Evidence-Based Communication Assessment and Intervention

Evidence-Based Communication Assessment and Intervention

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tebc20

Selecting the proper Tau-U measure for single-case experimental designs: Development and application of a decision flowchart

Joelle Fingerhut, Xinyun Xu & Mariola Moeyaert

To cite this article: Joelle Fingerhut, Xinyun Xu & Mariola Moeyaert (2021) Selecting the proper Tau-U measure for single-case experimental designs: Development and application of a decision flowchart, Evidence-Based Communication Assessment and Intervention, 15:3, 99-114, DOI: 10.1080/17489539.2021.1937851

To link to this article: https://doi.org/10.1080/17489539.2021.1937851

Evidence-Based Communication Assessment and Intervention, 2021
Vol. 15, No. 3, 99–114, https://doi.org/10.1080/17489539.2021.1937851

Routledge
Taylor & Francis Group

EBP Advancement Corner

Selecting the proper Tau-U measure for single-case experimental designs: Development and application of a decision flowchart

Joelle Fingerhut, Xinyun Xu & Mariola Moeyaert
Department of Educational and Counseling Psychology, University at Albany-SUNY, Albany, NY, USA

Abstract
A variety of measures have been developed to quantify intervention effects for single-case experimental design studies. Within the family of non-overlap indices, the Tau-U measure is one of the most popular indices. There are several Tau-U variants, each one calculated differently. The appropriateness of each Tau-U variant depends upon the data characteristics present within the study (e.g. number of measurement occasions, the within-case variability, and baseline trend). However, inconsistent terminology is used to refer to the Tau-U variants, and researchers can overlook the attributes of the different Tau-U variants. As a result, the Tau-U variants can be applied inappropriately, and this can result in invalid conclusions of intervention effectiveness. This paper proposes a Tau-U flowchart that can assist the decision-making process when using Tau-U with single-case experimental designs that incorporate baseline-intervention (AB) comparisons (e.g. multiple-baseline designs, withdrawal/reversal designs, etc.). The flowchart can help researchers select the appropriate Tau-U variant to use based on their data and research questions. The flowchart is applied to two single-case experimental studies to demonstrate its use.

Keywords: *Decision-making tool; Effect size; Meta-analysis; Single-case experimental designs; Tau-U; Tau-U decision flowchart.*

The single-case experimental design (SCED) is one frequently utilized methodology in the fields of communication disorders and special education (Cakiroglu, 2012; Ganz & Ayres, 2018; Schlosser, 2009). This type of research design is unique in that it allows for repeated measurement of the outcome variable while each subject serves as its own control. The dependent variable is measured at multiple points over time across different conditions, both when the independent variable is not present and when the independent variable is present (What Works Clearinghouse, 2020). These different conditions are also known as baseline (i.e. phase A) and intervention phase(s) (i.e. phase B). Thus, one baseline and intervention phase together are termed an AB phase contrast or AB design. What Works Clearinghouse (2020) guidelines recommend for studies to include at least three attempts to demonstrate the effect of the intervention across three points in time. As a result, several different types of SCEDs have been developed that are based upon the basic AB design logic; examples of these design tactics include the multiple-baseline design staggered across participants and withdrawal (i.e. reversal) design (What Works Clearinghouse, 2020).

For correspondence: Joelle Fingerhut jfingerhut@albany.edu Department of Educational and Counseling Psychology, University at Albany-SUNY, Albany, NY, USA; Xinyun Xu xxu5@albany.edu

SINGLE-CASE EXPERIMENTAL DESIGN MEASURES

A variety of different analysis techniques have been developed to evaluate intervention effects using SCED data. Visual analysis

has historically been used to determine intervention effectiveness in SCED research (Horner et al., 2012). However, within the past few decades there has been an increased interest in using quantitative analyses (Fingerhut et al., 2021). One possible reason for this is due to criticisms of visual inspection. Reliance on visual analysis alone may lead to Type II errors, as visual analysis is less sensitive to small data changes (McClain et al., 2014). Other research has found visual analysis to lack interrater agreement (Brossart et al., 2006; Ninci et al., 2015; Wolfe et al., 2016). Another reason for the increase in popularity of quantitative analyses is that it better allows for results to be aggregated; for example, the average estimated intervention effect across participants (and even studies) can be calculated. This aggregation of results permits more generalized conclusions about intervention effectiveness; the quantification of the effect aids in summarizing findings and allows for easier dissemination of findings (Shadish, 2014). Furthermore, legislation such as the Every Student Succeeds Act (2015) calls for the use of evidence-based practices in the classroom, and consequently there has been a greater need for researchers to quantitatively represent, summarize, and interpret results from SCEDs (Solomon et al., 2015). As a result, there has been an increased need for quantitative procedures to evaluate SCED data.

A variety of measures have been developed for use with SCEDs (e.g. percent of non-overlapping data, Scruggs et al., 1987; non-overlap of all pairs, Parker & Vannest, 2009; between-case standardized mean difference, Pustejovsky et al., 2014; percent of goal obtained; Ferron et al., 2020). One group of measures are the non-overlap indices, which reflect the amount of overlapping and/or non-overlapping data between the baseline and intervention phase. These measures are useful when the difference in mean between baseline and intervention phase would not be a meaningful representation of the data (Parker et al., 2011). Non-overlap indices are the most commonly used method for quantifying SCED estimates (Fingerhut, Moeyaert et al., 2020; Shadish et al., 2014). They are easily calculated, and can even be calculated by hand, unlike other quantification methods.

Each newly developed non-overlap index was developed with the intention to improve upon the previously developed non-overlap indices. One of the original non-overlapping measures is the percent of non-overlapping data (PND; Scruggs et al., 1987). This measure is criticized for being overly influenced by outliers, as the calculation and estimate depends on one data point (Lenz, 2013). As a result, percent of data points exceeding the median (PEM; Ma, 2006), the percent of all non-overlapping data (PAND; Parker et al., 2007), and improvement rate difference (IRD; Parker et al., 2009) were developed to improve upon PND. Although PEM, PAND, and IRD include more data points in the calculation than does PND, these measures cannot control for baseline trend. Baseline trend is a common characteristic in SCED data; Brossart et al. (2018) analyzed 115 published SCED data sets and found the average data set to have small to moderate trend in both the baseline and intervention phase (the average monotonic baseline trend was 0.32, and the average monotonic trend in the intervention phase was 0.43). Furthermore, ignoring baseline trend can have a negative impact on the calculation. Consider the hypothetical baseline and intervention phases with observation data points in the baseline as 1, 2, 3, 4, 5 and observation data points in the intervention as 6, 7, 8, 9, 10. In this example, which is graphically displayed in Figure 1, there is a trend, but no true intervention effect; because there is no

