Developing and testing a single-case experimental design tool: improving the way researchers choose and justify quantification techniques

Joelle Fingerhut

University at Albany, State University of New York, joellefingerhut@gmail.com

The University at Albany community has made this article openly available. Please share how this access benefits you.

Follow this and additional works at: https://scholarsarchive.library.albany.edu/legacy-etd

Part of the Special Education and Teaching Commons

Recommended Citation

https://scholarsarchive.library.albany.edu/legacy-etd/2682

This Dissertation is brought to you for free and open access by the The Graduate School at Scholars Archive. It has been accepted for inclusion in Legacy Theses & Dissertations (2009 - 2024) by an authorized administrator of Scholars Archive. Please see Terms of Use. For more information, please contact scholarsarchive@albany.edu.
Developing and Testing a Single-Case Experimental Design Tool: Improving the Way Researchers Choose and Justify Quantification Techniques

by

Joelle E. Fingerhut

A Dissertation
Submitted to the University at Albany, State University of New York
In Partial Fulfillment of
the Requirements for the Degree of
Doctor of Philosophy

School of Education
Division of Educational Psychology & Methodology
Spring 2021
Abstract

Certain quantification techniques may be more appropriate than others for single-case design analysis depending on the research questions, the data/graph characteristics, and other desired features. The purpose of this study was a.) to develop a user-friendly tool that could assist researchers in choosing, justifying, and calculating the estimate of single-case design quantification techniques and b.) to validate and test the effectiveness of the developed tool. A total of sixteen different quantification techniques and nine different research questions, data characteristics, and desired features that may affect the appropriateness of the single-case experimental design quantification technique were identified to be included in the tool. The resulting tool provides a ranking of recommended quantification techniques, from most appropriate to least appropriate, depending on the user input (i.e., the research questions, data characteristics, and desired features that are relevant depending on the graph and data set). A pretest and posttest design was utilized to test the effectiveness of the tool. Five AB graphs were displayed to twenty-five single-case design researchers/practitioners. These participants were asked to choose an appropriate SCED quantification technique, justify their use of the quantification technique, and calculate the resulting estimate both before and after accessing the tool. A rubric was developed to objectively score participants’ responses. Results of a two-tailed paired t-test showed that the tool caused a statistically significant change in ability to choose a quantification technique and justify the use of the quantification technique. There were no statistically significant differences in effectiveness of the tool across academic department affiliation and education level. Results show that those who use single-case designs may need support in choosing and justifying their use of appropriate quantification techniques, but that the developed tool can be a helpful resource. The free tool can be accessed at https://osf.io/7usbj/.
Acknowledgements

Thank you to my dissertation chair, Dr. Mariola Moeyaert, for your support and willingness to review the many ideas and drafts that eventually led to the final product. I am constantly inspired by you and working with you has been an honor in of itself. Thank you to my dissertation committee members, Drs. Benjamin Solomon and Kevin Quinn. You each offered a unique practitioner perspective that helped improve both the quality of the tool and the study, and I enjoyed our meetings. Creating the tool would not have been possible without the help from those who took the time to review the original version and also provide feedback to improve it. Thank you to Drs. Erin Barton, John Ferron, Rumen Manolov, Patrick Onghena, and Robyn Tate.

To my friends and colleagues: Xinyun Xu, I remember when I first told you about this dissertation idea and you were so genuinely excited. You have been a huge figure of support and also a great friend. Holly Meredith, thank you for listening to me practice my presentations, hearing my ideas, and for your constant support. Jess Murray and Panpan Yang, we did it! I am grateful for our lengthy conversations these last few months as we each wrapped up writing our dissertations.

And to my husband Khalid, this would absolutely not have been possible without you.
Table of Contents

Abstract ................................................................................................................................. ii
Acknowledgements ............................................................................................................. iii
Chapter 1: Introduction ........................................................................................................ 1
  Popularity of SCED Research ............................................................................................. 2
  Development of SCED Quantitative Analytical Techniques ............................................. 4
    Non-overlap Methods ....................................................................................................... 5
    Effect Size Metrics .......................................................................................................... 6
  Statement of the Problem .................................................................................................... 7
  Purpose of the Study ........................................................................................................... 11
  Significance of the Study .................................................................................................... 12
Chapter 2: Literature Review ............................................................................................... 14
  Visual Analysis .................................................................................................................... 14
    Masked Visual Analysis .................................................................................................... 17
  SCED Quantification Techniques ...................................................................................... 18
    Non-Overlap Measures .................................................................................................... 19
    Standardized Mean Difference and Related Quantification Techniques ....................... 27
    Regression ....................................................................................................................... 31
    Log Response Ratio and Related Quantification Techniques ......................................... 34
  Considerations for Picking an Appropriate Quantification Technique ............................ 36
    Research Question(s) ....................................................................................................... 37
    Data Characteristics ......................................................................................................... 43
    Desired Quantification Technique Features .................................................................... 58
  Present Study ..................................................................................................................... 66
Chapter 3: Method ............................................................................................................... 69
  Part 1: Development of Tool ............................................................................................. 69
    Creating the Tool: Sheet 1 (Instructions) ......................................................................... 70
    Creating the Tool: Sheet 2 (Tool) ................................................................................... 70
    Creating the Tool: Sheet 3 (Metric Details) ....................................................................... 73
    Creating the Tool: Sheet 4 (Notes and References) ........................................................ 74
    Creating the Examples Document ................................................................................... 74
    Application of Tool with Graph/Data Set ......................................................................... 76
Part 2: Using the Tool as an Intervention ................................................................. 77
  Participants ........................................................................................................... 78
  Materials ................................................................................................................ 81
  Measures ................................................................................................................ 82
  Procedures ............................................................................................................ 85
  Statistical Analysis ............................................................................................... 87
Chapter 4: Results .................................................................................................. 90
Descriptive Statistics ............................................................................................... 91
  Social Validity ....................................................................................................... 91
  Pretest Scores ...................................................................................................... 93
  Change in Scores ................................................................................................. 100
  Time to Complete Study ....................................................................................... 107
Part 3: Inferential Analyses .................................................................................... 108
  Change in Scores ................................................................................................. 109
  Academic Department Differences ...................................................................... 113
  Education Level Differences ................................................................................ 115
  Time to Complete Study ....................................................................................... 118
Chapter 5: Discussion .............................................................................................. 119
Effectiveness of Excel Tool for Choosing Quantification Techniques .............. 119
Effectiveness of Excel Tool for Justifying Quantification Techniques .............. 123
Perceived Effectiveness of the Excel Tool ............................................................ 124
Relation between Findings and Pretest Scores, Academic Department, and Education Level .................................................................................................. 125
  Education Level .................................................................................................. 128
  Academic Department ......................................................................................... 129
Implications of Current Study ............................................................................. 130
  Current Use of SCED Quantification Techniques .............................................. 130
  Changing the Way People Analyze SCED Data .................................................. 133
Limitations and Future Research ......................................................................... 134
  Limitations of the Study ...................................................................................... 134
  Limitations of the Tool ......................................................................................... 137
  Future Research: Summary ................................................................................ 140
Conclusion .............................................................................................................. 141
References ........................................................................................................................................ 143
Appendix 3.1 .................................................................................................................................... 163
Appendix 3.2 .................................................................................................................................... 166
Appendix 3.3 .................................................................................................................................... 172
Appendix 3.4 .................................................................................................................................... 173
Appendix 3.5 .................................................................................................................................... 174
Appendix 3.6 .................................................................................................................................... 179
Appendix 3.7 .................................................................................................................................... 180
Appendix 3.8 .................................................................................................................................... 181
Chapter 1: Introduction

A single-case experimental design\(^1\) (SCED) is a research design that allows repeated measures of the dependent variable within one case (What Works Clearinghouse, 2020). The dependent variable is measured at different sequential moments in time and across different levels of the independent variable (i.e., treatment indicator), which are also known as “phases” (e.g. baseline and treatment phase). The case serves as its own control rather than having multiple participants in an experimental or control group, as is typical for group comparison designs. The goal of using SCEDs is to find evidence for a functional relation (experimental control) between the researcher-manipulated independent variable and the dependent variable of interest. There are several types of SCEDs; these include, but are not limited to, the multiple-baseline design, reversal design and alternating treatment design (See: Ledford, & Gast, 2018; What Works Clearinghouse, 2020).

Two key features of SCEDs are replication and randomization, as these features help establish internal and external validity (Horner et al., 2005). Replication occurs when there are multiple opportunities to measure the effect of the independent variable on the dependent variable; this can occur within participants, between participants, or both. For example, the independent variable can be withdrawn and then reintroduced (reversal, withdrawal, or ABAB design), the implementation of two or more treatment conditions systematically altered within a participant (alternating treatment design), or the independent variable can be introduced at different points of time across different cases, settings, or behaviors (multiple-baseline design). Randomization can enhance the internal validity of SCEDs by ensuring that any results are caused by the intervention and not caused by confounding or underlying variables.

\(^1\) Also known as a single case, single subject, single-subject, interrupted time series, small \(n\), \(n\) of 1 trial, and \(n = 1\).
Randomization prevents experimenter bias because the introduction of the intervention is pre-determined. Furthermore, based on the randomization scheme used, a randomization distribution can be built for hypothesis testing without making any underlying data assumptions.

