Accounting for baseline trends in intervention studies: Methods, effect sizes, and software

Li-Ting Chen¹, Po-Ju Wu² and Chao-Ying Joanne Peng³

Abstract: Single-case experimental studies are relevant and important to substantiate the effectiveness of behavioral, clinical, and educational interventions. When a baseline trend is present in intervention studies, it is challenging for researchers to determine, visually or statistically, if an effect is due to the intervention or to the naturally occurring trend. In this paper, we demonstrated and appraised four methods that can assess an intervention effect in the presence of baseline trends. These four methods quantified an intervention effect as a phase change effect size (i.e., mean phase difference and the slope and level change) or a nonoverlap effect size (i.e., Tau-U and Tau_c). Empirical data from an intervention study were used in the demonstration. All methods were evaluated in terms of types of intervention effects assessed, assumptions, specialized computing tools, and application issues (e.g., missing scores). To empower researchers to use these four methods, we provide a summary of each method’s appropriate inferences about an intervention effect and its computing tools’ strengths and limitations. Based on the results, we recommend using multiple baseline trend-controlled methods to draw a conclusion about an intervention effect.

ABOUT THE AUTHOR
Our research team has been interested in investigating practical and theoretical issues surrounding the definition, application, and computation of effect sizes for between-subjects and within-subjects designs, including the single-case experimental designs. The issues we have investigated and contributed toward the effect size literature include the impact of APA and AERA guidelines on effect size reporting practices, alternative effect sizes beyond Cohen’s d, how to deal with baseline trends, the treatment of missing data, the treatment of ties, strengths and limitations of computing tools for the computation of effect sizes.

PUBLIC INTEREST STATEMENT
How can we know that an intervention worked, despite a natural improvement in a participant’s behavior? Clinicians, educators, service providers, and specialists often rely on a single-case design to determine if an intervention has effectively changed a specific behavior, such as disruptive behavior, in a participant. A basic single-case design consists of a baseline phase and an intervention phase. If a behavior shows an improvement trend during the baseline phase, such as a decrease in disruptive behavior or an increase in communication, it is difficult to conclude that such a behavior has improved due to the intervention alone, or due to a combination of the intervention and the behavioral trend. Using real-world data, this article demonstrates four methods that can untangle a baseline trend from an intervention effect. It is hoped that this article empowers practitioners and researchers to appropriately assess an intervention effect in the presence of baseline trends.
In intervention studies, the effectiveness of an intervention or treatment is often studied using single-case experimental designs (SCEDs). SCEDs have been used in subfields of psychology, such as, neuropsychology (Zermatten, Rochat, Manolov, & Van der Linden, 2018), health psychology (Sniehotta, Presseau, Hobbs, & Araújo-Soares, 2012), clinical psychology (Morgan & Morgan, 2001), and counseling psychology (Ellis, Hutman, & Chapin, 2015). A SCED study collects detailed information about a few participants’ target behavior over an extended period of time in order to determine the effect of an intervention. SCED studies typically consist of a baseline (or A) phase and an intervention (or B) phase in which the participants serve as their own controls. Consequently, data are correlated serially over time or sessions.

While visual analysis is the primary method of evaluating an intervention effect in SCEDs (Kazdin, 2011), statistical methods are increasingly employed to complement visual analysis (Manolov, Gast, Perdices, & Evans, 2014). For both visual and statistical analyses of SCED data, one challenge often encountered by practitioners and researchers was the presence of a baseline trend, because it may persist into the intervention phase (Sullivan, Shadish, & Steiner, 2015). If an undesirable behavior decreases (a downward trend), or a desirable behavior increases (an upward trend), during the baseline phase, it may be argued that the behavior has naturally improved without the intervention. According to Parker, Cryer, and Byrns (2006), 41% of the 165 data graphs, based on an AB design, showed an improvement trend in the baseline phase. When baseline trends were present, researchers could not reach a consensus based on visual analysis, or yield valid conclusions about the intervention effect (Lieberman, Yoder, Reichow, & Wolery, 2010; Mercer & Sterling, 2012).

Likewise, the statistical analyses (e.g., effect size, its statistical test, and interpretation) of SCED data are compounded by baseline trends (Campbell, 2004). It is therefore imperative that researchers be informed of ways of dealing with an existing baseline trend when assessing an intervention effect in SCEDs. A number of methods and computing tools have been proposed to fit a straight line as the baseline trend (Manolov, 2018), according to What Works Clearinghouse’s standards (Kratochwill et al., 2010). Yet these methods and computing tools have not been critically assessed in terms of their ability to deal with practical issues such as, how to properly interpret each method’s proposed effect size index when assumptions are violated, when missing data are present, or when scores are tied. Furthermore, the functionalities of specialized computing tools written for baseline trend-controlled methods are not widely known.

This study seeks to fill these voids by demonstrating and evaluating four baseline trend-controlled methods and their computing tools. Specifically, this study aims to

(I) use empirical data (Lambert, Cartledge, Heward, & Lo, 2006) to define and illustrate four methods that account for baseline trends in SCEDs;

(II) investigate each method’s proposed intervention effect size in terms of its proper interpretation, assumptions, and specialized computing tools; and

(III) resolve each method’s application issues, including unequal phase lengths, tied scores, missing data, out-of-range predicted scores, and unequally spaced sessions.

The empirical data is hereafter referred to as the Lambert data. The Lambert data have been used in methodological articles to illustrate statistical methods suitable for SCED studies (e.g., Chen, Peng, & Chen, 2015; Shadish, 2014).
1. The Lambert data

Lambert et al. (2006) implemented the use of response cards in two fourth-grade math classes to determine if the intervention minimized students’ disruptive behaviors while learning math. Their study employed a reversal design with two baseline phases (A1 and A2) and two intervention phases (B1 and B2). A disruptive behavior was recorded in 10 intervals of each study session. The dependent variable was the number of intervals in which at least one disruptive behavior was observed; hence, data ranged from 0 to 10 for each student in each study session.

To appraise the methods examined in this paper, we focused on data obtained from three students (S1, S2, and S3) during the second baseline (A2) and the second intervention (B2) phases (see Figure 1). These students’ data were well suited for evaluating the four methods for five reasons. First, the three students’ data exhibited different baseline trends during the A2 phase. According to Figure 1, Student S1 showed an upward trend, Student S2 showed a flat or no trend, and Student S3 showed a downward trend. The manner in which the four methods dealt with different trends, or lack thereof, was of interest to this study.

Second, the intervention was implemented in two math classes for different lengths. The different phase lengths in the Lambert data could present a challenge when student-level results are integrated into class-level results. Third, there were tied scores for all three students. These tied scores required adjustments when a trend-controlled method based on ranks of scores is applied. In this paper, we demonstrated how adjustments could be implemented. Fourth, several intervention sessions were missed; hence, missing scores occurred during those sessions. If and how each baseline trend-controlled method dealt with missing scores was of interest to our evaluation because missing scores frequently occurred in SCED studies (Chen, Feng, Wu, & Peng, 2019). Lastly, when a method used a trend line to account for the baseline trend, predicted scores derived from the trend line might be outside the legitimate score range of 0 to 10 for the Lambert data. It was, therefore, important to investigate if and how each method dealt with out-of-range predicted scores while controlling for a baseline trend.

2. Methods, effect sizes, and software accounting for baseline trends

To identify methods suitable for our examination, we used the search terms (“single-case” OR “single-subject”) AND (“trend”) in the Web of Science database, along with references located in published

---

Figure 1. The number of intervals with disruptive behaviors in A2 and B2 phases for Students S1, S2, and S3. S1’s and S2’s data were replicated from Lambert et al. (2006). S3’s data were suggested by one student in Lambert et al. (2006).
journal articles. We limited our search to methods proposed after 2009. Our search identified four methods that were suitable for our investigation and assessment. These four methods do not require a model for data or analysis and are applicable in simple as well as complex designs. Those methods that could not be fully implemented without specialized computing tools were excluded, such as, the graph rotation for overlap and trend method (Parker, Vannest, & Davis, 2014).