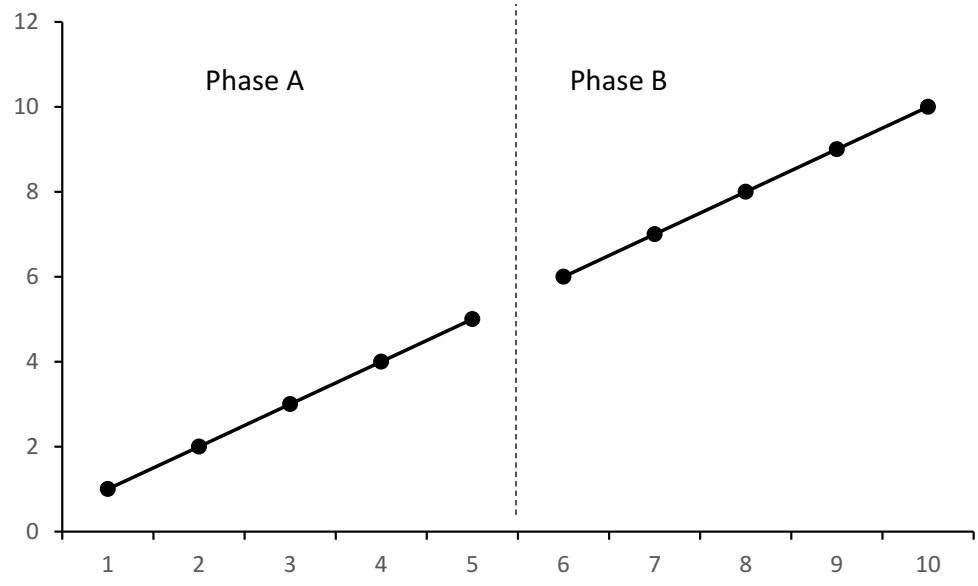

Figure 1. Example AB Graph with Trend.

Note. Graph was recreated using Microsoft Excel®.

overlap between observations in the baseline and intervention phase, a non-overlap measure would incorrectly indicate an intervention effect. For better consideration of trend within SCED data, newer measures such as percent of data points exceeding the median trend (PEM-T; White & Haring, 1980; Wolery et al., 2010), and the Tau-U variants termed Tau-U $_{\text{Trend A}}$ and Tau-U $_{\text{adj}}$ (two indices that are discussed in the next section) were developed to provide trend control.

Tau-U $_{A\ vs\ B}$ index and variants (Tau-U $_{Trend\ A}$ and Tau-U $_{adj}$)

Non-overlap of all pairs (NAP; Parker & Vannest, 2009) is another non-overlap index that was developed for use with SCEDs. Although it does not provide trend control, NAP makes comparisons between all data points in the calculation, unlike previously discussed non-overlap indices. As a result, the consequential estimate is based upon more information than SCED measures that do not make comparisons between all data points. The measure Tau-U $_{\text{A vs B}}$ (Parker et al., 2011) was developed to improve upon NAP; Tau-U $_{\text{A vs B}}$ makes NAP estimates more easily interpretable by linearly rescaling the estimate to the range −1.00 and 1.00. Tau-U $_{\text{A vs B}}$ (Parker et al., 2011) is one of the most popularly used indices (Fingerhut, Moeyaert et al., 2020; Tarlow, 2017), and has several variants. One variant of Tau-U $_{\text{A vs B}}$ is a baseline corrected version (Tau-U $_{\text{Trend A}}$; Parker et al., 2011), which was created to help account for baseline trend that can be present in data sets. While PEM-T is one non-overlap index that can control for baseline trend, the baseline trend control for this measure is unreliable (Brossart et al., 2014), demonstrating a need for an improved version. Tau-U $_{\text{Trend A}}$ is estimated by calculating the monotonic trend within the baseline phase, Trend A,

and then subtracting it from the original Tau-U $_{A\ vs\ B}$ calculation. Tau-U $_{Trend\ A}$ has several limitations, including that it can yield estimates that are not bound between −1.00 and 1.00, which can become an issue for researchers upon interpreting results (Brossart et al., 2018; Tarlow, 2017). The degree of adjustment for this measure is impacted by the length of the baseline and intervention phases, leading to further issues when interpreting results (Tarlow, 2017). As a result of these limitations, a second baseline-corrected Tau-U was developed (Tau-U $_{adj}$; Tarlow, 2017). This measure removes the baseline trend by using Theil-Sen robust regression. The Theil-Sen adjustment is a slope estimate, which is the median of the calculated slopes of the time variable and the baseline outcome scores in a dataset. This measure is bound between −1.00 and 1.00 and can account for baseline trend. However, it fails to control trend when there are too few measurement occasions (e.g. n= 5) in the baseline phase (Tarlow, 2017). Tarlow (2017) provides a power table to help users determine the statistical power of Tau-U $_{adj}$ depending on the degree of trend and number of measurement occasions.

Each of the Tau-U indices may be appropriate for use depending on different research questions and data characteristics. One should not conclude that Tau-U $_{adj}$ is the better option because it was most recently developed. For example, Brossart et al. (2018) demonstrate that baseline-trend controlling measures can bias Tau-U $_{adj}$ estimates when there is no true baseline trend. Therefore, it is essential to think carefully about the research questions and data characteristics before selecting a measure for use.

COMPARING AND CONTRASTING NON-OVERLAP INDICES

There have been a multitude of methodological studies conducted comparing the statistical qualities of different measures, including Tau-U and other non-overlap indices. Giannakakos and Lanovaz (2019) conducted a Monte Carlo simulation study to test the validity of PEM, IRD, and Tau-U in detecting changes in AB designs. Findings showed that Tau-U and IRD can successfully control for Type I errors, and also have sufficient power, unlike PEM. Pustejovsky (2019) conducted a simulation study to determine the procedural sensitivities of different SCED measures. In this study, results showed PND, PAND, and IRD to be sensitive to the number of baseline and intervention sessions, unlike PEM and NAP. However, all aforementioned measures were sensitive to the observations session length. Fingerhut et al. (2021) conducted a Monte Carlo simulation study to determine the effect of data characteristics such as within-case variability and trend on the three Tau-U variants. Findings demonstrated Tau-U $_{adj}$ to be less affected by trend, and all Tau-U variants to be affected by within-case variability relative to the size of the intervention effect. Jaksic et al. (2018) used real data sets to compare PND, PEM, and regression-based hierarchical linear modeling (Van Den Noortgate & Onghena, 2003a, 2003b) in their ability to estimate intervention effects. PND appeared to underestimate treatment response, while PEM was found to be valid and correctly able to indicate a treatment response in the majority of the cases. Rakap (2015) applied PND, IRD, PEM-T, and Tau-U $_{Trend\ A}$ to different data sets. Results indicated Tau-U $_{Trend\ A}$ to have better discriminability and sensitivity than the other measures. Brossart et al. (2018) compared how different Tau-U variants performed when applied to 115 different SCED data sets. Results showed the importance of applying each Tau-U variant carefully (i.e. only applying trend control variants when trend is present). Tarlow (2017) applied Tau-U $_{adj}$, Tau-U $_{Trend\ A}$, extended celeration line (White & Haring, 1980), regression

(Allison & Gorman, 1993), and mean phase difference (Manolov & Solanas, 2013) to 65 different data sets, with results highlighting the benefits of Tau-U $_{adj}$ and mean phase difference over other measures. These are just a few examples of the methodological studies that have been done to compare SCED measures with each other.