**Popularity of SCED Research**

The uniqueness of SCEDs presents several benefits over group comparison designs, where inferences are made at the group level rather than at the individual level. SCEDs allow detailed information to emerge about a single case. Repeated measures occur within phases, which allows researchers to evaluate how the behavior or outcome measure changes over time (What Works Clearinghouse, 2020). Thus, detailed information emerges in regards to how the independent variable affects an individual, and within-subject trends can be tracked (Manolov & Moeyaert, 2017b). SCEDs can be useful in certain fields like behavior modification (Shadish & Sullivan, 2011) or medicine (Gabler et al., 2011) when the problem of interest occurs at an infrequent rate and/or it is difficult to find a large number of participants (Shadish, 2014a). An example of this is with autism research. It can be difficult to find enough participants for group design studies, since autism occurs at a low rate (Centers for Disease Control and Prevention, 2019). Therefore, SCEDs are a popular option for autism research (Matson et al., 2012).

Furthermore, SCEDs can have a high degree of external and internal validity when SCEDS are well-designed (e.g., include replication and randomization [Horner, et al., 2005]).

Due to the unique benefits of SCEDs, researchers have used this research design for over 60 years (Horner et al., 2005). Especially since the turn of the century in 2000, there has been an increase in utilization of SCEDs for research. When searching the database PsychINFO for peer-reviewed articles that contain the term(s) "single subject” OR “single-subject” OR “single case” OR “single-case” OR “multiple baseline” OR “multiple-baseline” OR “reversal design” OR
“ABAB design” OR “withdrawal design” OR “alternating treatment design” since 1990, there is an increase in search results beginning from 2002, which appears to taper off in 2015.

**Figure 1.1**

*Number of SCED Related Publications between 1990 and 2019*

![Graph showing the number of SCED related publications between 1990 and 2019.](image)

*Note.* Increase in the number of publications between 1990 and 2019 when searching terms "single subject” OR “single-subject” OR “single case” OR “single-case” OR “multiple baseline” OR “multiple-baseline” OR “reversal design” OR “ABAB design” OR “withdrawal design” OR “alternating treatment design” on PsychINFO database.

This increase in SCED popularity could be attributed to several advances within the field. For example, the development of different standards called for use of SCEDs in certain circumstances. The What Works Clearinghouse document (Kratochwill et al., 2010) recommended SCEDs for reviewing evidence-based practices, while the Council for Exceptional Children (2014) published standards for evidence-based practices in special education, advising researchers to use either group comparison research or single-subject research. Recently in 2018, the Council for Exceptional Children issued a statement further endorsing the use of SCEDs and
how to properly use them. Furthermore, organizations such as Institute of Education Sciences provide researchers with funding and grant money for SCED research, as well as offer yearly training institutes. As bodies of research have advocated and recommended SCEDs to be used to evaluate intervention effectiveness, more researchers may have started to utilize SCEDs as a research method.

It is also possible that SCEDs have become more widely used because fields that utilize SCEDs for research have increased in popularity. SCEDs are commonly used in the field of applied behavior analysis; as part of the course sequence to become a Board Certified Behavior Analyst, it is necessary to enroll in a verified course sequence that includes a course about using SCEDs (Behavior Analyst Certification Board, 2017). Applied behavior analysis has become increasingly popular, which has been demonstrated by the steady (and exponential, in some areas of the U.S.) growth of registered Board Certified Behavior Analysts since 2000 (Deochand & Fuqua, 2016). Thus, the growth in popularity of applied behavior analysis could also contribute to the overall increase in the use and publication of research that utilizes SCEDs.

**Development of SCED Quantitative Analytical Techniques**

Historically, SCEDs have been analyzed using visual analysis techniques (Horner et al., 2012). Visual analysis is the traditional method for analyzing data and is widely accepted in many fields as the primary method to determine if an effect is present in a SCED (Ledford et al., 2019). Visual analysis allows data to be closely examined, as it allows for detailed descriptions to be made regarding what is observed (e.g., immediate or delayed effect and data consistency). Visual analysis can be used throughout the study to determine if adjustments should be made\(^2\) (e.g. if baseline trend is present, researchers may decide to extend baseline phase instead of

---

\(^2\) Also known as “response-guided designs”. See Swan et al. (2020).
implementing the intervention), thus demonstrating its use as a type of formative assessment (Ledford et al., 2018). Although visual analysis is used frequently, research has demonstrated a lack of interrater agreement with visual analysis (Brossart et al., 2006; Ninci et al., 2015) and its tendency to result in Type II errors (McClain et al., 2014).

As SCEDs have become more popular, there has been a rise in the development of new quantitative analysis techniques for SCED researchers to use complementary with visual analysis. As certain U.S. federal government laws (e.g., No Child Left Behind Act of 2001; Individuals with Disabilities Improvement Act of 2004) were enacted that called for the use of evidence-based practices, researchers who used SCEDs needed to ensure that they could effectively analyze the outcomes from the data (Solomon et al., 2015). Furthermore, there has been an increase in outside pressure from evidence-based practice communities to use statistics that can also be understood by group comparison design researchers (Shadish, 2014b; What Works Clearinghouse, 2020). Shadish (2014b) also acknowledged pressure from journals and publishing groups to quantify SCED results, and that editors from journals such as the Journal of School Psychology and School Psychology Quarterly have expressed the desire for more researchers to use quantitative analyses with SCEDs. As a result of this increase in pressure to quantify SCED outcomes, many different statistical analysis techniques for SCEDs have been developed.

**Non-overlap Methods**

Non-overlap methods/indices allow for the quantification of an effect, unlike traditional visual analysis. Non-overlap indices express the treatment effect by means of percentages/proportions by determining the amount of overlap, non-overlap, or overlap minus non-overlap between data points in different phases (e.g., between baseline and treatment phase
or between two treatment phases [Parker et al., 2011a]). Most non-overlap indices do not require data assumptions (e.g., that errors are independent, homogeneous and normally distributed). Non-overlap indices can typically be calculated by hand or with online calculators (e.g., http://www.singlecaseresearch.org/; Vannest et al., 2016) and do not require extensive training to implement (Vannest & Ninci, 2015).

Several different non-overlap indices have been developed for use with SCEDs. The most frequently used quantification technique is percent of non-overlapping data (PND; Scruggs et al., 1987), which calculates the percentage of data that does not overlap between the baseline and treatment phase(s). Percentage of non-overlapping corrected data (PCND; Manolov & Solanas, 2009) is similar to PND, but controls linear trend that might be present. Another non-overlap index is improvement rate difference (IRD; Parker et al., 2009), which looks at how much improvement is present between the phases. Non-overlap of all pairs (NAP; Parker & Vannest, 2009) considers the comparison between all pairs of data points between the baseline and treatment phase. The newer-developed Tau-U (Parker et al., 2011a) builds off of NAP and is the percent of non-overlapping data minus the percent of overlapping data. Tau-U has also been adjusted for trend with Theil-Sen robust regression (baseline corrected Tau-U; Tarlow, 2017), yielding an interpretable estimate.

Effect Size Metrics

Effect size quantification techniques allow for quantifying the size of an effect. They also have a known sampling distribution, and so a standard error can be calculated (which is useful for statistical significance testing). Effect size quantification techniques include regression-based effect sizes (e.g. Center et al., 1985-1986), the standardized mean difference (SMD; Hedges et
al., 2012), the log-response ratio (Pustejovsky, 2018), and the design-comparable effect size estimator (Pustejovsky et al., 2014).

These more sophisticated methods can be too complex for some researchers to implement without proper training, but are beneficial over non-overlap indices in that they allow for the magnitude of the effect to be quantified (in addition, they account for other data complexities such as autocorrelation, trends, variability, etc.).

**Statement of the Problem**

While the development of quantification techniques has helped to further develop the SCED field, there remains a gap in knowledge as to how to select an appropriate quantification technique for use, as well as how to implement and report these quantification techniques appropriately (Solomon et al., 2015). Using quantification techniques improperly can influence the interpretation of results (e.g., whether an intervention is effective or not). Researchers may improperly use quantification techniques by not properly accounting for data characteristics. One example of this is when researchers utilize quantification techniques that cannot account for trend (e.g., Tau-U and PND) when trend is present in the data, which can influence the inference about treatment effectiveness. Brossart et al. (2018) found that there is a small to medium trend in the majority of SCED research graphs, therefore indicating that using quantification techniques like Tau-U and PND are likely inappropriate in most instances. Other research shows that certain quantification techniques (e.g., Tau) are impacted by within-case variability (Fingerhut et al., in press-a), and so other quantification techniques are preferable for use when variability is present. A recent study found that out of 161 SCED meta-analyses, 99 (61%) of studies used the average PND for synthesizing SCED meta-analyses, even though PND has several crucial limitations (Jamshidi et al., 2020). Although some quantification techniques are
preferable over others depending on the data characteristics and the research question(s) of interest (see: Manolov & Moeyaert, 2017b; Solomon, 2014), researchers may be unaware that different quantification techniques are available for use, and so they may simply use the one(s) with which they are most familiar.