The four methods quantified an intervention effect as a phase change effect size (slope change and/or level change), or a nonoverlap effect size, after adjusting it for a baseline trend. Below we define each method in terms of a step-by-step algorithm, followed by a demonstration. A discussion of each method’s effect size and its proper interpretation, assumptions, and application issues is presented at the end of each effect size section. An alpha level of .05 was pre-selected as a criterion for statistical significance for any statistical test of a null hypothesis. Table 1 summarizes each method’s effect size index, assumptions, computing tools, and application issues.

2.1. Phase change effect sizes
Two methods computed the phase change as an index of an intervention effect: the mean phase difference (MPD) method and the slope and level change (SLC) method.

2.1.1. The MPD method
This method was proposed by Manolov and Solanas (2013) to quantify the phase difference between the intervention scores and predicted intervention scores, predicted from a baseline trend. Since 2013, multiple variations of MPD have been proposed. In this paper, we refer to the MPD initially proposed by Manolov and Solanas (2013) as MPD1, and subsequent variations as MPD2 to MPD8. MPD1 to MPD4 are conceptually and computationally simpler than MPD5 to MPD8 that were recently proposed by Manolov, Solanas, and Sierra (2018). MPD1 to MPD4 are presented in this section, whereas MPD5 to MPD8 are presented in supplemental materials available from https://osf.io/h75fd/?view_only=0c40cd0678ed45798501de418cca0f44

2.1.2. MPD1
The computation of MPD1 consists of four steps. Step 1: Form differenced baseline scores. Step 2: Compute the mean of the differenced baseline scores and treat it as the slope (b_A) of the baseline trend. Step 3: Use the slope (b_A) to compute the predicted intervention scores. Step 4: Obtain residuals by subtracting the predicted intervention scores from the observed intervention scores. The mean of the residuals from Step 4 is the MPD1.

The differenced baseline score of Step 1 is the difference between a baseline score and its preceding baseline score. Thus, the first differenced baseline score is the difference of the second baseline score minus the first baseline score, and the second differenced baseline score is the difference of the third baseline score minus the second baseline score. This definition of differenced scores is referred to as the first-order differencing (Manolov & Solanas, 2013). The first-order differencing assumes that a score is influenced by its immediately preceding score. The number of differenced baseline scores is one fewer than that of baseline scores, or (n_A-1).

In Step 2, the mean of the differenced baseline scores is the slope (b_A) of the baseline trend. b_A can be directly calculated from the difference between the last and the first baseline scores, divided by (n_A - 1), without considering baseline scores in between. The slope is used in Step 3 to derive predicted intervention score (\( \hat{y}_{n_A+j} \)) at the time point of (n_A + j) according to Equation (1):

\[
\hat{y}_{n_A+j} = y_1 + b_A \times (n_A + j - 1),
\]

where \( y_1 \) = the first baseline score, and \( j \) = the time point in the intervention phase = 1, 2, 3, ..., n_A. Therefore, the first predicted intervention score (\( \hat{y}_{n_A+1} \)) = \( y_1 + b_A \times (n_A) \), and the last predicted intervention score (\( \hat{y}_{n_A+n_B-1} \)) = \( y_1 + b_A \times (n_A + n_B-1) \).
In Step 4, MPD1 = the mean of residuals 
\[ \text{MPD1} = \frac{\sum_{j=1}^{n_B} (y_{n_A+j} - \hat{y}_{n_A+j})}{n_B} \] 
where \( y_{n_A+j} \) = the observed intervention score at the \((n_A+j)\)th time point, and \( \hat{y}_{n_A+j} \) = the predicted intervention score obtained from Equation (1) at the same time point. By definition, MPD1 is a measure of phase change due to level change and/or slope change (Manolov & Solanas, 2013).

An empirical benchmark of MPD1 has been established to facilitate its interpretation. Solomon, Howard, and Stein (2015) compiled a distribution of empirical MPD1s based on 131 studies of different sorts (school-wide positive behavior support, teacher performance feedback, math interventions, and classroom-based individual behavior interventions). Solomon et al. (2015) reported the first quartile of the empirical MPD1 distribution to be 1.22, the second quartile =1.90, and the third quartile = 2.80.

| Characteristics                  | Phase change effect size | Nonoverlap effect size |
|----------------------------------|--------------------------|------------------------|
|                                  | MPD | SLC | Tau-\(U_{AB-A}\) | \(\text{Tauc}\) |
| Interpretations                  |     |     |                  |                    |
| Descriptive                      | √   | √   | √                | √                  |
| Inferential                      |     |     |                  |                    |
| Assumptions                      |     |     |                  |                    |
| At least ordinal-level data      | √   |     | √                |                    |
| At least interval-level data     |     | √   | √                |                    |
| Equally spaced sessions          | √   | √   | √                |                    |
| Baseline trend is linear\(^a\)   | √   | √   | √                |                    |
| Baseline trend continues into intervention phase\(^b\) | √   | √   | √                | √                  |
| Baseline trend is determined by the first and last baseline scores | √   | √   |          |                    |
| First-order differenced baselined scores can remove baseline trend | √   | √   |          |                    |
| Application issues               |     |     |                  |                    |
| Unequal phase length allowed     | √   |     | √                | √                  |
| Tied scores are adjusted or not an issue | √   | √   |          |                    |
| Missing data handled by computing tools\(^c\) | √   | √   | √                |                    |
| Out-of-range predicted scores are not an issue or adjusted by computing tools\(^d\) | √   | √   |          |                    |
| Unequally-spaced sessions adjusted by computing tools\(^e\) | √   | √   |          |                    |

\(^a\)\(\text{Tau-}\(U_{AB-A}\)\) and \(\text{Tauc}\) can account for monotonic baseline trends.

\(^b\)Although all four methods assume that the baseline trend continues to intervention phase, the length of baseline and intervention phases should be approximately equal. Parker et al. (2014) stated, “If [the baseline trend] extended through a long Phase B, the reliability or credibility of this slope quickly approaches zero” (p. 88).

\(^c\)For MPD, missing data can be handled by MPD R function in supplemental materials. For SLC, missing data can be handled by SLC R function in supplemental materials. For \(\text{Tau-}\(U_{AB-A}\)\), \(\text{Tauu}\) R function available at http://ktarlow.com/stats/tauu.txt handles missing data by omitting the missing scores. Likewise, for \(\text{Tauc}\), \(\text{Tauc}\) R function in supplemental materials and Tarlow’s tool available at http://www.ktarlow.com/stats/tau handle missing data by omitting the missing scores.

\(^d\)Out-of-range predicted scores are not an issue for SLC and \(\text{Tau-}\(U_{AB-A}\)\). For MPD3 and MPD4, out-of-range predicted scores can be adjusted in MPD R function available from the supplemental materials and the web-based tool available at https://manolov.shinyapps.io/MPDExtrapolation/. For \(\text{Tauc}\), \(\text{Tauc}\) R function available from supplemental materials can adjust out-of-range predicted scores.

\(^e\)Unequally spaced measurements may be resolved by the MPD R and SLC R functions available from the supplemental materials.
2.1.3. MPD2
Manolov and Rochat (2015) modified the computation of MPD1 to use the median of the baseline scores as the constant in Equation (1), instead of the first baseline score. According to Manolov and Rochat (2015), the modified predicted intervention score \( \hat{y}_{nA+j} \) at the time point of \( (nA+j) \) is obtained from Equation (2):

\[
\hat{y}_{nA+j} = y_{A,\text{Mdn}} + b_{A} \times |nA+j - (nA+1)/2|,
\]  

(2)

where \( y_{A,\text{Mdn}} \) = the median of baseline scores, \( b_{A} \) from Equation (1), and \( (nA+1)/2 \) is the median session of the baseline phase.