DECISION-MAKING PROCESSES FOR SINGLE-CASE EXPERIMENTAL DESIGN RESEARCHERS

Very few decision-making tools have been created based upon the findings from past methodological studies. Almost none of the decision-making tools that have been developed focus solely on non-overlap indices or Tau-U variants, specifically. Manolov and Moeyaert (2017) developed a chart including the criteria to be considered when choosing a SCED measure. Fingerhut, Marbou et al. (2020) created a user friendly Microsoft Excel® tool, based upon the chart developed by Manolov and Moeyaert (2017), to help users pick an appropriate measure depending on their research questions and data characteristics. Manolov et al. (2021) proposed a flowchart to help with a priori decisions regarding appropriate single-case measures and justification of the measures. Vannest et al. (2018) proposed criteria to help with decision-making and reporting of measures and visual analysis, acknowledging that there is no one best measure and that it is important to understand the data characteristics when picking a measure to use.

Tarlow (2017) developed a flowchart to help users appropriately use Tau-U $_{adj}$, but this flowchart does not include other Tau-U variants. Researchers may benefit from a decision-making tool that focuses solely on the three Tau-U variants. Tau-U is a preferred measure due to its advantages over other non-overlap indices, such as accounting for all data points, its ability to account for baseline trend, and it is less affected by ceiling effects (Tarlow, 2017). Tau-U is also a widely used measure; Tarlow (2017) found Tau-U to be cited in over 200 published SCED articles, while Fingerhut, Moeyaert et al. (2020) found Tau-U to be the most commonly used measure in SCED research published during the year 2019.[1] Therefore, researchers may benefit from a tool that helps them choose which Tau-U measure to use. Brossart et al. (2018) acknowledges that past reporting methods of Tau-U are confusing; researchers sometimes do not specify which Tau-U variant they are using, and consequently, a flowchart that explicitly distinguishes between the different variants may encourage appropriate reporting. Fingerhut et al. (2021) evaluated the effects of different data characteristics (i.e. number of measurement occasions, within-case variability, trend, and size of intervention effect) on Tau-U variant estimates. Based on these results, a flowchart can be developed to guide SCED researchers.

RESEARCH AIMS

The aim of this paper is to develop a Tau-U decision flowchart and demonstrate the applicability of the flowchart using two published SCEDs, using the findings from Fingerhut et al. (2021), which can assist researchers in selecting an appropriate Tau-U measure for their SCED data. Such a flowchart can be applied to the basic AB design, and any SCED design that incorporates AB design elements (e.g. multiple-baseline design, withdrawal design, and their variants). This flowchart considers typically encountered data characteristics and includes steps and guidelines for choosing appropriate Tau-U indices.

[1]This review was limited to the database PsycINFO.

FLOWCHART FOR TAU-U INDEX SELECTION

First, researchers are recommended to evaluate their research question(s) and decide if Tau-U can provide the desired quantification. As Tau-U is a non-overlap measure, it cannot provide an estimate to answer research questions concerning change in level, change in slope, and/or the magnitude of the effect across cases and/or studies. Therefore, if the research question includes any of these aspects, researchers are encouraged to use a different analytic approach (e.g. hierarchical linear modeling; Moeyaert et al., 2014; Van Den Noortgate & Onghena, 2003a, 2003b).

Results from Fingerhut et al. (2021) indicated that all Tau-U indices are significantly affected by the size of the intervention effect; large intervention effects cause the Tau-U across case estimates to be more toward 1.00 (which is desired). When there is no effect, Tau-U $_{A\ vs\ B}$, Tau-U $_{Trend\ A}$, and Tau-U $_{adj}$ yield a wide range of estimates instead of approximating zero as theoretically expected. Due to the potential issue in determining if there is an effect or not, it would be beneficial if several researchers who are naïve to the study first conduct a visual analysis of the graphs (Vannest et al., 2018), focusing on the within-case variance, intervention effect, and trend within each AB phase contrast. For example, the researcher is encouraged to visually analyze these data characteristics within the baseline and intervention phase of each participant in a multiple-baseline design across participants, or each AB phase contrast within each participant in a withdrawal design. Researchers are encouraged to use structured visual analysis techniques such as stability envelopes (Lane & Gast, 2014) to help evaluate the amount of variability relative to the magnitude of the intervention effect. If after completing visual analysis these reviewers believe the graph reflects a small effect (according to the research field-specific expectations), the Tau-U estimate (when using the Critical Tau-U[2] to determine if there is an effect or not) may not be able to detect the true effect, and researchers should use Tau-U indices with caution (e.g. either use a different analysis technique and/or report multiple effect size measures). This is because the Critical Tau-U method is conservative in order to control for Type I error (around 5%; see Fingerhut et al., 2021). The Tau-U estimate needs to be relatively large (larger than the Critical Tau-U) to make an inference that there is an intervention effect. The intervention effect needs to be relatively large to maintain the power (i.e. ability to detect an effect when there is a true effect) while controlling for Type I error. However, when the intervention effect is small, the power of Tau-U estimates using the Critical Tau-U is limited. This means that researchers are unlikely to obtain a large enough Tau-U estimate (larger than the Critical Tau-U) to conclude that there is an intervention effect. This issue of low power can be seen in the results displayed in Appendix A of Fingerhut et al. (2021) in this issue. Appendix A shows that the power of Tau-U estimation is relatively small in order to control for Type I error.

Next, the researcher should determine if there is a baseline trend within each A phase (i.e. baseline phase); researchers can use online websites (e.g. http://ktarlow.com/stats/tau [Tarlow, 2016]; http://www.singlecaseresearch.org [Vannest et al., 2016]) to determine if there is a significant trend. Results from Fingerhut et al. (2021) demonstrate that Tau-U $_{adj}$ is less affected by trend than is Tau-U $_{Trend\ A}$. Therefore, if there is a trend, researchers are advised to use Tau-

[2]See Fingerhut et al. (2021) for details about the Critical Tau-U method.

