trend, consider using a correction method (e.g., PNCD) or another effect size measure.

Autocorrelation

As both an effect size measure and a statistical test, PND assumes that all possible permutations of rank scores are equally probable under the null hypothesis. It was demonstrated above that this assumption is violated when baseline trend is present in single-case data. The

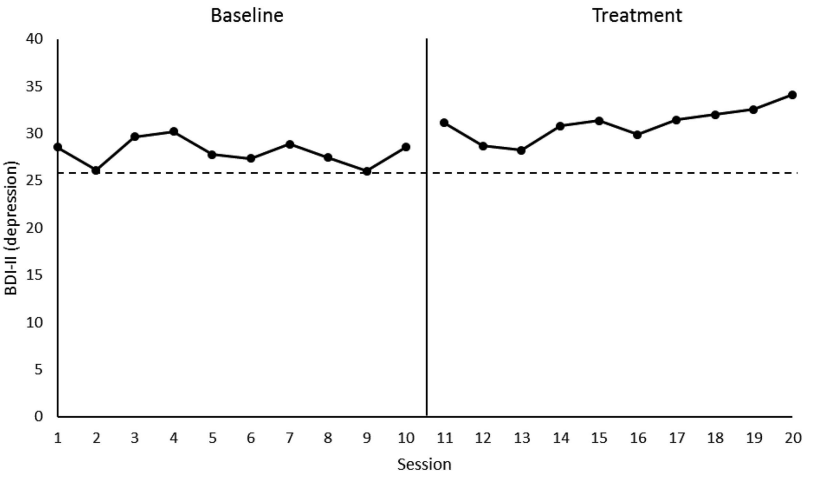

Figure 4. Percentage of nonoverlapping corrected data (PNCD; Manolov & Solanas, 2009) was used to correct baseline trend in Figure 3 single-case data. The result yields a more appropriate effect size (PNCD = 0%) and p value (p = 1.00).

same independence assumption is violated by autocorrelation, which in a time series occurs when variability or "disturbances" in data are carried from one observation into the next. Under these circumstances, the data are considered "serially dependent" because the observations are no longer completely independent of one another. Autocorrelation is a frequent topic of investigation in single-case and time series research because of the challenges it poses to statistical analysis (Bartlett, 1946; Brossart, Parker, Olson, & Mahadevan, 2006; Ferron, 2002; Manolov & Solanas, 2008; Matyas & Greenwood, 1996; Wampold, 1988).

Using simulation methods, Manolov and Solanas (2008) found that PND was the statistic least affected by the presence of autocorrelation when compared to several other single-case effect size measures. Another study by Manolov and Solanas (2013) found that PND's p values (estimated via simulation) were similarly unaffected by autocorrelation. PND is thus an attractive option for single-case investigators concerned about serial dependency in their data.

Small and Large n Time Series

Allison and Gorman (1994) stated that the value of PND is unacceptably dependent on the number of baseline observations. They argue the relationship between PND and the n of baseline observations is especially problematic with very small or very large time series. Allison and Gorman used simulation data to point out that with a small number of baseline observations, PND tends to yield moderate effect sizes even in the absence of experimental effect, that is, under the null hypothesis. They also show that as the baseline n grows very large, the PND value decreases. Together, these two tendencies may have the unwelcome effect of discouraging investigators from obtaining as many baseline scores as possible. Essentially, Allison and Gorman made an "important distinction between an effect size and a statistical test" (p. 887):

> An effect size is an estimate of the magnitude of a relationship, independent of sample size. A statistical test establishes the probability that a given effect size could have occurred beyond chance expectations for a given sample size. One can have large effect sizes that are statistically insignificant and statistically significant tests that accompany miniscule effect sizes. (p. 887)

Until now, PND could not account for the chance expectations that occur under the null hypothesis, that is, PND was an effect size but not a statistical test. Large PND effects occurring by chance, especially with few observations, would be indistinguishable from large PND effects with small probabilities under the null hypothesis. The problem is therefore the lack of a corresponding statistical test that can provide more context for interpreting the PND effect size.