Similar to MPD1, MPD2 = the mean of residuals \( = \sum_{j=1}^{nB} (y_{nA+j} - \hat{y}_{nA+j}) / nB \), where \( y_{nA+j} \) = the observed intervention score at the \( (nA+j) \)th time point, and \( \hat{y}_{nA+j} \) = the predicted intervention score defined in Equation (2) at the same time point.

2.1.4. MPD3 and MPD4
The predicted score from Step 3 of MPD1 (Manolov & Solanas, 2013) or MPD2 (Manolov & Rochat, 2015) may be outside its logical or legitimate range. When this occurs, Faith, Allison, and Gorman (1997) suggested to adjust the predicted out-of-range scores to the nearest maximum (=10 in the Lambert data), or minimum (=0 in the Lambert data). In this paper, we refer to MPD1 with Faith et al.’s (1997) adjustment as MPD3, and MPD2 with Faith et al.’s (1997) adjustment as MPD4. By definition, \(|\text{MPD3}| \leq |\text{MPD1}| \) and \(|\text{MPD4}| \leq |\text{MPD2}| \).

MPD1 to MPD4 can be computed using a free web-based tool available from https://manolov.shinyapps.io/MPDExtrapolation/, under the tab “Mean phase difference”. When missing scores occur, researchers can obtain MPD1 to MPD4 from the MPD R function included in supplemental materials.

2.1.5. Demonstration
For S1 and S2, MPD1 and MPD4 were obtained from the web-based tool mentioned above. For S3, we took the missing sessions (27 and 32) into consideration when computing MPD1 to MPD4 using the MPD R function.

According to Table 2, Student S1’s MPD1 (= -9.57), MPD2 (= -12.07), MPD3 (= -7.79) and MPD4 (= -8.00). For Student S2, both MPD1 and MPD2 were identical (= -9.43) and both MPD3 and MPD4 were identical (= -8.00) because S2’s baseline data displayed a flat or no trend in Figure 1. These MPDs were interpreted as follows: for S1 and S2, there was a decrease of at least seven intervals containing disruptive behaviors during the intervention phase from what was predicted from the baseline trend. Furthermore, the MPD1s for the two students were both greater, in absolute values, than the third quartile of 2.80 (Solomon et al., 2015). We, therefore, concluded that the response card intervention was effective in decreasing disruptive behaviors for Students S1 and S2, after accounting for an upward trend for S1, but no trend for S2.

Student S3’s MPD1 (= 3.57), MPD2 (= 3.07), MPD3 (= 0.50), and MPD4 (= 0.39) were all positive suggesting an increase in disruptive behaviors during the intervention phase, after accounting for a downward baseline trend. These results did not support an effective intervention for S3, beyond a decreasing trend already occurring during the baseline phase. Because there is no statistical test of a sample MPD1 to MPD4, we interpreted these MPDs descriptively.

2.1.6. The SLC method
The SLC method was proposed by Solanas, Manolov, and Onghena (2010) to quantify slope change and level change separately. The SLC method consists of three steps. Step 1: Remove the baseline trend from baseline and intervention scores. Step 2: Compute the slope change based on detrended intervention scores. Step 3: Compute the level change based on detrended and slope-change-controlled intervention scores and detrended baseline scores.
### Table 2. MPD1, MPD2, MPD3, and MPD4 computed for Students S1, S2, and S3 of the Lambert data

| Student | $b_A$ = baseline slope | MPD1<sup>a</sup> | MPD2<sup>b</sup> | MPD3<sup>c</sup> | MPD4<sup>d</sup> |
|--------|------------------------|------------------|------------------|------------------|------------------|
| S1     | $(8-3)/7 = 0.71$       | ![Graph](image)  | -9.57            | -12.07           | -7.79            |
| S2     | $(10-8)/7 = 0.29$      | ![Graph](image)  | -9.43            | -9.43            | -8.00            |
| S3     | $(2-7)/6 = -0.83$      | ![Graph](image)  | 3.57             | 3.07             | 0.50             |

Values under each data graph show the mean of residuals for its corresponding MPD. The data graphs were copied from the free web-based tool available at [https://manolov.shinyapps.io/MPDExtrapolation/](https://manolov.shinyapps.io/MPDExtrapolation/), except for S3. In addition to the data graphs, the returned results from the web-based tool show the values of baseline trend ($b_A$) and MPD in red font. When “Aim of the intervention” is selected based on the intended results (either increase or reduce), the green vertical lines show the residuals of supporting the effectiveness of the intervention effect and the red vertical lines show the residuals of not supporting the effectiveness of the intervention.

- MPD1 was computed under the tab “Mean phase difference” with (1) “Slope = differencing. Intercept definition =?” selected “2013” and (2) “To avoid out-of-bounds forecasts, specify the minimal possible value” entered based on the minimum possible value on the measured scores (such as 0 in Lambert data) to replace “NA”, and (4) “How to control for out-of-bounds extrapolations?” selected “flattened trend”.
- MPD2 was computed under the tab “Mean phase difference” with (1) “Slope = differencing. Intercept definition =?” selected “2015” and (2) “To avoid out-of-bounds forecasts, specify the minimal possible value” entered based on the maximum possible value on the measured scores (such as 10 in Lambert data) to replace “NA”, and (4) “How to control for out-of-bounds extrapolations?” selected “flattened trend”.
- MPD3 was obtained under the tab “Mean phase difference” with (1) “Slope = differencing. Intercept definition =?” selected “2013”, (2) “To avoid out-of-bounds forecasts, specify minimal possible value” entered based on the minimum possible value on the measured scores (such as 0 in Lambert data) to replace “NA”, (3) “To avoid out-of-bounds forecasts, specify maximal possible value” entered based on the maximum possible value on the measured scores (such as 10 in Lambert data) to replace “NA”, and (4) “How to control for out-of-bounds extrapolations?” selected “flattened trend”.
- MPD4 was obtained under the tab “Mean phase difference” with the same input as MPD3, except for selecting 2015 instead of 2013 under “Slope = differencing. Intercept definition =?”.
To remove the baseline trend in Step 1, Equation (3) is applied:

\[
D_t = y_t - (b_A \times t), \text{ for } t = 1, 2, \ldots, (n_A + n_B)
\]  

where \(D_t\) is the detrended score at the \(t\)th time point, \(y_t\) is the observed score at the same time point \(t\), and \(b_A\) is the slope in Equation (1).

Step 2's slope change (sc) is the difference between the last detrended intervention score and the first detrended intervention score, divided by \((n_B - 1)\). Thus,

\[
sc = \frac{D_{(n_A + n_B)} - D_{(n_A + 1)}}{(n_B - 1)}.
\]  

Step 3's level change (lc) is the difference between the mean of the detrended and slope-change-controlled intervention scores and the mean of the detrended baseline scores. The detrended and slope-change-controlled intervention scores, or \(D_{SCC}\), are given by Equation (5):

\[
D_{SCC} = D_t - sc \times (t - n_A - 1), \text{ for } t = (n_A + 1), (n_A + 2), \ldots, (n_A + n_B).
\]  

where \(D_t\) is the detrended intervention score at the \(t\)th time point in the intervention phase. Thus,

\[
lc = \text{mean}\{D_{SCC}, t = (n_A + 1), \ldots, (n_A + n_B)\} - \text{mean}\{D_t, t = 1, \ldots, n_A\}.
\]  

Both \(sc\) and \(lc\) can be computed using a free web-based tool at https://manolov.shinyapps.io/Change/, under the tab “Slope and level change”. When there are missing scores, researchers can obtain \(sc\) and \(lc\) from the SLC R function included in supplemental materials.