Furthermore, researchers may report quantification techniques incorrectly and make wrong interpretations; or they may use one quantification technique and report that they used a different one. Brossart et al. (2018) reported this could be an issue when researchers use the *Single Case Research* website (http://www.singlecaseresearch.org/; Vannest et al., 2016). This online calculator uses vague language in regards to which Tau is calculated (traditional or trend adjusted), and so researchers that use this popular calculator may erroneously report using the original Tau-U index even though they may have used the baseline-corrected Tau-U.

Other issues related to using SCED quantification techniques include confusion over scales for interpreting results (i.e., what indicates a small, medium, or large effect). For example, although Vannest and Ninci (2015) urge researchers to consider the context of their own research when interpreting results, many researchers reference Vannest and Ninci’s proposed benchmarks stringently (i.e., 0.20 as a small change, 0.20 to 0.80 a moderate/large change, and above 0.80 a very large change). This misuse of the Tau benchmark scale shows how SCED researchers may interpret results incorrectly. Although Fingerhut and colleagues (in press-a) propose an alternative method for interpreting Tau results, a similar issue remains for interpreting the results of other quantification techniques.

Lastly, researchers rarely report their justification for the quantification technique that they use (Solomon et al., 2015) or use quantification techniques for reasons that do not align with
the quantification technique founders (Fingerhut et al., 2020). This can contribute to the improper use of these quantification techniques.

Still, other researchers do not use SCED quantification techniques at all and instead rely primarily on visual analysis to determine if an effect is present. This overreliance on visual analysis can occur for several reasons including that researchers cannot decide which quantification technique to use, they use visual analysis habitually, and/or do not know the advantages of using quantitative measures (Shadish, 2014b). An examination of 74 articles published in 2018 in the *Journal of Applied Behavioral Analysis* demonstrates that 100% of the articles published within this journal do not report an effect size measure; only one reported a non-overlap statistic (PND), while the remaining reported visual analysis and/or means and ranges. Other studies (e.g., Peng et al., 2013) found that only half of group design and SCED research studies report an effect size measure.

The overuse of traditional visual analysis poses several issues. As previously mentioned, several studies have demonstrated that visual analysis is unreliable and that it can be difficult to identify small effects when using visual analysis (Brossart et al., 2006; McClain et al., 2014; Ninci et al., 2015). Therefore, more recent research calls for researchers to move away from using only visual analysis to determine evidence of an effect, and instead to use it as a supplement along with other SCED quantification techniques (Heyvaert et al., 2015; Vannest et al., 2018).

These issues demonstrate a need for greater collaboration and communication between two scientific communities: researchers and methodologists. Although numerous SCED quantification techniques have been developed, researchers do not use them appropriately and/or do not use them at all. It is possible that researchers remain unaware of recent developments in
analyzing SCED data, even though efforts have been made by those in the methodology community to disseminate this information. For example, several special issues regarding SCED analytical techniques have been published within applied journals (e.g., “Single Subject Causal Mediation Analysis” in *Evaluation and the Health Professions*, 2020; “Advances in Statistical Analysis and Meta-analysis of Single-Case Experimental Designs” in *Evidence-Based Communication Assessment and Intervention*, 2020; “Advances in Single-Case Research Design and Analysis” in *Developmental Neurorehabilitation*, 2018). Online web calculators have also been developed for researcher and practitioner use (e.g., MultiSCED Tool [Cools et al., 2017]; http://www.singlecaseresearch.org/ [Vannest et al., 2016]). Perhaps this method of disseminating methodological information to those who would use it in practical settings is not effective. It is also possible that researchers are aware of these quantification techniques, but have not learned how to appropriately use them. More complex methods that have been developed for use like piecewise regression (Center et al., 1985-1986) and HPS d statistic (Hedges et al., 2012) are advantageous quantification techniques, but the difficulty in calculating them may prevent researchers from using them. Researchers may easily understand other quantification techniques that have simpler formulas (e.g., PND), but remain unaware of how to use these quantification techniques appropriately. Lastly, it should be considered that researchers may be more likely to use certain quantification techniques over others depending on their affiliated research institution, which may routinely use one quantification technique over another. Overall, the lack of communication between different research communities hinders the advancement of research in fields such as special education and applied behavior analysis, which rely heavily on the use of SCEDs.
Purpose of the Study

The first purpose of this study was to develop and validate a free and user-friendly Excel tool for researchers to utilize to help them choose an appropriate SCED quantification technique for use in quantifying SCED results. This tool was developed based upon previous research (e.g., Manolov & Moeyaert, 2017b) that outlined how certain quantification techniques are more appropriate than others depending on the SCED data characteristics and research questions of interest. The goal of this tool was to assist researchers in determining appropriate SCED quantification techniques for use, as well as providing a justification for these quantification techniques, taking into account the data/graph characteristics, as well as the research question(s). The tool focused on quantification methods rather than visual analysis techniques, and focused specifically on within-case quantification techniques rather than between-case quantification techniques. Different dimensions identified by Manolov et al. (2021) including research question(s), data characteristics, and desired quantification technique features were presented in a digital chart, along with different facets of these dimensions (“trend in baseline”, “number of observation points in baseline”, etc.). These were shown alongside options such as “yes” or “no” for researchers to mark accordingly, if they needed or wanted the SCED quantification technique to account for each facet. The tool was programmed to produce an output based on the research questions(s) and data characteristics reported. This output lists viable quantification techniques for researchers to use with the SCED data.

The second purpose of this study was to test the effectiveness of the tool in assisting users to choose an appropriate SCED quantification technique, provide a justification for the quantification technique, and perform the correct calculation. First, users completed a demographic survey. This survey asked questions such as academic department affiliation, education level, etc. Next, the users analyzed five AB graphs (i.e., pretest graphs), and provided
their choice of quantification technique along with a justification for using that quantification technique. Completing the calculation was optional

After analyzing the AB graphs, the users were introduced to the tool (i.e., the intervention). The participants were given time to practice using and interacting with the tool, and were given five new graphs, data sets, and research questions to practice using with the tool. After practicing using the tool, participants were given the same five AB graphs/ data sets and research questions used for pretest, and asked to use the tool to determine the most appropriate quantification technique, write a justification, and perform the correct calculation. At the end, participants completed a short survey asking their opinion of the tool (i.e., “was the Excel tool useful for finding an appropriate quantification technique to use?”).

A rubric was used to rate both the appropriateness of the quantification technique, the justification of the quantification technique, and the calculation provided by the participant. If there was a significant positive change between pretest and posttest of the intervention (i.e., the tool) across the three indicators of the rubric (quantification technique appropriateness, quantification technique justification, and quantification technique calculation), it was concluded that the tool was an effective tool for picking, justifying, and calculating a SCED quantification technique. The demographic survey results were examined to determine if there were relationships between ability to analyze SCED graphs and academic departments or education level, and the social validity survey results were analyzed to determine acceptability of the tool amongst users.

**Significance of the Study**

Teaching researchers that SCED quantification techniques should be carefully chosen for use helps to prevent inaccurate implementation of these quantification techniques in the future.
Teaching researchers to use SCED quantification techniques in a more intentional way can help others understand the magnitude of the effect, as well as the clinical significance (McClain et al., 2018). Furthermore, the tool proposed in this study can ultimately prevent confusing reporting methods and encourage more accurate reporting. Ultimately, these changes help researchers more accurately determine which interventions are evidence-based (Heyvaert et al., 2015), and are helpful for SCED instructors and those training others how to use SCED quantification techniques appropriately.

Furthermore, encouraging appropriate use of SCED quantification techniques helps ensure that SCEDs are respected as a legitimate form of research. This is a concern because many agencies do not consider SCEDs to be rigorous designs, exempting them from receiving grant money and funding (Matson et al., 2012). Encouraging the appropriate use of quantification techniques may improve the credibility of SCEDs within the larger research community (Parker & Hagan-Burke, 2007).

Teaching researchers how to use appropriate quantification techniques demonstrates that it is possible to change the behavior of researchers who appear to be routinely using inappropriate techniques. This study demonstrates that it is possible for there to be effective communication between researchers and methodologists, and that it is possible to change the routine behavior of researchers. Ultimately, teaching researchers to use more appropriate analytical techniques demonstrates that it is possible to effectively explain these quantification techniques to researchers, and that effective dissemination of information is possible.
Chapter 2: Literature Review

Many analytical techniques have been established for the quantitative synthesis of SCED data (Manolov & Moeyaert, 2017b). These techniques can be classified into several different categories. The first category includes visual analysis-related techniques such as masked visual analysis (Ferron & Jones, 2006) and structured visual analysis (Ledford et al., 2018). A second category are non-overlap indices, such as percent of non-overlapping data (Scruggs et al., 1987) and baseline corrected Tau-U (Tarlow, 2017). A third category are the standardized mean difference effect size quantification techniques (Hedges et al., 2012, 2013; Manolov & Solanas, 2013). A fourth category are regression-based techniques (e.g., Center et al., 1985-1986; Maggin et al., 2011; Van den Noortgate & Onghena, 2003). The last category are log response ratio and related techniques (Ferron et al., 2020; Pustejovsky, 2018).

In this literature review, an overview of these analytical techniques is provided. First, visual analysis and related techniques are discussed, followed by a discussion of the quantification techniques that can provide a quantification (i.e., non-overlap, standardized mean difference, regression, and log-response ratio). In the second half of this literature review, the different components that can be considered when selecting an appropriate SCED quantification technique are discussed.