Allison and Gorman (1994) demonstrate that when the number of baseline points is three, the estimated average PND $\approx$ 21% under the null hypothesis (p. 886). Without a p value, this might erroneously lead an investigator to believe their intervention yielded a small-but-real effect even when no effect existed, and their results were merely the product of chance variability in their sample. However, when a p value for these data is calculated with the method presented above, for example, a time series with $n_A = 3$, $n_B = 10$, and $k = 2$, then PND = 20% with $p = .50$. With the added context of a statistical test, the investigator would fail to reject the null hypothesis. This holds true even when large treatment phases are conducted with short baseline observations: for $n_A = 3$, $n_B = 100$, and $k = 20$, then PND = 20% with $p = .51$. This leads to the second recommendation for PND with p values:

> 2. **Maximize the number of baseline observations to increase the statistical power of the test**.

Allison and Gorman (1994) also used simulation data to show that, as the number of baseline observations grows larger, the probability of an outlier in the baseline phase grows more likely, and thus the value of PND asymptotically approaches zero as a function of n_A. However, although these findings are true under the conditions of a computer simulation, they may not accurately describe the real world circumstances under which single-case clinical research is often conducted. In research areas where single-case experimental designs are popular (e.g., behavior therapy and school psychology), the risk of baseline outliers is overstated. For example, when clinicians wish to introduce a new target behavior that has never been performed by the patient, the likelihood of the patient spontaneously demonstrating the target behavior before treatment is extremely low.

In addition, as stated above, PND time series with more baseline observations have more statistical power (i.e., a greater likelihood of rejecting the null hypothesis). Consider a relatively short time series, with $n_A = 5$, $n_B = 5$, $k = 3$, and PND = 60% ($p = .06$); and a longer time series, $n = 20$, $n_A = 10$, $n_B = 10$, $k = 4$, and PND = 40% ($p = .03$). Here the longer series has a PND effect size smaller by 20%, yet the increased n also yields a statistically significant result, whereas the shorter series with larger effect was not statistically significant at the $\alpha = .05$ level. For many investigators, it is expected that the compromise of marginally smaller effect sizes for statistically significant results will be acceptable. The addition of p values may in fact encourage single-case investigators to obtain longer baselines.

Limitations of AB Single-Case Experimental Designs

Investigators should also consider (and, when possible, attempt to minimize) the experimental limitations of all AB single-case designs, regardless of analytic method. Time series experiments cannot completely control for historical threats to internal validity, that is, confounding events that occur independent of the treatment which nonetheless influence the outcome variable. Investigators may implement one or several tactics to control for historical threats to internal validity, such as the reversal (ABA, ABAB) and multiple baseline designs (Barlow & Hersen, 1984). However, practical limitations often rule out these methods, particularly for independent practitioners whose primary goal is service delivery rather than experimental research. External validity limitations should also be considered when interpreting experimental results. In single-case designs, the unit of analysis is the individual, rather than the randomly sampled group. Although investigators may demonstrate a treatment response beyond the degree expected by chance for the individual, one cannot assume from a sample of $n = 1$ that the result would generalize to all individuals. As in other between-groups experimental designs, study replication is essential for generalizability, particularly for single-case researchers. For further exploration of these limitations, the interested reader is directed to Campbell and Stanley's (1963) brief but illuminating discussion of interrupted time series experiments and their strengths and weaknesses relative to other experimental designs.

Conclusions

In their systematic review of PND reporting, Schlosser et al. (2008) noted the appeal of simple and accessible effect size measures in single-case research:

> PND continues to be a frequently employed metric for aggregating outcomes across studies using [single-case experimental designs]. In fact, it is unlikely that the field has similar implementation experiences for any other metric at this point in time . . . At present, most scholastic energy is directed toward the development of new metrics. Although these efforts are worthwhile, metrics should not be viewed as a panacea. When metrics are discussed in terms of their theoretical strengths and weaknesses alone, divorced from issues of implementation and application, we jeopardize the capability of a particular metric to realize these strengths or perhaps minimize weaknesses, whatever they may be. (p. 184)

PND is an imperfect measure of effect size, and, like every other statistical measure, it is not appropriate for all single-case data and designs. This last, important point leads to the third recommendation for PND with p values: (the bolded statement below is quoted almost verbatim from Allison & Gorman, 1994, p. 888):

> 3. **PND does not describe complex behavior patterns;** when research questions require complex modeling of change over time, use other more appropriate methods.