U $_{adj}$ rather than Tau-U $_{Trend\ A}$, because Tau-U $_{Trend\ A}$ is more affected by trend and can result in estimates over 1.00 (Fingerhut et al., 2021). However, researchers may prefer to use Tau-U $_{Trend\ A}$ if there are less than five measurement occasions in the baseline phase, as Tau-U $_{adj}$ fails to correct for baseline trend in this circumstance (Tarlow, 2017).

After determining the appropriate Tau-U estimate to use for each AB phase contrast,[3] researchers can proceed to calculate the index. When interpreting results, researchers are not recommended to use the commonly cited benchmarks by Vannest and Ninci (2015), as the benchmarks are commonly interpreted out of context of the original study. For further guidance on proper benchmarking procedures see Vannest and Sallese (2021). Also, the different Tau-U indices produce different ranges of estimates (e.g. Tau-U $_{adj}$ has a ceiling effect around 0.80, while Tau-U $_{Trend\ A}$ can exceed 1.00). Researchers are not encouraged to use uniform guidelines to interpret the estimates of different Tau-U indices. As an alternative, the Critical Tau-U can be used to determine if there is evidence in support of an intervention effect or not. It is advised to rule in favor of an intervention effect only when the Tau-U estimate is more extreme than the Critical Tau-U value. This value controls the Type I error rate to about 5%, which means that there is only about 5% probability that researchers will make a false positive conclusion (i.e. researchers obtain a statistically significant estimate under the null hypothesis that there is no effect). Therefore, the researcher can confidently conclude there is a statistically significant effect when the Tau-U estimate is larger than the Critical Tau-U value. Researchers are also encouraged to consider participant characteristics and data characteristics when interpreting results, as advised by current research (e.g. Vannest et al., 2018).

Another approach is for researchers to pre-estimate the anticipated size of the effect, the degree of baseline trend, and within-case variability prior to conducting the study. Then, before conducting the study, the researchers can use these hypothesized parameter values to follow the flowchart accordingly and to determine the recommended analytical technique for use. Researchers are encouraged to report the analytical technique recommended for use based on the anticipated data characteristics. If after conducting the study the data characteristics are different than anticipated, researchers can make note of these differences and use the flowchart (using the actual data characteristics) to determine the appropriate analytical technique. This procedure is recommended by Manolov et al. (2021) and can help prevent experimenter bias. The resulting flowchart can be seen in Figure 2.

EMPIRICAL DEMONSTRATION

Two peer-reviewed studies in special education were selected to illustrate how the Tau-U flowchart can be used to analyze SCED data. Two multiple baseline-design studies were selected due to their varying data characteristics (number of measurement occasions, number of cases, amount of within-case variability, and degree of baseline trend), allowing for the flowchart to be applied in different manners. Data from graphs were extracted using the data retrieval software program WebPlotDigitizer (Rohatgi, 2016). The flowchart in Figure 2

[3]It is possible that different Tau-U variants would be most appropriate for use for different AB phase contrasts. For example, if one AB block has trend in the A (baseline) phase, but the other AB block does not have trend in the A phase, the researcher would need to use two different Tau-U variants for the estimates of each separate AB phase contrast.

is followed to guide the appropriate use of Tau-U.

Figure 3(a) is from a published study evaluating the effectiveness of an intervention to improve participants' implementation of discrete trials (McKinney & Vasquez, 2014). There are three participants. Participants 1, 2 and 3 have 6, 10, and 12 measurement occasions in the baseline phase, respectively, and 12, 14, and 9 measurement occasions in the intervention phase, respectively.

Following the guidelines in the flowchart, visual aids are used to initially evaluate the trend, intervention effect, and within-case variability. An envelope projected around the split middle trend is used to determine the possible intervention effect and direction of trend (Lane & Gast, 2014). This visual aid tool is applied using the R code

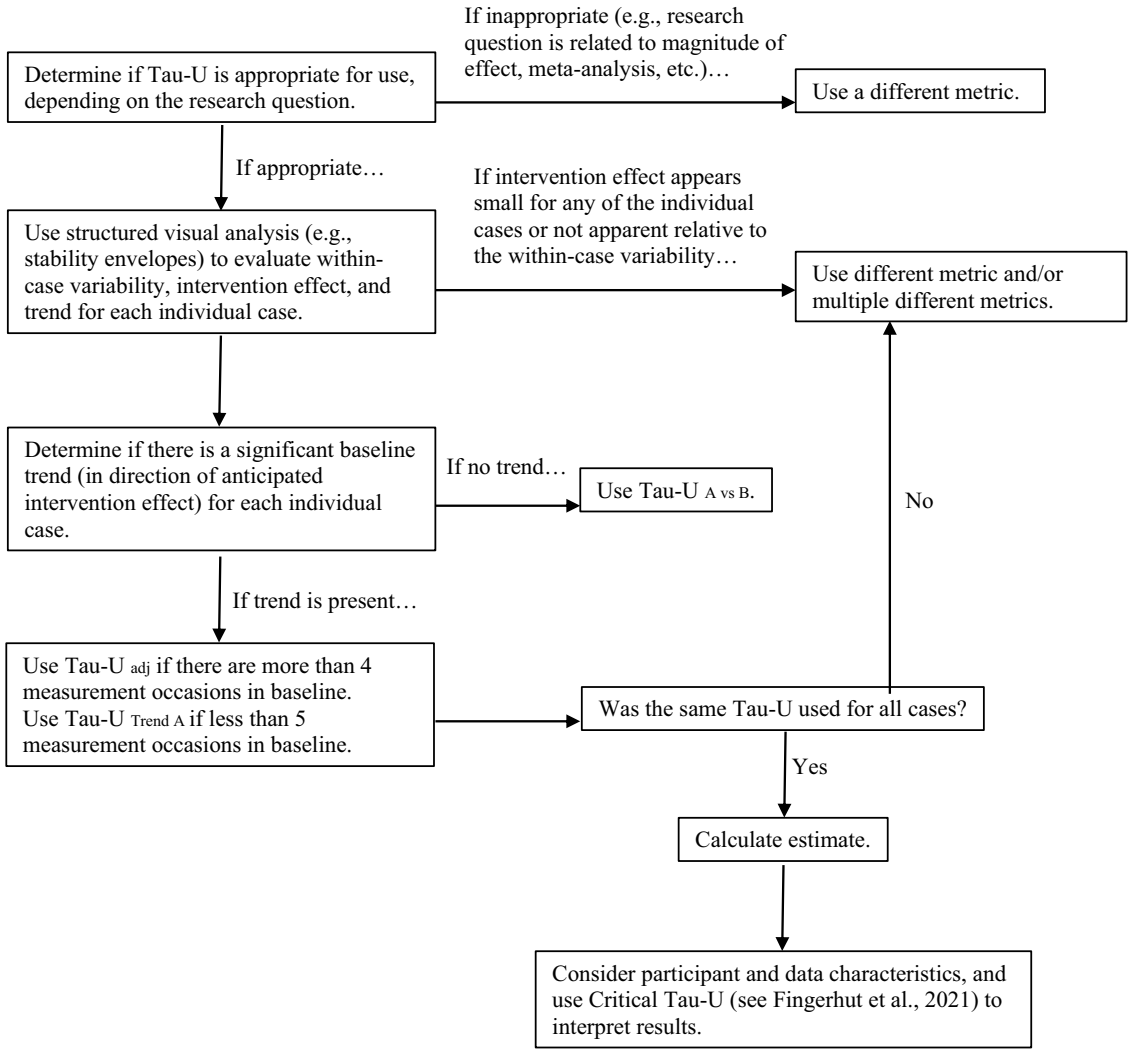

Figure 2. Tau-U Decision Flow Chart.