Visual Analysis

Visual analysis is a historically popular method for determining the presence of an effect (Wolfe et al., 2019), and is still viewed as a critically important method for analyzing SCED data in certain fields (e.g., special education; Ledford et al., 2018). Visual analysis has been described as a “holistic” method of evaluating the presence of an intervention effect (Ninci, 2019; What Works Clearinghouse, 2020). The newer What Works Clearinghouse (2020) standards no longer
recommend relying on visual analysis for determining if an effect is present; however, six
different aspects of SCED data have commonly been visually analyzed to determine if an effect
is present. These six aspects include: level, trend, variability (e.g., Ledford & Gast, 2018),
immediacy of effect, overlap of data points, and consistency of data patterns across similar
phases (Barton et al., 2018). Level refers to the mean (average) for the data within each phase.
Trend is the best fitting line for the data within a phase. Variability is the range of data around a
best fitting line for the average. The immediacy of the effect can be defined as the change in
level between the last three data points in one phase and the first three data points in the next
phase (Kratochwill et al., 2010). Overlap refers to the amount of data that overlaps with data in
the next phase. Consistency of data in similar phases means looking at the data from phases of
the same condition and comparing them to determine the consistency of data patterns.

Researchers such as Ledford and colleagues (2019) recommend using visual analysis for
formative decision-making, demonstrating the still current and important role visual analysis has
in analyzing data. However, relying solely on visual analysis to determine if an effect is present
in a SCED graph can be problematic for several reasons. Dart and Radley (2017) found that the
way data is displayed can affect the results of visual analysis, and Barton et al. (2019) found that
researchers use visual analysis terms inconsistently. Other research (Wolfe et al., 2016) indicates
visual analysis has weak reliability, meaning that researchers can reach different conclusions
when analyzing the same graph. Visual analysis does not allow for the quantification of the
treatment effect, which can be a problem in certain fields like special education where evidence
that an intervention is effective is needed to establish programs for students (Odom et al., 2018).
Furthermore, visual analysis does not allow for research findings to be communicated reliably.
Meta-analyses are also not possible, which further hampers the ability to identify evidence-based interventions.

Taken together, these problems with visual analysis indicate that it may be best to use visual analysis paired with quantification techniques (e.g., Barton et al, 2019). However, visual analysis is still used independently (i.e., not used with a quantitative measure) and as a summative measure (meaning it is used to make inferences about the effect after the study is complete). Fingerhut et al. (2020) found that 53% of studies related to SCEDs published in 2019 used visual analysis without quantifying the intervention effect. Researchers may be unaware of the criticisms of over relying on visual analysis to determine if an effect is present, or may be unaware of the different analytical techniques that can also be used for SCEDs. In the following sections, two different types of visual analysis are discussed (structured and masked visual analysis), which are more complex methods than traditional visual analysis, previously described.

**Structured Visual Analysis**

There have been efforts to encourage researchers to use visual analysis more systematically in order to combat some of the aforementioned issues with visual analysis. Structured, or systematic, visual analysis is when SCED graphs are visually analyzed with tools to help support the reliability and validity of determining if an intervention effect is present. Reliability is defined as the consistency of obtaining the same outcome, and validity is the extent to which the outcome actually represents the variable that is being measured. Structured visual analysis increases the likelihood that two different researchers will reach the same conclusions, thus increasing the reliability, as well increasing the likelihood that results will align with results from statistical analyses (Fisher et al., 2003), which helps increase external validity. One such
tool for structured visual analysis is known as stability envelopes (Lane & Gast, 2014); lines are drawn a specified distance away from the median, above and below the median. Stability envelopes can be structured around the split-middle trend line (a trend line centered around the median or mean data point across all measurement occasions) in order to evaluate trend stability. Another tool known as conservative dual criteria (Fisher et al., 2003) extends the mean and split-middle line into the intervention phase, and then adjusts these lines in the direction of the estimated effect. A more recently proposed method, referred to as visual aid implying an objective rule, has nine steps that can be followed to help strengthen visual analysis (see: Manolov & Vannest, 2019). Manolov et al. (2016) explain how stability envelopes and other related tools can be used for systematic visual analysis.

**Masked Visual Analysis**

Masked visual analysis is a systematic technique that can be used to help prevent Type I errors when using a response-guided method (Ferron & Jones, 2006). A randomization distribution is built upon the different randomization strategies that are incorporated into the SCED. Thus, randomization tests can be used, controlling for Type I error (Ferron & Jones, 2006).

The steps for using masked visual analysis for a multiple-baseline design are as follows:

a.) The study parameters are outlined. Researchers determine the design type, minimum number of observations per phase, and the randomization scheme.³ 

b.) The researchers split into an analysis team and an intervention team. 

c.) The study begins; outcome data is collected and the intervention team sends the data to the analysis team. 

d.) The analysis team analyzes the data and

---

³ Randomization algorithms include order of participants, start point of the intervention, blocked randomization, etc.).
determines when a random assignment should be made. e.) The intervention team makes random assignments, as indicated by the analysis team. Data is sent back to the analysis team, but the random assignment results are not shared with the analysis team. f.) This process continues until the analysis team determines that the study should be finished. g.) The analysis team announces which participants they believe were introduced first, second, etc. h.) The analysis team reveals the order of random assignment (e.g., which participant was introduced first, second, and third to the intervention). If the analysis team was wrong, the intervention team indicates to the analysis team to continue to specify the possible order or random assignment until they are correct. i.) A $p$-value is computed by dividing the number of specifications made by the analysis team by the total number of possible random assignments (i.e., the total possible combinations of order of participants. If there were 4 participants total, there are 24 possible ways that the participants could have been staggered and introduced to the intervention [4! = 24]).

Although structured visual analysis and masked visual analysis are useful visual analytical techniques, the remainder of this literature review and larger study will focus on quantitative SCED quantification techniques.

**SCED Quantification Techniques**

A variety of SCED quantification techniques have been established for use with SCEDs. These quantification techniques are outlined in the next sections, starting with non-overlap indices, the standardized mean difference and related quantification techniques, regression-based quantification techniques, and log-response ratio and related measures. The quantification techniques and the quantification technique founders, steps for calculating the quantification techniques, free and open source tools to help calculate the quantification technique, and examples of applied studies using the quantification techniques are provided in tables, as well.
**Non-Overlap Measures**

Non-overlap indices are non-parametric measures, with the exception of a few non-overlap techniques that are parametric. Non-overlap indices reflect the amount of overlapping or non-overlapping data between baseline and intervention phases. Several non-overlap indices have been developed for use with SCEDs. Percent of non-overlapping data (PND; Scruggs et al., 1987) is commonly thought of as the easiest and most popular non-overlap quantification technique to calculate (Parker et al., 2014). Percent of all non-overlapping data (PAND; Parker et al., 2007) and percentage of data points exceeding the median (PEM; Ma, 2006) are two non-overlap indices, among others, that were created to improve upon issues with PND. The improvement rate difference (IRD; Parker et al., 2009) is a quantification technique that was created to improve upon both PND and PAND. The extended celebration line (ECL; White & Haring, 1980), also known as percentage of data points exceeding median trend (PEM-T), is similar to PEM, but considers baseline trend. The percent of zero data (PZD; Scotti et al., 1991) can be used to determine the effectiveness of an intervention when reducing the frequency of a behavior. Non-overlap of all pairs (NAP; Parker & Vannest, 2009) is defined as the percentage of data that improve from the baseline phase to the intervention phase, and considers all data points in the calculation. Tau indices expand upon the original calculation of NAP by removing the amount of overlap from the percentage of non-overlapping data. Tau-U Trend A (Parker et al., 2011a) and baseline corrected Tau-U (Tarlow, 2017) are versions of Tau-U that account for trend in the baseline phase. Tau-U Trend A accounts for baseline trend by calculating the monotonic trend within the baseline phase, and then subtracting this from the original Tau-U calculation. Baseline corrected Tau-U removes the baseline trend with Theil-Sen robust regression. The Tau quantification techniques have more complex calculations; Tau quantification techniques do not
truly reflect the amount of overlapping or non-overlapping data due to their formulas, which include both non-overlapping data points and overlapping data points.