Not all single-case investigators intend to study complex behavior patterns. Many focus on achievement of straightforward behavioral criteria or clinical cutoffs. It is mostly for this reason that PND's enduring popularity should not be ignored. PND's accessibility and utility to a wide range of practitioner-researchers are strengths just as its statistical assumptions are weaknesses. This article has aimed to augment the popular use of PND with long awaited p values. To make this method as accessible to practitioners as possible, a Web-based application and free R software script were made available to those who wish to quickly and easily apply statistical tests to their clinical outcomes. In addition, three PND "best practice" guidelines were presented, including reminders to obtain a stable baseline when possible and to maximize the number of baseline observations. It was also demonstrated that incorporating p values into PND measurement can encourage investiga-

tors to follow some of these best practice guidelines.

References

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research and Therapy, 31,* 621–631. http://dx.doi.org/10.1016/0005-7967(93)90115-B

Allison, D. B., & Gorman, B. S. (1994). "Make things as simple as possible, but no simpler." A rejoinder to Scruggs and Mastropieri. *Behaviour Research and Therapy, 32,* 885–890. http://dx.doi.org/10.1016/0005-7967(94)90170-8

Barlow, D. H., & Hersen, M. (1984). *Single-case experimental designs: Strategies for studying behavior change* (2nd ed.). Elmsford, NY: Pergamon Press.

Bartlett, M. S. (1946). On the theoretical specification and sampling properties of autocorrelated time-series. *Journal of the Royal Statistical Society, 8*(Suppl.), 27–41. http://dx.doi.org/10.2307/2983611

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck Depression Inventory manual* (2nd ed.). San Antonio, TX: Psychological Corporation.

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

Brossart, D. F., Parker, R. I., Olson, E. A., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification, 30,* 531–563. http://dx.doi.org/10.1177/0145445503261167

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

Campbell, J. M. (2004). Statistical comparison of four effect sizes for single-subject designs. *Behavior Modification, 28,* 234–246. http://dx.doi.org/10.1177/0145445503259264

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003. http://dx.doi.org/10.1037/0003-066X.49.12.997

Ferron, J. (2002). Reconsidering the use of the general linear model with single-case data. *Behavior Research Methods, Instruments & Computers, 34,* 324–331. http://dx.doi.org/10.3758/BF03195459

Gast, D. L. (2010). *Single subject research methodology in behavioral sciences*. New York, NY: Routledge.

Goldfried, M. R., & Wolfe, B. E. (1996). Psychotherapy practice and research: Repairing a strained relationship. *American Psychologist, 51,* 1007–1016. http://dx.doi.org/10.1037/0003-066X.51.10.1007

Gorman, B. S., & Allison, D. B. (1996). Statistical alternatives for single-case designs. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 159–214). Mahwah, NJ: Erlbaum.

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York, NY: Oxford University Press.

Kazdin, A. E. (2008). Evidence-based treatment and practice: New opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *American Psychologist, 63,* 146–159. http://dx.doi.org/10.1037/0003-066X.63.3.146

Kratochwill, T. R., & Levin, J. R. (1992). *Single-case research design and analysis: New directions for psychology and education*. Hillsdale, NJ: Erlbaum.

Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject researches: Percentage of data points exceeding the median. *Behavior Modification, 30,* 598–617. http://dx.doi.org/10.1177/0145445504272974

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: A Special Education Journal, 19,* 109–135.