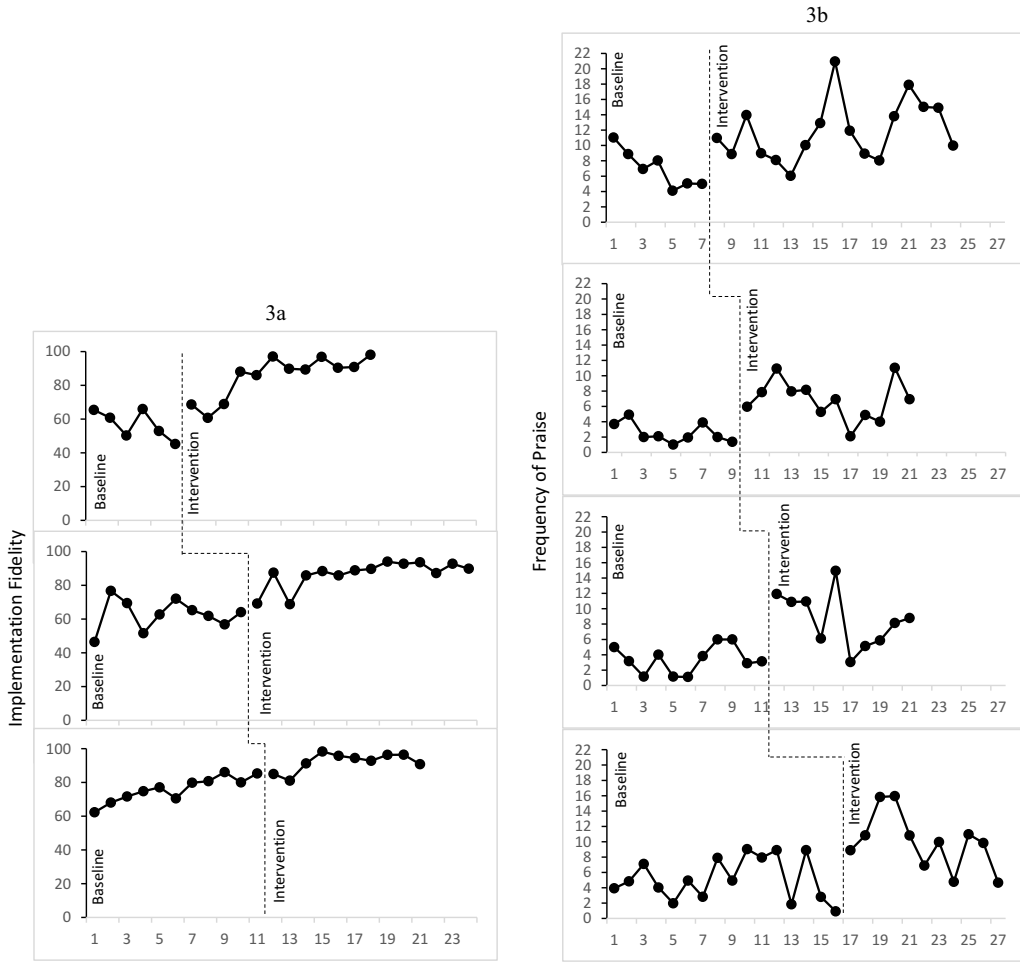

Figure 3. Empirical Demonstration Data: Two Multiple-Baseline Design Studies.

Note. Raw data to create the graphs in Figure 3a were retrieved from the original study of McKinney and Vasquez (2014) using WebPlotDigitizer (Rohatgi, 2016). Raw data to create the graphs in Figure 3b were retrieved from the original study of Gage et al. (2018) using WebPlotDigitizer (Rohatgi, 2016). Graphs were recreated using Microsoft Excel®.

by Manolov (2014). If the majority of the data points in the intervention phase are above the projected envelope, this indicates a possible effective intervention. The results from applying the visual aid indicate that there is possibly no effect for Participant 3 and possibly a small effect for Participants 1 and 2. Participants 2 and 3 possibly have a small trend and Participant 1 has a trend in the opposite direction of the intervention effect. The within-case variability is evaluated with stability envelopes (Wolery et al., 2010). The envelope is structured above and below 25% of the median in the baseline and intervention phase for each participant, and the number of data points that fall outside of this envelope is considered. This visual aid indicates that within-case variability appears to be small for Participant 1 and medium for Participant 2

and 3. As no participants appear to have large within-case variability relative to a small effect, Tau-U measures are acceptable to use.

Next, the Baseline Corrected Tau Calculator (http://ktarlow.com/stats/tau; Tarlow, 2016) is used to determine if a control for baseline trend is needed. The calculator indicates the need to control for trend for Participant 3, as the baseline trend is 0.81. This participant has 11 data points in the baseline phase, and the power table proposed by Tarlow (2017) indicates that there is enough power to correctly identify this baseline trend. The other two participants do not have a statistically significant trend. Participant 1 has only six data points in the baseline phase, thus demonstrating that there may be a lack of power to accurately detect the amount of trend (according to the power table proposed by Tarlow, 2017). However, visual analysis of the graph for Participant 1 demonstrates a potentially small trend in the opposite direction of the intervention effect, and the Baseline Corrected Tau Calculator (http://ktarlow.com/stats/tau; Tarlow, 2016) confirms this. Therefore, it is reasonable to assume that it is not necessary to control for baseline trend. Regarding Participant 2, there are 10 data points in the baseline phase. According to the power table proposed by Tarlow (2017), there is significant power to detect a baseline trend greater than 0.40, which is also the degree of trend Parker et al. (2011) use to indicate the need for trend adjustment. Using the Baseline Corrected Tau Calculator (http://ktarlow.com/stats/tau; Tarlow, 2016), the baseline trend is reported to be −0.07. Therefore, it is reasonable to assume that a trend adjustment is not needed for this participant. After the trend has been determined in this manner, the appropriate Tau-U index can be chosen. Tau-U $_{adj}$ is the appropriate Tau-U measure to use for Participant 3, and Tau-U $_{A\ vs\ B}$ is the appropriate measure for Participants 1 and 2. Different Tau-U indices are advised for use to calculate the individual estimates for the participants. Therefore, as the flowchart indicates, an overall estimate cannot be calculated due to the between-case variability (i.e. differences in trend between participants). A different measure should be used to calculate an overall effect, such as hierarchical linear modeling (e.g. Van Den Noortgate & Onghena, 2003a, 2003b) or between-case standardized mean difference (Pustejovsky et al., 2014).