### 2.1.7. Demonstration

For Students S1 and S2, we computed \(sc\) and \(lc\) using the web-based tool listed above (See Table 3). For S3, we took the missing sessions (27 and 32) into consideration and used the SLC R function to compute \(sc\) and \(lc\). For Student S1, an \(sc\) of −1.09 was interpreted as an average decrease of 1.09 intervals with disruptive behaviors from a previous session at any given intervention session. An \(lc\) of −7.59 was interpreted as S1 exhibiting an average of 7.59 fewer intervals with disruptive behaviors in the intervention phase than in the baseline phase, after accounting for an upward baseline trend. For Student S2, an \(sc\) of −0.16 was interpreted as an average decrease of 0.16 intervals with disruptive behavior from a previous session at any given intervention session. An \(lc\) of −8.66 was interpreted as S2 exhibiting an average of 8.66 fewer intervals with disruptive behaviors in the intervention phase than in the baseline phase with no trend. For both S1 and S2, their \(lcs\) suggested a noticeable decrease from A2 to B2 phases. Their \(scs\) suggested a downward trend during the intervention phase. Both results supported an effective intervention for S1 and S2, after considering the baseline trend.

Student S3’s \(sc\) (= 0.73) was interpreted as an average increase of 0.73 intervals with disruptive behavior from a previous session at any given intervention session. Student S3’s \(lc\) (= −0.94) was interpreted as S3 exhibiting an average of 0.94 fewer intervals with disruptive behaviors in the intervention phase than in the baseline phase, after accounting for a downward baseline trend. These results did not support an effective intervention for S3, beyond a decreasing trend already occurring during the baseline phase.

### 2.2. Discussion of the MPD and the SLC methods

MPD1 to MPD4, \(sc\), and \(lc\) are applicable to interval-level or ratio-level scores. These six effect sizes can be computed for individuals, as well as for a class/group as an average, even if phase lengths are unequal. These six effect sizes can be interpreted directly and descriptively, if measurements are meaningful, as in the Lambert data. MPD1 can also be compared relatively to quartiles of empirical MPD1s compiled by Solomon et al. (2015). Because the sampling distributions of these six effect sizes are not available, researchers cannot interpret them inferentially.
Table 3. SLC computed for Students S1, S2, and S3 of the Lambert data

| Student | Original Data Points and Detrended Asterisks | First and Last Data Points and Detrended Asterisks | Detrended Data with No Phase B Slope |
|---------|---------------------------------------------|-----------------------------------------------|-------------------------------------|
| S1      | ![Graph S1](image1.png)                     | ![Graph S1](image2.png)                      | ![Graph S1](image3.png)            |
| S2      | ![Graph S2](image4.png)                     | ![Graph S2](image5.png)                      | ![Graph S2](image6.png)            |
| S3      | ![Graph S3](image7.png)                     | ![Graph S3](image8.png)                      | ![Graph S3](image9.png)            |

Note. The graphs were generated in a row for each student under the tab “Slope and level change” available at [https://manolov.shinyapps.io/Change/](https://manolov.shinyapps.io/Change/), except for S3. The first graph with the caption “Original data points and detrended asterisks” presents the observed scores as black dots and the detrended scores of Equation (3) as red asterisks. “Baseline trend =”, shown in the green font is the value of the baseline trend (or $b_A$). The second graph with the caption “First and last data points and detrended asterisks” presents a green line connecting the first and the last detrended baseline scores, and a red line connecting the first and the last detrended intervention scores. “Change in slope =”, shown in red font is the value of $sc$. The third graph with the caption “Detrended data with no phase B slope” presents the detrended baseline scores and the detrended and slope-change-controlled intervention scores in black dots. The two blue-dashed lines represent the mean of the detrended baseline scores and the mean of the detrended and slope-change-controlled intervention scores, respectively. “Net change in level =”, shown in blue font is the value of $lc$. 

Chen et al., Cogent Psychology (2019), 6: 1679941. https://doi.org/10.1080/23311908.2019.1679941
The calculation of MPD1 to MPD4, sc, and lc assumes that (1) the baseline trend is linear and persists throughout the entire intervention phase; (2) the baseline trend (i.e., $b_A$) can be determined by the first and the last baseline scores; (3) baseline and intervention sessions are equally spaced; and (4) the first-order-differenced baseline scores can be used to effectively remove the baseline trend (Manolov & Solanas, 2013).

The third assumption is derived from the fact that differenced scores, defining $b_A$ in Equations (1 and 3), should correspond to time points, not session numbers. Linking scores with real time points is also recommended in the visual analysis according to “The Single-Case Reporting Guideline In BEhavioural Interventions (SCRIBE) 2016: Explanation and Elaboration” (Tate et al., 2016, p. 23).

When baseline sessions are not equally spaced, $b_A$ should be adjusted for the unequal spaces between sessions. For example, assuming four baseline scores of (3, 5, 7, 9) were obtained from Day 1, 3, 4, and 5, respectively; the mean of the differenced baseline scores (namely, $b_A$) should be adjusted to $(9-3)/(5-1) = 1.5$, rather than $(9-3)/(4-1) = 2$. Likewise, predicted intervention scores of Equations (1 and 2) need to be adjusted accordingly by treating the intervention sessions, that would equalize the unequal-spaced sessions, as missed sessions in the MPD and the SLC R functions.

Tied scores are acceptable to the MPD and the SLC methods because they are not based on ranks of scores. These methods’ web-based tools cannot accept missing scores. Missing scores from Sessions 27 and 32 by S3 were entered as “NA” into the MPD and the SLC R function before they computed MPD1 to MPD4, or sc and lc, respectively. Missing scores may also be imputed by statistical methods in order to maintain the design structure of an intervention study (Chen et al., 2019; Peng & Chen, 2018, in press).

Out-of-range predicted scores of Equations (1 and 2) were adjusted for MPD3 and MPD4 by the MPD’s web-based tool and by the MPD R function to user-specified maximum or minimum, according to Faith et al. (1997). The SLC method does not have to deal with out-of-range predicted scores.

### 2.3. Nonoverlap effect sizes

Two methods computed the degree of nonoverlap as an index of the intervention effect: the Tau-U$_{AB-A}$ method and the baseline-corrected Tau (Tau$_u$) method. Nonoverlap is defined as the percentage of intervention scores that do not overlap with baseline scores. The higher the nonoverlap, the greater is the intervention effect.

#### 2.3.1. The Tau-U$_{AB-A}$ method

The Tau-U$_{AB-A}$ method adjusts the Tau-U$_{AB}$ index for the baseline trend. Both Tau-U$_{AB}$ and Tau-U$_{AB-A}$ are nonoverlap effect sizes proposed by Parker, Vannest, Davis, and Sauber (2011). Both are derived from Kendall’s Tau and Mann–Whitney U, hence, the name Tau-U. The Tau-U$_{AB-A}$ method consists of three steps. Step 1: Compute Tau-U$_{AB}$ between baseline scores and intervention scores. Step 2: Compute the baseline trend in terms of Kendall’s S$_A$. Step 3: Adjust Tau-U$_{AB}$ for the baseline trend to yield Tau-U$_{AB-A}$.

Tau-U$_{AB}$ in Step 1 is defined as a ratio of Kendall’s S divided by the product of the number of baseline scores ($n_A$) and the number of intervention scores ($n_B$):

$$\text{Tau-U}_A = \frac{S_{AB}}{n_A \times n_B}, \text{ where } S_{AB} = \text{Kendall’s S}$$

(7)

$S_{AB}$ in Equation (7) is calculated from pairwise comparisons of baseline scores with intervention scores using a matrix suggested by Parker et al. (2011). Figure 2 presents such a matrix based on S1 data. In Figure 2 matrix, baseline (A2) scores are listed as rows in an ascending order of sessions; the intervention (B2) and baseline scores (A2) are listed as columns in a reversed order of sessions. Cell entries are results of pairwise comparisons. When a column value is greater than a row value, a “+” is recorded in their intersection cell. Conversely, when a row value is greater than
When the column value and the row value are tied, a "T" is recorded. The shaded area in the matrix is used to compute $S_{AB}$ for Students S1, S2, and S3 to be

$$S_{AB} = \frac{S_{AB} - S_A}{n_A \times n_B + n_A \times (n_A - 1)/2}$$

(8)

The Tauu R function from http://ktarlow.com/stats/r/tauu.txt (Tarlow, 2017b, March) computes Tau-U_{AB} under "ab mina" and Tau-U_{AB} under "ab", according to Parker et al. (2011). The definition of Tau-U_{AB} permits its interpretation to be relative to Tau-U_{AB}, or to Tau-U_{AB}'s empirical distribution (Solomon et al., 2015). According to Solomon et al. (2015), the first quartile of Tau-U_{AB} = .28, the second quartile = .47, and the third quartile = .57.