Manolov, R., & Solanas, A. (2008). Comparing N = 1 effect size indices in presence of autocorrelation. *Behavior Modification, 32,* 860–875. http://dx.doi.org/10.1177/0145445508318866

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

Manolov, R., & Solanas, A. (2013). Assigning and combining probabilities in single-case studies: A second study. *Behavior Research Methods, 45,* 1024–1035. http://dx.doi.org/10.3758/s13428-013-0332-3

Matyas, T. A., & Greenwood, K. M. (1996). Serial dependency in single-case time series. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 215–243). Mahwah, NJ: Erlbaum.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834. http://dx.doi.org/10.1037/0022-006X.46.4.806

Messer, S. B. (2004). Evidence-based practice: Beyond empirically supported treatments. *Professional Psychology: Research and Practice, 35,* 580–588. http://dx.doi.org/10.1037/0735-7028.35.6.580

Morgan, D. L., & Morgan, R. K. (2001). Single-participant research design. Bringing science to managed care. *American Psychologist, 56,* 119–127. http://dx.doi.org/10.1037/0003-066X.56.2.119

Parker, R. I., Brossart, D. F., Vannest, K. J., Long, J. R., De-Alba, R. G., Baugh, F. G., & Sullivan, J. R. (2005). Effect sizes in single-case research: How large is large? *School Psychology Review, 34,* 116–132.

Parker, R. I., & Vannest, K. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy, 40,* 357–367. http://dx.doi.org/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,* 303–322. http://dx.doi.org/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,* 284–299. http://dx.doi.org/10.1016/j.beth.2010.08.006

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57,* 416–428. http://dx.doi.org/10.1037/h0042040

Salzberg, C. L., Strain, P. S., & Baer, D. M. (1987). Meta-analysis for single-subject research: When does it clarify, when does it obscure? *RASE: Remedial & Special Education, 8,* 43–48. http://dx.doi.org/10.1177/074193258700800209

Schlosser, R. W., Lee, D. L., & Wendt, O. (2008). Application of the percentage of nonoverlapping data (PND) in systematic reviews and meta-analyses: A systematic review of reporting characteristics. *Evidence-Based Communication Assessment and Intervention, 2,* 163–187. http://dx.doi.org/10.1080/17489530802505412

Scruggs, T. E., & Mastropieri, M. A. (1994a). The effectiveness of generalization training: A quantitative synthesis of single-subject research. In T. E. Scruggs & M. A. Mastropieri (Eds.), *Advances in learning and behavioral disabilities* (Vol. 8, pp. 259–280). Bingley, UK: Emerald.

Scruggs, T. E., & Mastropieri, M. A. (1994b). The utility of the PND statistic: A reply to Allison and Gorman. *Behaviour Research and Therapy, 32,* 879–883.

Scruggs, T. E., & Mastropieri, M. A. (1998). Summarizing single-subject research. Issues and applications. *Behavior Modification, 22,* 221–242. http://dx.doi.org/10.1177/01454455980223001

Scruggs, T. E., & Mastropieri, M. A. (2001). How to summarize single-participant research: Ideas and applications. *Exceptionality: A Special Education Journal, 9,* 227–244.

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *RASE: Remedial and Special Education, 8,* 24–33. http://dx.doi.org/10.1177/074193258700800206

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

Shadish, W. R. (2014b). Analysis and meta-analysis of single-case designs: An introduction. *Journal of School Psychology, 52,* 109–122. http://dx.doi.org/10.1016/j.jsp.2013.11.009

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

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods, 17,* 510–550. http://dx.doi.org/10.1037/a0029312

Tarlow, K. R., & Penland, A. (2016). *Percentage of Nonoverlapping Data (PND) Calculator*. Retrieved from http://www.ktarlow.com/stats/pnd

Thompson, B. (1999). If statistical significance tests are broken/misused, what practices should Suppl. or replace them? *Theory & Psychology, 9,* 165–181. http://dx.doi.org/10.1177/095935439992006

Wampold, B. E. (1988). Introduction [to special autocorrelation issue]. *Behavioral Assessment, 10,* 227–228.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604. http://dx.doi.org/10.1037/0003-066X.54.8.594