Figure 3(b) is from a published study examining the effectiveness of a professional development intervention to increase teachers' use of behavior specific praise (Gage et al., 2018). Four participants participated in this multiple-baseline design study. Participants 1, 2, 3, and 4 have 7, 9, 11, and 16 measurement occasions in the baseline phase, respectively, and 17, 12, 10, and 11 measurement occasions in the intervention phase, respectively. Unlike the data in Figure 3(a), the data shown in Figure 3(b) are count/binomial data rather than continuous data. Therefore, it is not reasonable to assume that errors are independently and normally distributed, and that there are equal variances (assumptions that are made within this paper). However, for consistency and for the purposes of this empirical demonstration, it is assumed that the data in Figure 3(b) are continuous. Readers are advised to approach the analysis of binomial data differently [see Declercq et al. (2019) for more information regarding binomial data analysis].

The same visual aids used for Figure 3(a) are applied again. An envelope projected around the split middle trend is used (Lane & Gast, 2014). The results from evaluating the graphs with the envelope projected around the split middle trend indicates that there is possibly a large effect

for Participants 1, 2 and 3, and possibly a medium effect for Participant 4. There appears to be no trend in the direction of the intervention effect for any of the participants. The within-case variability is evaluated with the same stability envelopes used for Figure 3(a) (Wolery et al., 2010). This visual aid indicates that within-case variability appears to be large for all participants. As the intervention effects appear to be medium to large, no participants have large within-case variability relative to a small effect, and so Tau-U measures are acceptable for use.

Following the flowchart in Figure 2, the baseline trend is then estimated with Baseline Corrected Tau Calculator (http://ktarlow.com/stats/tau/; Tarlow, 2016). The power table proposed by Tarlow (2017) is used in a similar manner as described for Figure 3(a). It can be concluded that no participants have a significant trend. This indicates that Tau-U $_{\text{A vs B}}$ can be used for all participants, and the overall Tau-U $_{\text{A vs B}}$ across case estimate can be calculated. The Tau-U $_{\text{A vs B}}$ across case estimate (i.e. the unweighted average) is 0.78.

Next, the Critical Tau-U $_{\text{A vs B}}$ (see Appendix A of Fingerhut et al., 2021) indicates that there is an effect. The Critical Tau-U $_{\text{A vs B}}$ is determined by examining the most similar condition (20 measurement occasions, 4 cases, no trend, and large within-case variance; $\sigma_e^2 = 2.00$). The Critical Tau-U $_{\text{A vs B}}$ is 0.54. The obtained Tau-U $_{\text{A vs B}}$ across case estimate is 0.78, which is larger than the Critical Tau-U $_{\text{A vs B}}$, and so it can be concluded that there is an intervention effect across cases (with about 5% risk for Type I error).

DISCUSSION

A flowchart was developed based upon results from Fingerhut et al. (2021) and related research (e.g. Tarlow, 2017a; Vannest et al., 2018) to assist researchers in selecting appropriate Tau-U indices. The flowchart helps researchers determine which Tau-U indices are appropriate for use based upon their data characteristics. The Tau-U metric has been one of the most widely used SCED measures (Fingerhut et al., 2021; Tarlow, 2017), perhaps due to the fact that it is able to control for trend, considers all data points in the calculation, and there are easy to use calculators available (e.g. http://ktarlow.com/stats/tau, Tarlow, 2016; http://www.singlecaseresearch.org, Vannest et al., 2016). Although highly utilized for research, there are no consistent terms for the different Tau-U variants (Brossart et al., 2018; Tarlow, 2017). No previous studies have systematically provided clear guidance for selecting an appropriate Tau-U measure. Using consistent terminology is essential, as it is the foundation to mutual understanding of the different Tau-U variants. Without clear distinction and guidance around Tau-U variants, applied researchers are very likely to misuse Tau-U and make inaccurate conclusions, in turn compromising the credibility of obtained findings. The current paper applies the terminology proposed by Fingerhut et al. (2021) and encourages the SCED field to become aware of the different Tau-U measures that are appropriate for different scenarios.

The current paper makes a comprehensive comparison between the different Tau-U variants and identifies the influential factors on Tau-U selection. As a result, this flowchart can help researchers make more informed conclusions about intervention effectiveness by encouraging researchers to pick an appropriate Tau-U variant depending on the data characteristics. Researchers and practitioners in various fields (e.g. communication disorders, behavioral sciences, special education, among others) rely on evidence-based interventions to guide their work.

Choosing inappropriate measures can result in invalid interpretations of intervention outcomes. Practitioners who use data-based decision-making and use inappropriate measures may end up using ineffective or problematic interventions. This flowchart can help researchers make more informed decisions about which Tau-U variant to use. Researchers are also encouraged to use other decision-making tools, such as the a priori flowchart proposed by Manolov et al. (2021) and the SCED measure ranking tool proposed by Fingerhut, Marbou et al. (2020). If, after careful consideration, Tau-U is the measure of choice, applied researchers and practitioners can then use the flowchart proposed within this paper to determine the appropriate Tau-U variant for use.

The flowchart is advantageous because it is straightforward in its application. Researchers follow the direction of the flowchart and consider each of the influential data characteristics, such as within-case variability, intervention effect, trend, and the design conditions such as baseline length. By answering the flowchart questions one by one, researchers are also encouraged to think critically about their SCED data characteristics. In this way, the flowchart serves as a learning tool for applied researchers. This allows researchers to more easily provide appropriate justifications for using measures, which Manolov et al. (2021) and Fingerhut, Moeyaert et al. (2020) have identified as issues within the field of SCED. The flowchart also promotes the use of the Critical Tau-U interpretation method, which can help researchers more accurately interpret estimates as small, medium, or large. Thus, the flowchart is simultaneously easy to use and also promotes more accurate reporting of SCED Tau-U results.

Limitations

The proposed flowchart has several limitations. Firstly, the flowchart is developed to distinguish solely between different Tau-U variants. Thus, the flowchart is only applicable when researchers intend to use Tau-U, which can be limiting. For example, if the researcher has a research question related to magnitude of the intervention effect, Tau-U can immediately no longer be considered for use and the researcher has to select a different measure. This limitation is acknowledged in the flowchart, as it has been incorporated into the first step of the process. Researchers should only use the flowchart when the Tau-U has been determined as a suitable measure for their research.