The non-overlap quantification techniques with the original source, directions for calculating the quantification technique, calculators and tools to help calculate the quantification technique, and examples in the literature are displayed in Table 2.1.
Table 2.1

Non-Overlap Quantification Techniques: Quantification Technique, Original Source, Calculation, Free and Open-Source Tools, and Applied Examples

<table>
<thead>
<tr>
<th>Quantification technique</th>
<th>Original reference</th>
<th>How it is calculated</th>
<th>Free/open-source tools</th>
<th>Examples in applied studies</th>
</tr>
</thead>
</table>
| PND                      | Scruggs et al. (1987) | 1. Calculate the number of data in the intervention phase that are greater than the highest data point in the baseline phase. 2. Divide by the total number of data points in the intervention phase. 3. Multiply by 100. | Online Calculators:  
http://ktarlow.com/stats/pnd/ (Tarlow & Penland, 2016)  
https://manolov.shinyapps.io/Overlap/ (Manolov, n.d.)  
R Tools:  
SingleCaseES (Pustejovsky & Swan, 2017)  
Scan (Wilbert & Lueke, 2019)  
| PAND                     | Parker et al. (2007) | 1. Determine how many data points need to be removed so there is no between-phase overlap. 2. Divide the percentage of data points that remain by the original total. | R Tools:  
SingleCaseES (Pustejovsky & Swan, 2017)  
Scan (Wilbert & Lueke, 2019) | Guzman et al. (2016) |
<table>
<thead>
<tr>
<th>Quantification technique</th>
<th>Original reference</th>
<th>How it is calculated</th>
<th>Free/open-source tools</th>
<th>Examples in applied studies</th>
</tr>
</thead>
</table>
| PEM                      | Ma (2006)          | 1. Determine the median data point in the baseline phase.  
2. Extend a line from this point into the intervention phase.  
3. Calculate the proportion of data points in the intervention phase that fall above this line.  
4. Multiply the proportion by 2 and subtract 1. | **Online Calculators:**  
https://manolov.shinyapps.io/Overlap/ (Manolov, n.d.)  
**R Tools:**  
SingleCaseES (Pustejovsky & Swan, 2017)  
Scan (Wilbert & Lueke, 2019)  
SCDA plug-in for R (Bulté & Onghena, 2012) | Hurwitz et al. (2020)  
Taylor et al. (2019) |
| ECL/ PEM-T               | White & Haring (1980)  
Wolery et al. (2010) | 1. Extend a median-based trend line from baseline phase into the intervention phase.  
2. Count the number of data points in the intervention phase that are above and below the trend line. | **Online Calculators:**  
http://ktarlow.com/stats/pnd/ (Tarlow & Penland, 2016)  
https://manolov.shinyapps.io/Overlap/ (Manolov, n.d.)  
**R Tools:**  
<table>
<thead>
<tr>
<th>Quantification technique</th>
<th>Original reference</th>
<th>How it is calculated</th>
<th>Free/open-source tools</th>
<th>Examples in applied studies</th>
</tr>
</thead>
</table>
| IRD                      | Parker et al. (2009) | 1. Determine how many data points need to be removed so there is no between-phase overlap.  
2. Divide the number of data points that remain in each phase by the original total number of data points in the same phase.  
3. Determine the difference between these the two outcomes. | **Online Calculators:**  
| PZD                      | Scotti et al. (1991) | 1. Identify the first data point within the intervention phase that is 0.  
2. Calculate the percentage of data points in the intervention phase that is 0. | **R Tools:**  
SingleCaseES (Pustejovsky & Swan, 2017)  
<table>
<thead>
<tr>
<th>Quantification technique</th>
<th>Original reference</th>
<th>How it is calculated</th>
<th>Free/open-source tools</th>
<th>Examples in applied studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAP</td>
<td>Parker &amp; Vannest (2009)</td>
<td>intervention phase after the first 0 data point that remain at 0 (including the first 0 data point that was identified).</td>
<td><strong>Online Calculators:</strong> <a href="http://www.singlecaseresearch.org/calculators/nap">http://www.singlecaseresearch.org/calculators/nap</a> (Vannest et al., 2016)</td>
<td>Wendt et al. (2019)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1. Make pairwise comparisons between all data points in the baseline and intervention phase.</td>
<td><strong>R Tools:</strong> Scan (Wilbert &amp; Lueke, 2019)</td>
<td>Viskochil (2017)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. Mark the pairs as a positive change (pos), no change (ties), or negative change (neg).</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. Calculate: ((\text{pos} + 0.5 \times \text{ties}) / \text{total number of pairs}).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tau-U</td>
<td>Parker et al. (2011a)</td>
<td>1. Order the observed values in baseline and treatment phases in a time ordered manner of occurrence in a row.</td>
<td><strong>Online Calculators:</strong> <a href="http://www.singlecaseresearch.org/calculators/tau-u">http://www.singlecaseresearch.org/calculators/tau-u</a> (Vannest et al., 2016)</td>
<td>Birri (2018)</td>
</tr>
<tr>
<td>Quantification technique</td>
<td>Original reference</td>
<td>How it is calculated</td>
<td>Free/open-source tools</td>
<td>Examples in applied studies</td>
</tr>
<tr>
<td>--------------------------</td>
<td>--------------------</td>
<td>----------------------</td>
<td>------------------------</td>
<td>----------------------------</td>
</tr>
</tbody>
</table>
| Tau-U Trend A            | Parker et al. (2011a) | 1. Do pairwise comparisons as outlined for Tau-U, but for Trend A only.  
2. Calculate: (#pos - #neg) / #total pairs.  
**R Tools:**  
Scan (Wilbert & Lueke, 2019)  
| Baseline corrected Tau-U | Tarlow (2017) | 1. Estimate the trend in the baseline phase with the Theil-Sen robust regression estimate (the median of the time variable and the baseline variable. | http://ktarlow.com/stats/tau/ (Tarlow, 2016)  
**R Tools:**  
<table>
<thead>
<tr>
<th>Quantification technique</th>
<th>Original reference</th>
<th>How it is calculated</th>
<th>Free/open-source tools</th>
<th>Examples in applied studies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2. Calculate: (#pos - #neg) / #total pairs.</td>
<td>Scan (Wilbert &amp; Lueke, 2019)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. Subtract the outcome obtained from step 1.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Standardized Mean Difference and Related Quantification Techniques

The standardized mean difference is an effect size quantification technique that expresses the size of the intervention effect in relation to the standard deviation. One of the original standardized mean difference quantification techniques that is meant for use with group design studies rather than SCEDs is Cohen’s \( d \) (Cohen, 1992), which is the difference between the two group means divided by a pooled standard deviation. Hedges’ \( g \) (Hedges, 1981) is a version of Cohen’s \( d \) that has a bias adjustment (the standard deviation for each group is weighted by the sample size), making it appropriate for group design studies with a small number of cases. A version of Hedges’ \( g \) was developed for use with SCEDs, and is known as HPS-\( d \) (Hedges et al., 2012, 2013). This updated quantification technique uses a simple hierarchical model that depends on moment estimation techniques (Wolfe et al., 2019). However, this quantification technique assumes that there is no baseline trend, that the treatment effect is immediate, and that the effect is consistent across all cases (Hedges et al., 2013). An updated version of this quantification technique, known as between-case standardized mean difference (BC-SMD; Pustejovsky et al., 2014), removes the restriction that the intervention effect is constant across cases by using a restricted maximum likelihood estimation method. BC-SMD can also be characterized as a regression-based quantification technique, as it uses a two-level model; level 1 is a within-case regression model and level 2 is between-case variation in regression coefficients (Pustejovsky et al., 2014). Thus, BC-SMD is preferable for use across cases due to its ability to account for between-case differences. Furthermore, BC-SMD is on the same scale of group designs, and so Cohen’s \( d \) scale can be used for interpretation.

Two other SCED quantification techniques that relate to standardized mean difference are the mean phase difference (MPD; Manolov & Solanas, 2013) and mean baseline reduction (MBLR; Campbell, 2004). Both formulas for MPD and MBLR are similar to Cohen’s \( d \), but
without the standard deviation denominator. Thus, they are technically unstandardized mean difference quantification techniques.

Table 2.2 shows the original reference, calculation, sources, and examples in the literature for the standardized mean difference related quantification techniques.
Table 2.2

*Standardized Mean Difference and Related Quantification Techniques: Quantification Technique, Original Source, Calculation, Free and Open-Source Tools, and Applied Examples*

<table>
<thead>
<tr>
<th>Quantification technique</th>
<th>Original reference</th>
<th>How it is calculated</th>
<th>Free/open source tools</th>
<th>Examples in applied studies</th>
</tr>
</thead>
</table>
| MPD                      | Manolov & Solanas (2013) | 1. Calculate the trend in the baseline phase and project into the intervention phase.  
<table>
<thead>
<tr>
<th>Quantification technique</th>
<th>Original reference</th>
<th>How it is calculated</th>
<th>Free/open source tools</th>
<th>Examples in applied studies</th>
</tr>
</thead>
</table>
2. Subtract the outcome from step 1 from the average baseline phase observation.  
3. Divide outcome from step 2 by the average baseline observation.  
4. Multiply by 100. | R Tools:  
https://www.dropbox.com/s/wt1qu6g7j2ln764/MBLR.R?dl=0. (Manolov, 2015b) | Boyd, et al. (2011) |
Regression

Different types of regression-based approaches have been used for SCEDs. Estimation methods and modeling approaches include, but are not limited, to ordinary least squares (OLS; Gorsuch, 1983; Huijtema & McKean, 2000), generalized least squares (GLS; Swaminathan, Rogers, & Horner, 2014; Swaminathan, Rogers, Horner, et al., 2014), and multilevel models (Moeyaert et al., 2013; Van den Noortgate & Onghena, 2003). OLS is a parametric approach that, in the case of SCED, can quantify the change in level between baseline and intervention phase, as well as the trend (Huijtema & McKean, 2000). Additional parameter estimates can be added to the model depending on the research question (e.g., it can be used to estimate changes in trend and/or level). However, OLS makes several assumptions, including the assumption of normality and homoscedasticity of errors. In addition, the errors are assumed to be independent (i.e., absence of autocorrelation). In instances where these assumptions may be inappropriate to make (e.g., if heteroscedasticity or autocorrelation is present), GLS approach is a more viable and appropriate estimation method. GLS is similar to OLS, except that it can reflect count data and account for autocorrelation by adjusting the residuals (Swaminathan, Rogers, Horner, 2014). Like OLS, GLS is able to reflect both change in slope and change in level together in one model and provide separate estimates of these components.