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,* 18–28. http://dx.doi.org/10.1177/0022466908328009

(*Appendices follow*)

Appendix A

Derivation of the Formula for $p_{PND}(n_A,n_B,k)$

If n is the total number of observations, the sample space of interest consists of all n^n possible sequences of length n with values in 1, 2, . . . , n, with the first n_A observations corresponding to the baseline phase and the remaining n_B observations corresponding to the treatment phase. We use the following notation and terminology:

- A – set of observations in the baseline phase
- B – set of observations in the treatment phase
- success—an observation in B which is strictly greater than the largest observation in A

The sequences with PND at least $\frac{k}{n_B}$ are exactly the sequences which have at least k successes. Counting these sequences can be accomplished by partitioning the sample space in a suitable manner. Let $\boldsymbol{E}_{i,M,r}$ represent the set of sequences which satisfy the following properties:

1. Each sequence in $\boldsymbol{E}_{i,M,r}$ has exactly i successes (i is an integer between 1 and n_B)
2. Each sequence in $\boldsymbol{E}_{i,M,r}$ takes the maximum value M on A (M is an integer between 1 and n)
3. Each sequence in $\boldsymbol{E}_{i,M,r}$ has exactly r observations in A which take the maximum value M (r is an integer between 1 and n_A).

It is obvious that we can calculate unambiguous values of i, M, and r for any element in the sample space. Now it is necessary to count the number of sequences in each $\boldsymbol{E}_{i,M,r}$. For fixed i, M, and r, an element of $\boldsymbol{E}_{i,M,r}$ may be constructed as follows:

Task description	Number of possibilities
Choose r slots in A to take value M	$C(n_A, r)$
Choose values ($\leq M - 1$) for remaining slots in A	$(M - 1)^{n_A - r}$
Choose i slots in B for successes	$C(n_B, i)$
Choose values ($> M$) for successes	$(n - M)^i$
Choose values ($\leq M$) for remaining slots in B	$M^{(n_B - i)}$

By the multiplication principle, the total number of elements of $\boldsymbol{E}_{i,M,r}$ is given by the product of the entries in the second column of the above table. So the probability that a randomly chosen element of the sample space lies in $\boldsymbol{E}_{i,M,r}$ is

$$\frac{C(n_A, r)(M - 1)^{n_A - r} C(n_B, i)(n - M)^i M^{(n_B - i)}}{n^n}$$

To obtain $P(PND(x) \geq k)$, simply sum over all possible values of i which are between k and n, and all possible values of M and r.

(*Appendices continue*)

Appendix B

R Code for Calculating *p* Values for PND

After installing the free R statistical computing software (R Core Team, 2014; available at http://www.r-project.org), the code below may be copied into the R console. PND *p* values are then calculated by executing the function `pnd(nA,nB,k)`, where n_A, n_B, and k are replaced, respectively, with the number of baseline phase observations, the number of treatment phase observations, and the number of treatment phase observations which exceed completely the maximum value in the baseline phase. For example, with the example AB time series {4, 2, 3, 5, 5 / 7, 6, 5, 2, 9}, the user would type `pnd(5,5,3)`.

```
C <- function(n, k) {
  return(factorial(n)/(factorial(n - k) * factorial(k)))
}

pnd <- function(nA, nB, k) {
  n <- nA + nB
  p <- 0
  i <- k
  while (i < (nB + 1)) {
    M <- 1
    while (M < (n + 1)) {
      r <- 1
      while (r < (nA + 1)) {
        p <- p + C(nA, r) * (M - 1)^(nA - r) * C(nB, i) * (n - M) ^ i * M ^ (nB
- i)
        r <- r + 1
      }
      M <- M + 1
    }
    i <- i + 1
  }
  p <- p/n^n
  PND <- k / nB * 100
  results <- data.frame(PND, p)
  return(results)
}
```

Received July 1, 2016
Revision received August 6, 2016
Accepted August 8, 2016 ■