Another limitation of the flowchart is that it fails to consider all factors that may affect the appropriateness of Tau-U as a measure for use. For instance, autocorrelation is a common data characteristic of SCED studies, but the flowchart does not take autocorrelation into consideration. Autocorrelation is not included to keep the flowchart more easily interpretable by applied researchers, but researchers are encouraged to refer to the original papers (i.e. Parker et al., 2011; Tarlow, 2017) and methodological studies that have been conducted to determine if Tau-U variants are appropriate with various amounts of autocorrelation.

Users of the flowchart should keep in mind that Tau-U is not a perfect measure. The flowchart is helpful for ensuring that Tau-U is used and reported properly, especially due to the popularity of Tau-U. However, while Tau-U is advantageous over previous non-overlap indices, other measures that are not non-overlap indices may be preferable in certain scenarios (e.g. hierarchical linear modeling). Furthermore, the flowchart is only applicable to SCED designs that are built upon AB design logic

(e.g. multiple-baseline designs, withdrawal designs, etc.) The flowchart would not be appropriate for use with alternating treatment designs, for example.

Tau-U $_{adj}$ is recommended for use if there is evidence in support of a monotonic time baseline trend in the anticipated direction of the intervention effect. However, if the trend in the baseline phase is strong, the intervention should not be introduced at all. If there is already a significantly large improvement in the baseline phase, trend in the intervention phase is likely a continuation of the strong trend that is present in the baseline phase. It is always important to consider the entire context of the study and to critically evaluate the SCED data prior to analysis and before choosing and using a measure such as Tau-U.

Future research

The current paper is motivated by the misuse of Tau-U indices, and the flowchart is intended to support applied researchers in their future studies. However, it is also vital to evaluate the quality and reliability of published SCED studies. Thus, future studies can use the flowchart to evaluate the current use of Tau-U indices among published SCED studies. In this manner, the quality of past research findings that utilized Tau-U as a measure can be evaluated. Furthermore, there are other commonly used procedures, such as standardized mean difference (Pustejovsky et al., 2014) and regression-based measures (e.g. Swaminathan et al., 2014). These measures each have advantage and disadvantages; future research can investigate how these measures are applied, and decision-making tools can be developed for selecting and applying other measures.

A more inclusive flowchart could be developed in the future. It may be beneficial if a flowchart considers other stages of the research process, such as the type of SCED (e.g. multiple baseline design, alternating treatment design, etc.), the design conditions (e.g. the number of participants), and the data collection process (e.g. observation session length, recording system, etc.). This would provide additional guidance for applied researchers to facilitate informed decisions along all stages of the research process.

CONCLUSION

As the field of SCED research grows, it has become more essential that SCEDs are analyzed correctly so that the resulting estimates are valid and in turn support effective data-based decision-making. While quantification techniques (i.e. measures) are helpful for inferring intervention effects and for disseminating findings, it is essential that appropriate measures are selected and used as intended. Researchers need to carefully consider the research questions and the data characteristics present within the SCED graphs when choosing a SCED measure. The flowchart proposed within this study is one decision-making tool that can be used to guide the use of the measure Tau-U, and as the field grows, it can be expected that similar helpful tools will be created for use with further SCED measures.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author(s).

FUNDING

This research was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R305D190022. The content is solely the responsibility of the author and does not necessarily represent the official views of the Institute of Education Sciences, or the U.S. Department of Education.

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*(6), 621–631. https://doi.org/10.1016/0005-7967(93)90115-B

Brossart, D. F., Laird, V. C., & Armstrong, T. W. (2018). Interpreting Kendall's Tau and Tau for single-case experimental designs. *Cogent Psychology*, *5*, 1–26. https://doi.org/10.1080/23311908.2018.1518687

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*(5), 531–563. https://doi.org/10.1177/0145445503261167

Brossart, D. F., Vannest, K. J., Davis, J. L., & Patience, M. A. (2014). Incorporating nonoverlap indices with visual analysis for quantifying intervention effectiveness in single-case experimental designs. *Neurological Rehabilitation*, *24*(3–4), 464–491. https://doi.org/10.1080/09602011.2013.868361org/10.1080/09602011.2013.868361

Cakiroglu, O. (2012). Single subject research: Applications to special education. *British Journal of Special Education*, *39*(1), 21–29. https://doi.org/10.1111/j.1467-8578.2012.00530.x

Declercq, L., Jamshidi, L., Fernández-Castilla, B., Beretvas, S., Moeyaert, M., Ferron, J., & Van Den Noortgate, W. (2019). Analysis of single-case experimental count data using the linear mixed effects model: A simulation study. *Behavior Research Methods*, *51*(6), 2477–2497. https://doi.org/10.3758/s13428-018-1091-y

Every Student Succeeds Act. (2015), Pub. L. No.114-95, §4014, 129 Stat. 1802

Ferron, J., Goldstein, H., Olszewski, A., & Rohrer, L. (2020). Indexing effects in single-case experimental designs by estimating the percent of goal obtained. *Evidence-Based Communication Assessment and Intervention*, *14*(1), 6–27. https://doi.org/10.1080/17489539.2020.1732024

Fingerhut, J., Marbou, K., & Moeyaert, M. (2020). Single-Case Quantification Technique Ranking Tool (Version 1.2) [Microsoft Excel® tool]. https://www.doi.org/10.17605/OSF.IO/7USBJ

Fingerhut, J., Moeyaert, M., & Manolov, R. (2020). Literature review of single-case quantitative analysis techniques. https://doi.org/10.31234/osf.io/7yt4g

Fingerhut, J., Xu, X., & Moeyaert, M. (2021). Impact of within-case variability on Tau-U and regression-based effect size measures for single-case experimental data. *Evidence-Based Communication and Intervention*.