Hierarchical linear modeling (HLM; Van den Noortgate & Onghena, 2003) is an extension of single-level regression analysis and can be used to estimate change in level and/or slope for individual participants and an overall estimate across participants. HLM is able to

---

4 If only change in level is considered and the data is also standardized, this can be classified as a standardized mean difference approach. When slope and level change are both estimated, and the trend is estimated as the mean of the differenced measurements in the first phase (Solanas et al., 2010), this is known as slope and level change (SLC; Solanas et al., 2010). This metric is considered to be part of the general regression-based quantification technique described in this section.
account for between-case differences (i.e., heterogeneity of variance), and can account for the
nesting of participants (i.e., measurement occasions within participants and participants within
studies). Although fixed effects at the participant or student level could be calculated using OLS,
GLS is the more appropriate estimation method for use with HLM, as GLS meets the assumption
of homoscedasticity.

Table 2.3 displays more information about the regression-based quantification technique
(GLS estimation method) and HLM. OLS is not included, as it is similar to GLS except for the
aforementioned differences.
Table 2.3

Regression: Quantification Technique, Original Source, Calculation, Free and Open-Source Tools, and Applied Examples

<table>
<thead>
<tr>
<th>Quantification technique</th>
<th>Original reference</th>
<th>How it is calculated</th>
<th>Free/open source tools</th>
<th>Examples in applied studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression-based measure (GLS estimation method)</td>
<td>Huietma &amp; McKean (2000)</td>
<td>1. Determine the model. (^1)</td>
<td><strong>R Tools:</strong> lme4 (Bates et al., 2015)</td>
<td>Gimeno et al. (2019)</td>
</tr>
<tr>
<td></td>
<td>Swaninathan, Roger, Horner, et al. (2014)</td>
<td>2. Use the calculator or software tool to run the analysis.</td>
<td>nlme (Pinheiro et al., 2020)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Swaninathan, Roger, Horner (2014)</td>
<td></td>
<td>Scan (Wilbert &amp; Lueke, 2019)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><a href="https://www.dropbox.com/s/dni9qq5pqi3pc23/GLS.R?dl=0">https://www.dropbox.com/s/dni9qq5pqi3pc23/GLS.R?dl=0</a> (Manolov, 2015a)</td>
<td></td>
</tr>
<tr>
<td>HLM (analytical technique)</td>
<td>Moeyaert et al. (2013)</td>
<td>1. Determine the model. (^1)</td>
<td><strong>Online Calculators:</strong> MultiSCED (Cools et al., 2017)</td>
<td>Brosnan, et al. (2016)</td>
</tr>
<tr>
<td></td>
<td>Van de Noortgate (2003)</td>
<td>2. Use the calculator or software tool to run the analysis.</td>
<td><strong>R Tools:</strong> lme4 (Bates et al., 2015)</td>
<td>Harpøth et al. (2019)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>nlme (Pinheiro et al., 2020)</td>
<td></td>
</tr>
</tbody>
</table>

Note. \(^1\)Define the parameters to be included in the model. For example determine if trend, change in trend, autocorrelation, etc. will be included in the model or not.
Log Response Ratio and Related Quantification Techniques

The log response ratio (LRR; Pustejovsky, 2018) provides a SCED quantification technique in terms of percentage change. Referring to Table 2.4, as part of the calculation for LRR, the denominator of the formula for calculating LRR is the mean level of the baseline phase. As it is impossible to divide by zero, it is impossible to use LRR if the mean level of the baseline phase is zero. The percent of goal obtained (POGO; Ferron et al., 2020) is similar to the log response ratio, but can be used when the baseline level is zero. POGO considers the progress made towards a goal level.
Table 2.4

*Log Response Ratio and Related Quantification Techniques: Quantification Technique, Original Source, Calculation, Free and Open-Source Tools, and Applied Examples*

<table>
<thead>
<tr>
<th>Quantification technique</th>
<th>Original reference</th>
<th>How it is calculated</th>
<th>Free/open source tools</th>
<th>Examples in applied studies</th>
</tr>
</thead>
</table>
| LRR                      | Pustejovsky (2018) | 1. Divide mean level of intervention phase by mean level of baseline phase.  
| POGO                     | Ferron et al. (2020) | For desired behavior increase:  
                           |                    | 1. Subtract the expected level of behavior without the intervention from the obtained level of behavior.  
                           |                    | 2. Divide the outcome from step 1 by the goal level of behavior minus the expected level of behavior without the intervention.  
                           |                    | 3. Multiply by 100. | | |
|                          |                    | For desired behavior decrease:  
                           |                    | 1. Subtract the obtained level of behavior from the expected level of behavior without the intervention.  
                           |                    | 2. Divide the outcome from step 1 by expected level of behavior without the intervention minus the goal level of behavior.  
                           |                    | 3. Multiply by 100. | | |
Considerations for Picking an Appropriate Quantification Technique

Quantification techniques must be carefully considered, as certain quantification techniques are more appropriate than others depending on several SCED factors (Manolov & Moeyaert, 2017b). For example, using quantification techniques that control for baseline trend when there is no baseline trend present in the data set can lead to erroneous conclusions (Brossart et al., 2018). Applying baseline trend correction when unnecessary can skew the obtained results by making them greater or smaller than the true estimate. However, few researchers actually justify their choice of SCED quantification techniques; thus, when these quantification techniques are used in published research the reader can be left questioning the appropriateness of the technique used (Solomon, 2014). Manolov and others (2021) recommend for researchers to include the quantification technique and justification of the quantification technique as part of the pre-registered research protocol. This would require researchers to think critically about the quantification techniques they use, and would prevent researchers from reporting the largest quantification of an effect (i.e., selective reporting bias). Fingerhut and others (2020) also found that across all quantification techniques, 75% of the SCED articles published in 2019 included appropriate quantification technique justifications (i.e., rationales that aligned with rationales provided by the quantification technique founders). This indicates that the remaining 25% of articles that provided quantification technique justification provided inappropriate justifications, and thus may not have been used appropriately. It is possible that researchers are unaware of the

---

5 Parker et al. (2011a) only used baseline correction in their data sets when trend was greater than .40 in Phase A, while Tarlow (2017) recommends only using trend correction if the trend is statistically significant. This indicates another issue: there is inconsistency concerning when to use a baseline adjusted technique or not.

6 For example, if the researcher provided the justification “baseline corrected Tau-U was used to control for trend”, this justification would match with the quantification technique founder justification: “baseline corrected Tau-U is recommended to single-case investigators as a flexible and superior alternative to Tau-U when autocorrelation or baseline trend may be present in their time-series data” (Tarlow, 2017, pg. 462). Both justifications mention using baseline corrected Tau-U when baseline trend is present.
need to consider characteristics of the data when picking a quantification technique, or simply that researchers consistently use the same quantification technique because it is a tradition to do so in their field.

In the next section, different aspects that can be considered when choosing a SCED quantification technique are discussed. The components for consideration are based upon recommendations made by previous researchers (e.g., Manolov & Moeyaert, 2017b; Manolov et al., 2021). The components are organized into three categories, which are based upon dimensions identified by Manolov et al. (2021). The three categories discussed are: research question(s), data characteristics, and quantification technique desired features.

**Research Question(s)**

Different quantification techniques are able to answer different research questions, and thus different quantification techniques might be preferable over others depending on the research question (Manolov & Moeyaert, 2017b; Manolov et al., 2021). Different research questions that may possibly be used are discussed in this section.

**Non-Overlapping Data.** A research question might be about the amount of non-overlapping data (i.e., “What is the percentage of non-overlap between study phases?”), in which case non-overlap indices (e.g., IRD; Parker et al., 2009) would be appropriate for use. An example of such a research question can be seen in the study by Miles et al. (2019), where Tau-U Trend A is used to test for the percentage of non-overlap between the phases.

However, the formula for some non-overlap indices are complex, making them more difficult to interpret because they do not truly reflect the percentage of non-overlap. For example, the Tau indices consider both the amount of non-overlapping data and overlapping data together, whereas other non-overlap indices simply reflect the amount of overlapping or non-overlapping
data separately (see: Table 2.1 for complete formulas). For example, PND is simply the percentage of non-overlapping data, and thus can be interpreted as so. Although Tau indices are commonly reported as “non-overlap indices” (e.g., Curran & Van Horne, 2019), in reality these indices do not solely reflect the amount of non-overlapping data and researchers should be careful to interpret estimates in this manner.

**Magnitude of Change in Level.** One common research question concerns the magnitude of the change in level between baseline and intervention phase, sometimes framed simply as “change in level”. Quantification techniques that can estimate the magnitude of change in level between baseline and intervention phase include regression-based techniques (e.g., GLS; Swaminathan, Rogers, & Horner, 2014) and standardized mean difference (Hedges et al., 2012, 2013; Shadish et al., 2014a). Non-overlap indices are sometimes inappropriately used to quantify the magnitude of change in level (also called “effect size”) between baseline and intervention phase (e.g., Carr et al., 2015; Sucuoglu & Demir, 2012). However, these quantification techniques truly reflect the amount of overlapping or non-overlapping data points and not the magnitude of the change in level. Thus, they are not advised for use to answer research questions regarding the magnitude of the change in level.