2.3.2. Demonstration

Using the Tauu R function, we computed Tau-U_{AB} for Students S1, S2, S3 to be −.92, −1.00, and −.98, respectively (see Table 4). These Tau-U_{AB}S were interpreted as, 92%, 100%, and 98% of pairs of data showed improvement from the A2 phase to the B2 phase for Students S1, S2, and S3, respectively.

Using the Tauu R function, we computed Tau-U_{AB} for Students S1, S2, and S3 to be −.78, −.79, and −.56, respectively. Student S1's Tau-U_{AB} meant that 78% of pairs of data showed improvement after the upward baseline trend was removed from Tau-U_{AB}. It was a decrease in absolute value from its corresponding Tau-U_{AB} = −.92, but greater, in absolute value, than the third quartile of Solomon et al.'s (2015) distribution of Tau-U_{AB}. Student S2's Tau-U_{AB} meant that 79% of pairs of data showed improvement after the somewhat flat baseline trend was removed. It was a decrease in absolute value from its corresponding Tau-U_{AB} = −1.00, but greater, in absolute value, than the third quartile of Solomon et al.'s (2015) distribution. Student S3's Tau-U_{AB} meant that 56% of pairs of data showed improvement after the downward baseline trend was removed. It was a decrease in absolute value from its corresponding Tau-U_{AB} = −.98 and was similar, in absolute value, to the third quartile of Solomon et al.'s (2015) distribution. We, therefore, concluded that the response card intervention improved S1's, S2's, and S3's disruptive behaviors.
Table 4. Tau-U_{AB}, Tau-U_{A}, and Tau-U_{AB,A} computed for S1, S2, and S3 of the Lambert data

| Tau-U Equations | Student S1 | Student S2 | Student S3 |
|-----------------|------------|------------|------------|
| \(\text{Tau-U}_{AB}^{\text{ab}}\) | \(\frac{s_{AB}}{n_{AB}}\) | \(-\frac{46}{72} = -0.92\) | \(-\frac{52}{72} = -1.00\) | \(-\frac{52}{72} = -0.98\) |
| \(\text{Tau-U}_{A}^{\text{Trenda}}\) | \(\frac{s_{A}}{n_{A}}\) | \(\frac{23}{78} = 0.30(p = 0.149)\) | \(\frac{22}{78} = 0.28(p = 0.437)\) | \(\frac{22}{78} = 0.25(p = 0.437)\) |
| \(\text{Tau-U}_{AB,A}^{\text{ab.mina}}\) | \(\frac{s_{AB}}{n_{AB}} - \frac{s_{A}}{n_{A}}\) | \(-\frac{70}{99} = -0.71\) | \(-\frac{70}{99} = -0.79\) | \(-\frac{70}{99} = -0.56\) |

Kendall’s \(S_{AB}\), \(S_{A}\), \(n_{A}\), and \(n_{B}\)
\(S_{AB} = \#\text{positive-}\#\text{negative}\)
\(S_{A} = \#\text{positive-}\#\text{negative}\)
\(n_{A} = \#\text{baseline (A2) scores}\)
\(n_{B} = \#\text{intervention (B2) scores}\)

\(S_{AB} = 2 - 68 = -66, S_{A} = 17 - 5 = 12, n_{A} = 8, n_{B} = 9\)

Note: Tau-U_{AB}, Tau-U_{A}, and Tau-U_{AB,A} are obtained from the Tauu R function available at [http://ktarlow.com/stats/r/tauu.txt](http://ktarlow.com/stats/r/tauu.txt) (Tarlow, 2017b, March)

- \(^{\text{a}}\)Tau-U_{AB} is shown under “ab”.
- \(^{\text{b}}\)Tau-U_{A} is shown under “Trenda”.
- \(^{\text{c}}\)Tau-U_{AB,A} is shown under “ab.mina”.
2.3.3. The Tau\(_U\) method

The Tau\(_U\) method consists of four steps. Step 1: Quantify a baseline trend as a Kendall’s Tau (\(\tau\)) correlation coefficient (Kendall, 1962) by correlating baseline scores and their time points. Step 2: Test Kendall’s \(\tau\) statistically using Kendall’s S test (Kendall, 1962). Step 3: If Kendall’s S test is statistically significant according to a researcher’s pre-specified criterion, perform the Theil–Sen regression on the baseline data to remove the baseline trend from both the baseline and the intervention scores. Step 4: Compute a Kendall’s Tau correlation coefficient (Kendall, 1962) between residuals, obtained from the Theil–Sen regression line, and a dummy variable, coded as 0 for all baseline residuals and 1 for all intervention residuals. Because the residuals are uncorrelated with the baseline trend, Kendall’s Tau from Step 4 is called the baseline-corrected Tau in Tarlow (2017a), or Tau\(_u\) in this paper.

If the Step 2 test is not statistically significant, Tau\(_u\) is not computed and an uncorrected Tau (or Tau\(_noc\) in this paper) is computed instead. If Tau\(_noc\) is computed in Step 2, Steps 3 and 4 are not performed. Kendall’s \(\tau\) from Step 1, Tau\(_noc\) from Step 2, and Tau\(_u\) from Step 4 are all Kendall’s Tau correlation coefficients. They can be calculated using Tarlow’s tool from http://www.ktarlow.com/stats/tau.

Kendall’s \(\tau\) from Step 1 is identical to Tau-U_A (Parker et al., 2011) in Table 4, if baseline scores are not tied. If some of baseline scores are tied, as in Students S1, S2, and S3 data, Equation (9) corrects tied scores for Kendall’s \(\tau\):

\[
\text{Kendall’s } \tau = \frac{S_A}{\sqrt{\frac{1}{2} n_A(n_A - 1) - U} \sqrt{\frac{1}{2} n_A(n_A - 1)}} \tag{9}
\]

where \(S_A\) is identical to the numerator of Tau-U_A in Table 4, and \(U\) = the correction variable for ties in baseline scores. The correction variable \(U\) is determined from Equation (10):

\[
U = \frac{1}{2} \sum u(u - 1), \tag{10}
\]

where \(u\) = the number of tied baseline scores in each set, and \(\frac{1}{2} u(u-1)\) is the pairwise comparisons among tied scores in each set. For example, for Student S1, there were three 8s (the first set of tied baseline scores) and three 10s (the second set of tied baseline scores). Thus, \(U = \frac{1}{2} \sum u(u - 1) = \frac{1}{2} \times 3 \times (3 - 1) + \frac{1}{2} \times 3 \times (3 - 1) = 6\) for S1. The second square root of the denominator in Equation (9) contains no correction for ties in the baseline time points because there are none among time points.

In Step 2, the S test of Kendall’s \(\tau\) is performed as a normal approximate test when the number of baseline scores \((n_A)\) is 10 or more (Kendall, 1962). If the S test of Step 2 turns out to be statistically insignificant, according to a pre-specified \(\alpha\), Tau\(_noc\) is computed, instead of Tau\(_u\). Tau\(_noc\) is computed between the observed baseline and intervention scores and a dummy variable, coded as 0 for all baseline scores and 1 for all intervention scores. Because the residuals are uncorrelated with the baseline trend, Tau\(_noc\) in this paper is computed instead. If Tau\(_noc\) in Equation (9) is identical to the numerator of Tau-U_A in Equation (7), and \(U_{AB} = \) the correction variable for ties in baseline and intervention scores. \(U_{AB}\) is determined from Equation (12):

\[
U_{AB} = \frac{1}{2} \sum u_{AB}(u_{AB} - 1), \tag{12}
\]

where \(u_{AB}\) = the number of tied scores within and across baseline and intervention phases in each set. For Student S1, there were six sets of tied scores, including three 3s, three 8s, three 10s, two 4s, two 1s, and two 0s. Thus, \(U_{AB} = \frac{1}{2} \sum u_{AB}(u_{AB} - 1) = \frac{1}{2} \times 3 \times (3 - 1) \times 3 + \frac{1}{2} \times 2 \times (2 - 1) \times 3 = 12\) for S1. The second square root of the denominator in Equation (11) contains the correction for ties in the dummy variable, namely, \([\frac{1}{2} n_A(n_A - 1) - \frac{1}{2} n_B(n_B - 1)]\); because the dummy variable is tied on 0s and...
1s. Tau_{noc} is tested using a normal approximate test when the sample size of (n_A + n_B) is 10 or more (Kendall, 1962).