Gage, N. A., Grasley-Boy, N. M., & MacSuga-Gage, A. S. (2018). Professional development to increase teacher behavior-specific praise: A single-case design replication. *Psychology in the Schools*, *55*(3), 264–277. https://doi.org/10.1002/pits.22106

Ganz, J., & Ayres, K. (2018). Methodological standards in single-case experimental design: Raising the bar. *Research in Developmental Disabilities*, *79*, 3–9. https://doi.org/10.1016/j.ridd.2018.03.003

Giannakakos, A. R., & Lanovaz, M. J. (2019). Using AB designs with nonoverlap effect size measures to support clinical decision-making: A Monte Carlo validation. *Behavior Modification*. Advance online publication. 1–16. https://doi.org/10.1177/0145445519860219

Horner, R. H., Swaminathan, H., Sugai, G., & Smolkowski, K. (2012). Considerations for the systematic analysis and use of single case research. *Education and Treatment of Children*, *35*(2), 269–290. https://doi.org/10.1353/etc.2012.0011

Jaksic, H., Vause, T., Frijters, J. C., & Feldman, M. (2018). A comparison of a novel application of hierarchical linear modeling and nonparametric analysis for single-subject designs. *Behavior Analysis: Research and Practice*, *18*(2), 203–218. https://doi.org/10.1037/bar0000091

Lane, J. D., & Gast, D. L. (2014). Visual analysis in single case experimental design studies: Brief review and guidelines. *Neuropsychological Rehabilitation*, *24*(3–4), 445–463. https://doi.org/10.1080/09602011.2013.815636

Lenz, A. (2013). Calculating effect size in single-case research: A comparison of nonoverlap methods. *Measurement and Evaluation in Counseling and Development*, *46*(1), 64–73. https://doi.org/10.1177/0748175612456401

Ma, H. (2006). An alternative method for quantitative synthesis of single-subject research: Percentage of data points exceeding the median. *Behavior Modification*, *30*(5), 598–617. https://doi.org/10.1177/0145445504272974

Manolov, R. (2014). *SD_band.R* [R Syntax] https://dl.dropboxusercontent.com/s/elhy454ldf8pij6/SD_band.R

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

Manolov, R., Moeyaert, M., & Fingerhut, J. (2021). A priori metric justification in single-case experimental designs research. *Perspectives on Behavior Science*. https://doi.org/10.1007/s40614-021-00282-2

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. https://doi.org/10.1016/j.jsp.2012.12.005

McClain, M. B., Otero, T. L., Haut, J., & Schatz, R. B. (2014). Single subject design in educational research: The use of effect size measures in analysis and interpretation. In *SAGE Research Methods Cases*. SAGE Publishing. https://doi.org/10.4135/9781446273050145341 73

McKinney, T., & Vasquez, E., III. (2014). There's a bug in your ear!: Using technology to increase the accuracy of DTT implementation. *Education and Training in Autism and Developmental Disabilities*, *49*(4), 594–601. https://www.jstor.org/stable/24582354

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

Ninci, J., Vannest, K. J., Wilson, V., & Zhang, N. (2015). Interrater agreement between visual analysis of single-case data: A meta-analysis. *Behavior Modification*, *39*(4), 510–541. https://doi.org/10.1177/0145445515581327

Parker, R., Hagan-Burke, S., & Vannest, K. (2007). Percentage of All Non-Overlapping Data (PAND): An alternative to PND. *Journal of Special Education*, *40*(4), 194–204. https://doi.org/10.1177/00224669070400040101

Parker, R., & Vannest, K. J. (2009). An improved effect size for single case research: Nonoverlap of all pairs (NAP). *Behavior Therapy*, *40*(4), 357–367. https://doi.org/10.1016/j.beth.2008.10.006

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

Parker, R., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau. *Behavior Therapy*, *42*(2), 284–299. https://doi.org/10.1016/j.beth.2010.08.006

Pustejovsky, J. E. (2019). Procedural sensitivities of effect sizes for single-case designs with directly observed behavioral outcome measures. *Psychological Methods*, *24*(2), 217–235. https://doi.org/10.1037/met0000179

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

Rakap, S. (2015). Effect sizes as result interpretation aids in single-subject experimental research: Description and application of four nonoverlap methods. *British Journal of Special Education*, *42*(1), 11–33. https://doi.org/10.1111/1467-8578.12091

Rohatgi, A. (2016). Webplotdigitizer ver.3.10. http://arohatgi.info/WebPlotDigitizer

Schlosser, R. W. (2009). The role of single-subject experimental designs in EBP times. *Focus: Technical Brief*, *22*, 1–8. https://ktdrr.org/ktlibrary/articles_pubs/ncddrwork/focus/focus22/

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single subject research: Methodology and validation. *Remedial and Special Education*, *8*(2), 24–33. https://doi.org/10.1177/074193258700800206

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

Shadish, W. R., Hedges, L. V., Pustejovsky, J. E., Rindskopf, D. M., Boyajian, J. G., & Sullivan, K. J. (2014). Analyzing single-case designs: D, G, hierarchical models, Bayesian estimators, generalized additive models, and the hopes and fears of researchers about analyses. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 247–281). American Psychological Association.

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*(4), 438–458. https://doi.org/10.1007/s10864-015-9221-4

Swaminathan, H., Rogers, J. J., & Horner, R. H. (2014). An effect size measure and Bayesian analysis of single-case designs. *Journal of School Psychology*, *52*(2), 213–230. https://psycnet.apa.org/doi/10.1016/j.jsp.2013.12.002

Tarlow, K. R. (2016). *Baseline corrected Tau calculator*. Retrieved June 16, 2021, from https://www.ktarlow.com/stats/tau

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

Van Den Noortgate, W., & Onghena, P. (2003a). Hierarchical linear models for the quantitative integration of effect sizes in single case research. *Behavior Research Methods, Instruments & Computers*, *35*(1), 1–10. https://doi.org/10.3758/BF03195492

Van Den Noortgate, W., & Onghena, P. (2003b). Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educational and Psychological Measurement*, *63*(5), 765–790. https://doi.org/10.1177/0013164403251027

Vannest, K. J., & Ninci, J. (2015). Evaluating intervention effects in single-case research designs. *Journal of Counseling and Development*, *93*(4), 403–411. https://doi.org/10.1002/jcad.12038

Vannest, K. J., Parker, R. I., Gonen, O., & Adiguzel, T. (2016). *Single case research: Web based calculators for SCR analysis. (Version 2.0) [Web-based application]*. Texas A&M University. http://www.singlecaseresearch.org

Vannest, K. J., Peltier, C., & Haas, A. (2018). Results reporting in single case experiments and single case meta-analysis. *Research in Developmental Disabilities*, *79*, 10–18. https://doi.org/10.1016/j.ridd.2018.04.029

Vannest, K. J., & Sallese, M. R. (2021). Benchmarking effect sizes in single-case experimental designs. *Evidence-based Communication Assessment and Intervention*.

What Works Clearinghouse. (2020). What works clearinghouse standards handbook, version 4.1. U. S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/handbooks

White, O. R., & Haring, N. G. (1980). *Exceptional teaching: A multimedia training package*. Merrill.

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *Journal of Special Education*, *44*(1), 18–29. https://doi.org/10.1177/0022466908328009

Wolfe, K., Seaman, M. A., & Drasgow, E. (2016). Interrater agreement of the visual analysis of individual tiers and functional relations in multiple baseline designs. *Behavior Modification*, *40*(6), 852–873. https://doi.org/10.1177/0145445516644699