**Magnitude of Change in Trend.** Some quantification techniques that can reflect a magnitude of change in trend include MPD (Manolov & Solanas, 2013) and regression (Maggin et al., 2011; Moeyaert et al., 2014; Swaminathan, Rogers, & Horner, 2014). The study by Brosnan et al. (2016) is one such study that used HLM to examine the change in trend between baseline and intervention phase.

**Magnitude of Change in Variability.** Change in variability can be defined as the difference in within-case variance between two different phases (e.g., baseline and treatment
If the research question considers the magnitude of change in variability (e.g., Winkens et al., 2014), researchers can use a regression-based method or technique, such as GLS or HLM (e.g., Baek et al., 2014; Swaminathan, Rogers, & Horner, 2014; Van den Noortgate & Onghena, 2007), respectively. GLS and HLM are both able to quantify the data variability in the baseline and the treatment separately; although they cannot provide an effect size parameter, they can measure the heterogeneity of variance between phases.

**Other Research Questions.** Researchers may also want to consider several of these research questions together (e.g., measure both change in level and change in trend); in these instances a regression-based technique (Maggin et al., 2011 Swaminathan, Rogers, & Horner, 2014) could be used. Still other quantification techniques are useful for answering different research questions. If the research question is about eliminating behaviors or a reduction of the outcome, PZD is a quantification technique that can be used (Scotti et al., 1991). If the researcher is concerned with the progress obtained towards a set goal, POGO is a quantification technique that is appropriate for use (Ferron et al., 2020).

These examples show how different quantification techniques are able to answer different research questions. Certain quantification techniques are unable to perform the calculations needed to answer certain research questions. Thus, researchers need to carefully consider the research question(s) when determining which quantification technique is appropriate for use.

The different quantification techniques and their abilities to account for different research question are displayed in Table 2.5. The categorizations of quantification techniques are based upon information provided by the founders of the quantification technique, unless otherwise specified.
Table 2.5

Research Questions and SCED Quantification Techniques

<table>
<thead>
<tr>
<th>Quantification technique</th>
<th>Overlapping data points?</th>
<th>Magnitude of change in level?</th>
<th>Magnitude of change in slope?</th>
<th>Magnitude of change in variability?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-overlap</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PND (Scruggs et al., 1987)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PAND (Parker et al., 2007)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>IRD (Parker et al., 2009)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PEM (Ma, 2006)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>ECL/ PEM-T (White &amp; Haring, 1980)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PZD (Scotti et al., 1991)</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>NAP (Parker &amp; Vannest, 2009)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Tau-U (Parker et al., 2011a)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Tau-U Trend A (Parker et al., 2011a)</td>
<td>Yes (not a true non-overlap statistic due to complex equation)</td>
<td>No</td>
<td>Yes (Cannot provide effect size parameter, but can measure the heterogeneity of trend between phases)</td>
<td>No</td>
</tr>
<tr>
<td>Baseline corrected Tau-U (Tarlow, 2017)</td>
<td>Yes (not a true non-overlap statistic due to complex equation)</td>
<td>No</td>
<td>Yes (Cannot provide effect size parameter, but can measure the</td>
<td>No</td>
</tr>
<tr>
<td>Quantification technique</td>
<td>Overlapping data points?</td>
<td>Magnitude of change in level?</td>
<td>Magnitude of change in slope?</td>
<td>Magnitude of change in variability?</td>
</tr>
<tr>
<td>--------------------------</td>
<td>--------------------------</td>
<td>-------------------------------</td>
<td>-----------------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>HPS-d (Hedges et al., 2012, 2013)</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes (Cannot provide effect size parameter, but can measure the heterogeneity of variance between phases)</td>
</tr>
<tr>
<td>BC-SMD (Shadish et al., 2014)</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes (Cannot provide effect size parameter, but can measure the heterogeneity of variance between phases)</td>
</tr>
<tr>
<td>MPD (Manolov &amp; Solanas, 2013)</td>
<td>No</td>
<td>Yes (overall difference; Manolov &amp; Moeyaert, 2017b)</td>
<td>Yes (overall difference; Manolov &amp; Moeyaert, 2017b)</td>
<td>Yes (Cannot provide effect size parameter, but can measure the heterogeneity of variance between phases)</td>
</tr>
<tr>
<td>MBLR (Campbell, 2004)</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>GLS (Swaminathan et al., 2014)</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>HLM (Van den Noortgate &amp; Onghena, 2008)</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>LRR (Pustejovsky, 2018)</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>POGO (Ferron et al., 2020)</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>
Across-Case SCED Quantification Techniques. Different quantification techniques are appropriate for analyzing intervention effectiveness across participants or studies (e.g., meta-analysis). Non-overlap indices can be used to determine if an effect is present across cases/studies. However, the majority of these quantification techniques do not have a sampling distribution and are commonly averaged to synthesize the effect across cases or studies. Jamshidi and others (2020) examined how different quantification techniques are used for meta-analyses, and how these quantification techniques are weighted. Their study found PND to be the most popular quantification technique for meta-analyses, which also mirrors findings from other studies (e.g., Maggin et al., 2011). The popularity of PND as a method to summarize findings across studies is concerning due to the known limitations of this particular quantification technique, including its inability to account for trend (Shadish & Rindskopf, 2007) and its overreliance on a single data point, making it unable to account for outliers (Parker & Vannest, 2014). Another concerning finding from Jamshidi et al. (2020) is that that meta-analyses that used non-overlap indices (e.g., Tau, NAP, etc.) reported a simple average rather than a weighted average, and that only a few meta-analyses reported checking statistical assumptions.

HLM technique and BC-SMD are most appropriate for calculating estimates across cases and studies because these quantification techniques can account for within and between-case differences, while other quantification techniques are unable to do so (Shadish et al., 2014). Other quantification techniques (e.g., non-overlap indices, standardized mean difference), while sometimes used for across case and across study analyses (e.g., Hong et al., 2017), do not have this flexibility. This can impact results, and researchers are encouraged to carefully consider the technique used when analyzing outcomes across cases and studies.
This literature review and larger study focuses on analysis at the participant level; however, researchers are encouraged to use HLM technique or the BC-SMD technique when conducting analyses across cases or studies.

**Data Characteristics**

Researchers are advised to consider the data characteristics when picking a quantification technique for use, as certain quantification techniques can better account for certain data characteristics than others. In this section, several different data characteristics are identified and discussed.

**Autocorrelation.** One such data characteristic to be considered is autocorrelation, or serial dependency. Autocorrelation is defined as the serial dependency of the errors in a data set, or how much one data point relates to the data points that come before and after it (Shadish & Sullivan, 2011; Solomon, 2014). Shadish and Sullivan (2011) found the average SCED data set to have an autocorrelation value of 0.20. Thus, it is likely that the average SCED data set contains some amount of autocorrelation. For this reason, it is important for researchers to consider whether the quantification technique picked for use can account for autocorrelation/if the quantification technique is robust to autocorrelation. There are also a variety of correlation structure models, including first-order, higher-order and first-order moving average (see: Joo et al. [2019] for more information regarding autocorrelation model specification).

Some quantification techniques are unable to model autocorrelation, (e.g., PEM [Ma, 2006]; IRD [Parker et al., 2009]), making them likely to be affected by autocorrelation if autocorrelation is present in the data set. Other non-overlap indices cannot account for autocorrelation, but research has demonstrated these quantification techniques to be somewhat robust to autocorrelation. Tarlow (2017) found baseline corrected Tau-U to be relatively robust
to autocorrelation, while Solomon (2014) found autocorrelation to have a large effect for the upper range of Tau-U \( \text{Trend A} \) estimates, and only a modest effect for the lower range of estimates. MPD cannot account for autocorrelation, but is demonstrated to be robust against autocorrelation (Manolov & Solanas, 2013; Solomon, 2014).

Still, other quantification techniques can factor autocorrelation directly into the model, meaning that they can more appropriately control for autocorrelation if it is present. Examples of such quantification techniques include regression-based techniques (e.g., Swaminathan et al., 2014; Van den Noortgate & Onghena, 2008), specifically the GLS method (Maggin et al., 2011). This benefit of regression over other quantification techniques may make it preferable for use when autocorrelation is present in the data. However, several assumptions must be made including that the amount of autocorrelation is the same across all phases (Maggin et al., 2011). Furthermore, the reliability of estimates increases with larger data sets (Maggin et al., 2011); thus, Swaminathan et al. (2010) suggest a minimum of 20 measurement occasions per case with a minimum of five data points per phase.

Due to the complexity of autocorrelation and the difficulty that applied researchers may have in understanding autocorrelation, the developed tool does not include autocorrelation as a data characteristic.

**Baseline Trend.** Baseline trend is another issue that is commonly present in SCEDs. Brossart et al. (2018) found the average SCED data set to have small to medium monotonic trend in the baseline phase and treatment phase (.32 and .43, respectively). Ignoring trend in the baseline means that it is possible the true effect may not be appropriately captured, and estimates may be inaccurate. Consider the hypothetical data set: 1, 2, 3, 4, 5 in the baseline phase, and 6, 7, 8, 9, 10 in the treatment phase. Without considering the baseline phase, the PND estimate is 1.00,
indicating a large treatment effect. However, it is likely that the trend in baseline simply extends into the treatment phase, and there may not be an effect at all. For this reason, it is important to control baseline trend if it is present in the data set.