If the S test of Step 2 is statistically significant, the Tau_c method proceeds to Step 3. Step 3 determines the slope and the intercept of the Theil–Sen regression line. The Theil–Sen regression is a nonparametric robust regression method suitable for skewed and heteroscedastic data. The Theil–Sen slope (b_{TS}) is the median based on all pairs of baseline scores, according to Equation (13):

\[ b_{TS} = \text{median of} \left( \frac{y_k - y_{k'}^{k'}}{k - k'} \right), \quad \text{for} \ k = 1, 2, \ldots, n_A \]  

(13)

where \( y_k \) is the baseline score at the kth session and \( y_{k'}^{k'} \) is the k’th session score, where \( k \neq k' \). The Theil–Sen intercept (\( a_{TS} \)) is computed by Equation (14):

\[ a_{TS} = \text{median of} (y_k - b_{TS} \times k), \quad \text{for} \ k = 1, 2, \ldots, n_A \]  

(14)

The \( a_{TS} \) and \( b_{TS} \) obtained from Theil–Sen regression make the correlation coefficient between the baseline scores and their residuals equal to approximately zero. The \( a_{TS} \) and \( b_{TS} \) are subsequently used to yield residuals (\( e_k \)) for all baseline and intervention scores, as in Equation (15):

\[ e_k = y_k - \text{predicted} y_k(= a_{TS} + b_{TS} \times k), \]

where \( k = 1, 2, \ldots, n_A, n_A + 1, \ldots, (n_A + n_B) \).  

(15)

In Step 4, Tau_c is computed according to Equation (16) between residuals derived in Step 3 and a dummy variable, coded as 0 for all baseline residuals and 1 for all intervention residuals:

\[ \text{Tau}_c = \frac{S_c}{\sqrt{1/2 N(N - 1)} - U_c \sqrt{1/2 N(N - 1) - 1/2 n_A(n_A - 1) - 1/2 n_B(n_B - 1)}}. \]  

(16)

where \( S_c \) is computed from shaded cells of a data matrix, similar to the shaded cells in Figure 2, for which the residuals of Equation (15) are listed as rows and columns, and \( U_c \) = the correction variable for ties in the residuals. \( U_c \) is determined from Equation (17):

\[ U_c = \sum \frac{1}{2} u_c(u_c - 1) \]  

(17)

where \( u_c \) = the number of tied residuals in each set across baseline and intervention phases. Similar to Tau_{noc}, the second square root of the denominator in Equation (16) contains the correction for ties in the dummy variable. Tau_c is tested using a normal approximate test when the sample size of (n_A + n_B) is 10 or more (Kendall, 1962).

2.3.4. Demonstration

Table 5 presents results obtained from Tarlow’s tool. In Step 1, Kendall’s \( \tau \)s of the baseline trend for Students S1, S2, and S3 were: \( \tau = .48 \) (\( p = .158 \)), \( .28 \) (\( p = .437 \)), \( -.75 \) (\( p = .031 \)), respectively. The \( \tau \)s of Students S1 and S2 were not statistically significant at \( \alpha = .05 \). Tarlow’s tool indicated that there was no need to correct for these students’ baseline trends. We therefore computed Tau_{noc} for S1 and S2, respectively, as \( -.70 \) (\( p = .002 \)) and \( -.75 \) (\( p = .001 \)). These results indicated that there was a statistically significant decrease in S1’s and S2’s intervals with disruptive behaviors from baseline to intervention phases. We, therefore, concluded that the response card intervention was effective for S1 and S2.

For S3, Tarlow’s tool computed Tau_c = .56 (\( p = .017 \)) because Kendall’s \( \tau \) of \( -.75 \) was statistically significant. Although Tau_c was statistically significant, its sign was positive indicating an increase in S3’s intervals with disruptive behaviors from baseline to intervention phases. Such a result contradicted S3’s data displayed in Figure 1. According to S3’s residual graph shown in the upper panel of Table 5, residuals obtained for intervention sessions (from 24 to 32) were increasingly larger than 0. This phenomenon is likely to occur when data are bounded (such as the Lambert data bounded...
by 0 and 10) (Tarlow, 2017b). It was, therefore, difficult to determine if the response card intervention was effective for S3, based on Tarlow’s $\tau_u = .56$.

2.4. Discussion of the Tau-$U_{AB-A}$ and the Tau$_c$ methods

Tau-$U_{AB-A}$ and Tau$_c$ are applicable to scores measured at least at the ordinal level. Both can be computed for individuals, as well as for a class/group as an average, even if phase lengths are unequal. Tau-$U_{AB-A}$ can be interpreted relatively by comparison to Tau-$U_{AB}$, or to quartiles of empirical Tau-$U_{AB,A}$ compiled by Solomon et al. (2015). Because there is no sampling distribution for a sample Tau-$U_{AB-A}$, it is inappropriate to interpret Tau-$U_{AB-A}$ inferentially (Peng & Chen, 2018, April). Tau$_c$ can be interpreted descriptively as well as inferentially, as demonstrated above.

In contrast to MPD and SLC, Tau-$U_{AB-A}$ and Tau$_c$ do not assume that the baseline trend is linear; they assume that the baseline trend persists throughout the intervention phase. The computation of Tau-$U_{AB}$, Tau-$U_{AB-A}$, Tau$_c$ or Tau$_{noc}$ is not affected by unequally spaced sessions because all are based on ranks of scores, not time points. The calculation of Tau-$U_{AB}$ or Tau-$U_{AB-A}$ is not adjusted for tied scores. Therefore, when tied scores occur, neither Tau-$U_{AB}$ nor Tau-$U_{AB-A}$ is adjusted for tied scores. Thus, when tied scores occur, neither Tau-$U_{AB}$ nor Tau-$U_{AB-A}$ can reach its theoretical maximum (Peng & Chen, 2018, April). The computation of Tau$_c$ is adjusted for tied scores, as shown in Equations 16 and 17.

The Tau-$U_{AB-A}$ method and the Tau$_R$ function, as well as the Tau$_c$ method and Tarlow’s tool, treated missed sessions, such as Sessions 27 and 32 in S3, as nonexistent. Consequently, in Table 5 for S3, the residual obtained for Session 28 is plotted for Session 27 and the residual obtained for Session 34 is plotted for Session 32.

Because Tau-$U_{AB-A}$ does not compute predicted scores, out-of-range predicted scores are not an issue for this method. It is an issue with the Tau$_c$ method. Out-of-range predicted scores in Equation (15) were not adjusted by Tarlow’s tool. They can be adjusted to user-specified maximum or minimum, according to Faith et al. (1997) in the Tau$_c$ tool, available from the supplemental materials. After adjusting the out-of-range predicted scores for S3, Tau$_c$ = .07 ($p = .818$) is statistically insignificant and different from Tarlow’s Tau$_c$ ($= .56, p = .017$). Based on the adjusted Tau$_c$ = .07, we concluded that the response card intervention was ineffective for S3.

When deciding whether to report Tau-$U_{AB-A}$ or Tau-$U_{AB}$, researchers are faced with different recommendations in the literature. Parker et al. (2011) recommended to remove the baseline trend, therefore reporting Tau-$U_{AB-A}$ if Tau-$U_A$ (Table 4) was at least .40. In Parker et al.’s (2011) analysis, Tau-$U_A = .40$ corresponded to the 75th percentile of 382 empirical Tau-$U_{AS}$ derived from published articles. We agree with Brossart, Laird, and Armstrong (2018) recommendation to report Tau-$U_{AB-A}$ when there are theoretical and empirical reasons for it. According to Brossart et al. (2018), a possible theoretical reason is that the researcher suspects a Hawthorne effect. A possible empirical reason is that the baseline trend is tested to be statistically significant, although the power of such a test is low when there is a small number of baseline scores. Similarly, the choice of Tau$_c$ or Tau$_{noc}$ depends on the significance test of Kendall’s $\tau$ in Step 1. Tarlow (2017a) recommended at least 10 baseline scores in order to achieve a reasonably large statistical power for testing Kendall’s $\tau$.

3. Conclusion

We defined and illustrated four methods that can account for baseline trends while assessing an intervention effect. We demonstrated these methods using three students’ data extracted from Lambert et al. (2006) and specialized computing tools. Three students’ data displayed an upward, flat, or downward trend in the baseline phase. All methods successfully incorporated three types of baseline trends into the assessment of the response card intervention for the students. Results revealed that the response card intervention was effective for Students S1 and S2. Such a conclusion was supported by all four methods investigated in this paper. Three methods (MPD, SLC, and Tau$_c$) concluded that the intervention was ineffective for S3. Because S3 showed a downward trend (or decrease in intervals with disruptive behaviors) during the baseline phase, a further decrease
### Table 5. Tau computed for Students S1, S2, and S3 of the Lambert data

| Student | Test for Baseline Trend\(^a\) | Tau | Residual Graph |
|---------|--------------------------------|-----|----------------|
| S1      | Baseline Trend:  
\quad • Tau (τ) = .48, p = .158  
Recommended Effect Size:  
\quad • Tau (No Baseline Correction): do not reject null hypothesis of no baseline trend | Tau (No Baseline Correction or \(\text{Tau}_{\text{noc}}\))  
\quad Effect Size:  
\quad • \(\text{Tau}_{\text{noc}} = -.70, p = .002\) | (Because the baseline trend was not significant, there was no need to look at the residual graph.) |
| S2      | Baseline Trend:  
\quad • Tau (τ) = .28, p = .437  
Recommended Effect Size:  
\quad • Tau (No Baseline Correction): do not reject null hypothesis of no baseline trend | Tau (No Baseline Correction or \(\text{Tau}_{\text{noc}}\))  
\quad Effect Size:  
\quad • \(\text{Tau}_{\text{noc}} = -.75, p = .001\) | (Because the baseline trend was not significant, there was no need to look at the residual graph.) |
| S3      | Baseline Trend:  
\quad • Tau (τ) = -.75, p = .031  
Recommended Effect Size:  
\quad • Baseline Corrected Tau; reject null hypothesis of no baseline trend | Baseline Corrected Tau (\(\text{Tau}_c\))  
\quad • \(a_{15} = 9, b_{15} = -1\)  
\quad \(\text{Tau}_c = .56, p = .017\) | Residual graph based on the residuals generated from Tarlow’s tool (PHASE A—CORRECTED and PHASE B—CORRECTED) |
Table 5. (Continued)

| Student | Test for Baseline Trend\(^a\) | Tau | Residual Graph |
|---------|-------------------------------|-----|----------------|
|         | Baseline Corrected Tau (\(\tau_c\)) adjusting for out-of-range predicted scores |     |                |
|         | \(\alpha_{TS} = 9, \beta_{TS} = -1\) | \(\tau_c = .07, p = .818\) |                |

The results are generated from Tarlow’s tool available at [http://www.ktarlow.com/stats/tau](http://www.ktarlow.com/stats/tau), except for the lower panel of S3. Edited results are shown in **bold**. To perform the significance test of Step 2, users need to first enter data into Tarlow’s tool, then click the button “Test for Baseline Trend”. Tarlow’s tool returns Kendall’s \(\tau\) in Equation (9) under “Baseline Trend” and presents the recommendation. According to the recommended effect size, users can click either the “Baseline Corrected Tau” button to compute \(\tau_c\), or the “Tau (No Baseline Correction)” button to compute \(\tau_{noc}\). By clicking on the button “See Additional Output”, users can obtain \(\alpha_{TS}\) under “\(a\) (int) =” and \(\beta_{TS}\) under “\(b\) =”. In addition, the residuals from the baseline phase can be obtained under “PHASE A- CORRECTED” and the residuals from the intervention phase can be obtained under “PHASE B- CORRECTED.” The upper residual graph for S3 was plotted in Excel using residuals generated from Tarlow’s tool. The lower residual graph for S3 was plotted in Excel, after adjusting for out-of-range scores.

\(^a\)After users click the button “Test for Baseline trend”, Tarlow’s tool returns either “Tau (No Baseline Correction): do not reject the null hypothesis of stable baseline” or “Baseline Corrected Tau; reject the null hypothesis of stable baseline”. We changed the wording to “Tau (No Baseline Correction): do not reject the null hypothesis of no baseline trend” or “Baseline Corrected Tau; reject the null hypothesis of no baseline trend.”
shown in the intervention phase was not sufficient, according to the three methods, to claim an effective intervention. Yet Tau-U_{AB-A} suggested that the intervention was effective for S3. In light of S3’s data display and proper interpretations of the results obtained from the four methods, we concluded that the intervention was not effective for S3. If the intervention were implemented over a longer phase for S3, the data and conclusion might be different. These findings highlight the importance of employing multiple baseline trend-controlled methods and of implementing an intervention sufficiently in order to fully assess an intervention effect for each participant. We agree with Kratochwill et al. (2013) that multiple statistical analyses should be used to assess an intervention effect in SCED studies.

Our study also investigated the proper interpretation of each method’s proposed effect size, assumptions, and ways in which several application issues can be resolved. Our findings are summarized in Table 1. The MPD and SLC methods require that data be at least interval-level measurements and the baseline trend be linear. The calculation of the baseline trend in MPD and SLC also assumes that the baseline trend can be determined solely by the first and last baseline scores. When applying these two methods, researchers need to attend to the equally spaced measurement issue. Unequally spaced measurements may be resolved by the MPD R and SLC R functions available from the supplemental materials. Although the web-based computing tools for MPD and SLC do not accept missing values, data graphs are generated to facilitate the interpretation of the results for datasets without missing scores.

The Tau-U_{AB-A} and Tau_{c} methods are both meaningful, even when there are outliers, because they are based on ranks of scores. The issue of out-of-range predicted scores is likely to be encountered with MPD1, MPD2, and Tau_{c}. In this paper, we provided a solution to adjust the out-of-range predicted scores to their nearest maximum or minimum. Manolov (2018) proposed another strategy to permit the baseline trend to level off at a certain point during the intervention phase (see supplemental materials).

Each of the four methods reviewed and evaluated in this paper has its unique merits. Of the four methods, Tau_{c} was the only index that allows for an inferential interpretation of the intervention effect. The other three methods yield effect sizes that should be interpreted descriptively, but not inferentially. It should be noted that the causal interpretation of Tau_{c} or other effect size measures, requires a sound study design. Researchers are encouraged to follow the SCED design standards presented in What Works Clearinghouse (Institute of Education Sciences, 2017) when designing an intervention study. Furthermore, according to What Works Clearinghouse, six data features should be examined to determine the effectiveness of an intervention, including level change, trend, variability, immediacy of the effect, overlap, and consistency of data in similar phases. To fully examine the six features of a participant’s data, researchers can consult Chen et al. (2015) and Manolov (2018). Given the importance of SCED studies in establishing and confirming evidence-based intervention practices, we hope that this paper has empowered practitioners and researchers to appropriately employ methods to disentangle baseline trends from intervention effects.

Acknowledgment
We thank Kathryn Allen, Brian Luckey, and Cong Zou for their comments on a previous draft.

Funding
The authors received no direct funding for this research.

Author details
Li-Ting Chen¹
E-mail: litingc@unr.edu
ORCID ID: http://orcid.org/0000-0001-5978-0548
Po-Ju Wu²
E-mail: pojwu@umail.iu.edu
ORCID ID: http://orcid.org/0000-0002-3608-5259
Chao-Ying Joanne Peng³
E-mail: peng@indiana.edu
ORCID ID: http://orcid.org/0000-0002-7282-5461
¹ Counseling and Educational Psychology, University of Nevada, Reno, NV, USA.
² Department of Statistics, Indiana University Bloomington, IN, USA.
³ Department of Psychology, National Taiwan University, Taipei, Taiwan.

Citation information
Cite this article as: Accounting for baseline trends in intervention studies: Methods, effect sizes, and software, Li-Ting Chen, Po-Ju Wu & Chao-Ying Joanne Peng. Cogent Psychology (2019), 6: 1679941.
References
Brossart, D. F., Laird, V. C., & Armstrong, T. W. (2018). Interpreting Kendall’s Tau and Tau-U for single-case experimental designs. *Cogent Psychology, 5*(1), 1–26. doi:10.1080/23311908.2018.1518607

Campbell, J. M. (2004). Statistical comparison of four effect sizes for single-subject studies. *Behavior Modification, 28*(2), 234–246. doi:10.1177/0145445503235296

Chen, L.-T., Peng, Y., Wu, P.-J., & Peng, C.-Y. J. (2019). Dealing with missing data by EM in single-case studies. *Behavior Research Methods. Advance online publication. doi:10.3758/s13428-018-01120-8.*

Chen, L.-T., Peng, C.-Y. J., & Chen, M.-E. (2015). Computing tools for implementing standards for single-case designs. *Behavior Modification, 39*(6), 835–869. doi:10.1177/0145445515603706

Ellis, M. V., Hutman, H., & Chapin, J. (2013). Reducing supervise anxiety: Effects of a role induction intervention for clinical supervision. *Journal of Counseling Psychology, 62*(4), 608–620. doi:10.1037/cou0000099

Faith, M. S., Allison, D. B., & Gorman, B. S. (1997). Meta-analysis of single-case research. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 119–138). New Jersey, NJ: Lawrence Erlbaum Associates.

Institute of Education Sciences. (2017). What works clearinghouse: Standards handbook (Version 4.0). Retrieved from https://ies.ed.gov/ncee/wcc/Handbooks

Kazdin, A. E. (2011). Single-case research designs: Methods for clinical and applied settings (2nd ed.). New York, NY: Oxford University Press.

Kendall, M. (1962). Rank correlation methods. New York, NY: Hafner Publishing Company.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). Single case designs technical documentation. In *What works clearinghouse: Procedures and standards handbook (Version 1.0).* Retrieved from https://ies.ed.gov/ncee/wcc/Document/229

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education, 34,* 26–38. doi:10.1177/0741932512452794

Lambert, M. C., Cartledge, G., Heward, W. L., & Lo, Y.-Y. (2006). Effects of response cards on disruptive behavior and academic responding during math lessons by fourth-grade urban students. *Journal of Positive Behavior Interventions, 8*(2), 88–99. doi:10.1177/1098300706080020701

Lieberman, R. G., Yoder, P. J., Reichow, B., & Wolery, M. (2010). Visual analysis of multiple baseline across participants graphs when change is delayed. *School Psychology Quarterly, 25,* 28–44. doi:10.1037/a0018600

Manolov, R. (2018). Linear trend in single-case visual and quantitative analyses. *Behavior Modification, 42*(5), 684–706. doi:10.1080/01454455.2017.13726301

Manolov, R., Gast, D. L., Perdices, M., & Evans, J. J. (2014). Single-case experimental designs: Reflections on conduct and analysis. *Neuropsychological Rehabilitation: An International Journal, 24,* 634–660. doi:10.1080/09602011.2014.903199

Manolov, R., & Rochat, L. (2015). Further developments in summarising and meta-analysing single-case data: An illustration with neurobehavioural interventions in acquired brain injury. *Neuropsychological Rehabilitation, 25,* 637–662. doi:10.1080/09602011.2015.1064452

Manolov, R., & Solanas, A. (2013). A comparison of mean phase difference and generalized least squares for analyzing single-case data. *Journal of School Psychology, 51*(2), 201–215. doi:10.1016/j.jsp.2012.12.005

Manolov, R., Solanas, A., & Sierra, V. (2018). Extrapolating baseline trend in single-case data: Problems and tentative solutions. *Behavior Research Methods. Advance online publication. doi:10.3758/s13428-018-1165-x.*

Mercer, S. H., & Sterling, H. E. (2012). The impact of baseline trend control on visual analysis of single-case data. *Journal of School Psychology, 50*(3), 403–410. doi:10.1016/j.jsp.2011.11.004.

Morgan, D. L., & Morgan, R. K. (2001). Single-participant research design: Bringing science to managed care. *American Psychologist, 56*(2), 119–127. doi:10.1037/0003-066X.56.2.1119

Parker, R. I., Cryer, J., & Byrns, G. (2006). Controlling baseline trend in single-case research. *School Psychology Quarterly, 21*(4), 418–444. doi:10.1037/h0084131

Parker, R. I., Vannest, K. J., & Davis, J. L. (2014). A simple method to control positive baseline trend within data nonoverlap. *The Journal of Special Education, 48*(2), 79–91. doi:10.1177/0022466912456430

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2017). Linear trend in single-case research. *Behavior Research Methods, 49*(3), 819–822. doi:10.3758/s13428-016-0804-4

Peng, C.-Y. J., & Chen, L.-T. (2018). Can Tau-U combine nonoverlap with trend? Paper presented at the American Educational Research Association Annual Meeting, New York, NY.

Peng, C.-Y. J., & Chen, L.-T. (2018). Handling missing data in single-case studies. *Journal of Modern Applied Statistical Methods, 17*(1), eP2488. doi:10.22237/jmasm/1525133280

Peng, C.-Y. J., & Chen, L.-T. (in press). Can missing data methods strengthen the validity of single-case studies? *Journal of Modern Applied Statistical Methods, 17*(2).

Shadish, R. W. (Eds.) (2014). Special issue: Analysis and meta-analysis of single-case designs. *Journal of School Psychology, 52*(2).

Snihottaf, F., Pesseau, J., Hobbs, N., & Arojuo-Soares, V. (2012). Testing self-regulation interventions to increase walking using factorial randomized N-of-1 trails. *Health Psychology, 31*(6), 733–737. doi:10.1037/a0027337

Solanas, A., Manolov, R., & Onghena, P. (2010). Estimating slope differences in multiple designs as a function of N = 1 designs. *Behavior Modification, 34*(3), 195–218. doi:10.1177/0145445510363306

Solomon, B. G., Howard, T. K., & Stein, B. L. (2015). Critical assumptions and distribution features pertaining to contemporary single-case effect sizes. *Journal of Behavioral Education, 24,* 438–458. doi:10.1007/s10864-015-9221-4

Sullivan, K. J., Shadish, W. R., & Steiner, P. M. (2015). An introduction to modeling longitudinal data with generalized additive models: Applications to single-case designs. *Psychological Methods, 20*(1), 26–42. doi:10.1037/met0000020

Tate, R. L., Perdices, M., Rosenkotter, U., McDonald, S., Togher, L., Shadish, W., ... Vohra, S. (2016). The Single-Case Reporting Guideline In BEhavioural Interventions (SCRIBE) 2016: Explanation and
elaboration. Archives of Scientific Psychology, 4(1), 10–31. doi:10.1037/arc0000027
Zermatten, V., Rochat, L., Manolov, R., & Van der Linden, M. (2018). Can an external device create and
trigger intention in a patient with a server brain injury? Neuropsychological Rehabilitation: An
International Journal, 28(7), 1211–1228. 10.1080/09602011.2016.1255230