Certain non-overlap indices are unable to account for trend due to the way they are calculated (e.g., PND [Scruggs et al., 1987]; PEM [Ma, 2006]; Tau-U [Parker et al., 2011a]). Some quantification techniques were created to address the inability of such quantification techniques to account for trend. For example, PEM-T was created to address the inability of PEM to account for baseline trend (White & Haring, 1980), while Tau-U Trend A and baseline corrected Tau-U were created to improve upon Tau-U (Parker et al., 2011a; Tarlow, 2017). However, some of these quantification techniques have other issues related to their control of trend. For example, Tau-U Trend A has conservative trend control and does not effectively control for trend if the trend is too large (Tarlow, 2017). Tarlow (2017) also found that baseline corrected Tau-U does not control for trend well if there are less than five data points in the baseline phase. Brossart et al. (2018) notes the importance of only using certain trend-controlling quantification techniques (e.g., Tau-U Trend A and baseline corrected Tau-U) when trend is present, as failure to do so can lead to skewed estimates.

A related issue is the inaccurate reporting of measures (e.g., Curran & Van Horne, 2019) making it difficult to determine if a quantification technique controlling for trend (e.g., baseline corrected Tau-U) was truly used or not. Many studies do not explicitly indicate whether a trend-control quantification technique was used or not (e.g., Curran & Van Horne, 2019). Adding to the confusion is that there are no set terms for referring to the different Tau indices. The failure to explicitly state which quantification technique has been used can cause misinterpretation of the data because trend-controlling quantification techniques like the Tau indices have different
scales (Brossart et al., 2018; Fingerhut et al., in press-b; Tarlow et al., 2017). Baseline corrected Tau-U has a ceiling effect at .80, while Tau-U $\text{Trend}_A$ can exceed 1.00 (Fingerhut et al., in press-a); if it is not clear which quantification technique was used, this can lead to issues with interpretation since different estimates are considered small, medium, or large effects for the different quantification techniques. Researchers should thus state clearly if a trend-control technique is used or not, which has been noted and advised by previous studies (e.g., Fingerhut et al., in press-b; Tarlow, 2017).

**Outliers.** All SCED quantification techniques are affected by outliers, but certain quantification techniques are more affected than others. PND is criticized for being highly effected by outliers, as it uses only one Phase A data point (Lenz, 2013); if that particular data point is an outlier, this will cause the estimate to be deflated or inflated, depending on the direction of the desired outcome. The inability of PND to account for outliers is a great criticism of the quantification technique; several recently published articles acknowledge this issue with PND and cite using a different quantification technique other than PND in response to this flaw (e.g., Dillman et al., 2019). Some non-overlap SCED quantification techniques were developed in response to this criticism of PND. PAND is one quantification technique that was developed to address the issue of over-relying on a single data point, but this quantification technique has still been criticized for not performing well if outliers are present (Lenz et al., 2013; Solomon, 2014). PZD is another quantification technique that is heavily influenced by outliers (Allison & Gorman, 1993; Campbell, 2004). Some non-overlap indices that are less influenced by outliers include Tau indices, NAP, and IRD (Barton et al., 2019; Maggin et al., 2011; Tarlow, 2017). NAP and Tau formulas consider all data points in the series, and thus are less likely to be overly influenced by outliers. Furthermore, Tarlow (2017) notes that rank order methods, like the one
that is used to calculate Tau indices, do not make assumptions regarding outliers, unlike other regression-based techniques. Furthermore, some non-overlap indices may be preferable over the standardized mean difference if outliers are present because non-overlap indices do not rely on mean, median or mode calculations (Barton et al., 2019; Maggin et al., 2011; Wolfe et al., 2019). Although all quantification techniques are somewhat affected by outliers, certain non-overlap indices such as NAP and Tau indices may be preferable over other quantification techniques if outliers are present, as these quantification techniques are less affected by outliers due to the way these quantification techniques are calculated.

**Variability.** Similarly, large amounts of variability (i.e., the amount that individual data points differ from each other within one individual case) can affect the estimates when certain quantification techniques are used. Most non-overlap indices are affected by large within-case variability because these quantification techniques reflect the amount of data overlap and non-overlap that is present. Data sets with large variability have larger amounts of overlapping data points. Fingerhut et al. (in press-a) found the estimates of Tau indices to be affected by within-case variability relative to the size of the intervention effect, while HLM is relatively unaffected by within-case variability, since within-case variability is estimated and included as part of the model. Thus, if there is large within-case variance, researchers are encouraged not to use a Tau index. However, it is also important to consider that the formulas for quantification techniques like Tau and NAP consider all data points in the series, making these particular non-overlap indices better options than other non-overlap indices if there is within-case variability. More research is needed to determine the systematic influence within-case variance has on SCED quantification techniques.
**Number of Data Points.** The number of data points in the baseline and/or intervention phase can also affect the estimates of different quantification techniques. It is important for researchers to consider the number of data points in the data set, especially since Shadish & Sullivan (2011) found only 55% of data sets to have more than five data points in the baseline phase. PND, PAND, IRD, and PEM are quantification techniques that have been found to have low statistical power for studies with a small number of data points (Brossart et al., 2014). NAP has been found to work well with shorter data sets, but is impacted by the number of data points that overlap across baseline and treatment conditions (Parker et al., 2014, p. 142). Tau indices are also found to be impacted by the number of data points in the baseline and intervention phase. Parker et al. (2011a) found that data sets with longer baseline phases and shorter intervention phases yield smaller Tau-U and Tau-U $\tau_{\text{TrendA}}$ estimates on average, even if the total number of data points is kept constant. This is an issue because researchers may continue to collect data points during the intervention phase in order to increase the effect captured by the estimate, indicating a potential ethical issue that is a result of using Tau indices. Baseline corrected Tau-U is not advised for use if there are less than five data points in the baseline phase (Tarlow, 2017) and so Tarlow (2017) proposes a power table for use to help determine if there are enough data points in the baseline phase to correctly detect the presence of an intervention effect. MPD and GLS have also been found to be affected by the phase length (Manolov & Solanas, 2013); when the intervention phase is longer than the baseline phase, these methods have worse performance in terms of efficiency. The bias-corrected LRR is one quantification technique that can be used with a small number of data points in the baseline and/or treatment phase (Pustejovsky, 2019).
**Type of Data.** Researchers need to consider the type of data that is being measured. If the dependent variable is a count or percentage durations of a behavior, then it is appropriate to use a quantification technique such as LRR or POGO (Ferron et al., 2020; Pustejovsky, 2018). If using regression, a generalized linear mixed model can account for count data (Declercq et al., 2019); Declercq et al. (2019) found that a miss-specified linear mixed model for count data has worse goodness of fit and power compared to a generalized linear mixed model. Other quantification techniques such as standardized mean difference and regression-based quantification techniques can be used for continuous data (Shadish et al., 2014; Swaminathan, Rogers, & Horner, 2014). In a study by Swan et al. (2020), researchers found when simulating count outcome data, the estimated baseline mean was unbiased for the data series that were determined to be stabilized, but baseline variance was underestimated.

The different quantification techniques and their abilities to account for data characteristics are displayed in Table 2.6. Each quantification technique is listed in the first column, and different data characteristics are listed in the first row. The categorizations of quantification techniques are based upon information provided by the quantification technique founders, unless otherwise specified.

---

7 This data characteristic is not included in Table 2.6., as all quantification techniques can account for continuous data, except regression, LRR and POGO, which can also be used with count data.
Table 2.6

Data Characteristics and SCED Quantification Techniques

<table>
<thead>
<tr>
<th>Quantification technique</th>
<th>Can model/robust against autocorrelation?</th>
<th>Account for trend in baseline/robust against trend in baseline?</th>
<th>Not overly affected by outliers/robust against outliers?</th>
<th>Can account for large variability/robust against high variability?</th>
<th>Not overly affected by number of data points in baseline or treatment phase?</th>
</tr>
</thead>
<tbody>
<tr>
<td>PND (Scruggs et al., 1987)</td>
<td>No; cannot model autocorrelation.</td>
<td>No.</td>
<td>No; highly affected.</td>
<td>No; highly affected due to reliance on only the most extreme data point for calculation.</td>
<td>No; sensitive to number of data points in baseline phase (Pustejovsky, 2019).</td>
</tr>
<tr>
<td>PAND (Parker et al., 2007)</td>
<td>No; cannot model autocorrelation.</td>
<td>No.</td>
<td>No; not overly influenced (Maggin et al., 2011), but criticized for not doing well with outliers (Solomon, 2014).</td>
<td>No; affected by within-case variability due to non-overlap formula.</td>
<td>No.</td>
</tr>
<tr>
<td>IRD (Parker et al., 2009)</td>
<td>No; cannot model autocorrelation.</td>
<td>No.</td>
<td>Not heavily influenced by outliers (Maggin et al., 2011).</td>
<td>No; affected by within-case variability due to non-overlap formula.</td>
<td>No; sensitive to number of data points in baseline and treatment phase</td>
</tr>
</tbody>
</table>

Non-overlap
<table>
<thead>
<tr>
<th>Quantification technique</th>
<th>Can model/robust against autocorrelation?</th>
<th>Account for trend in baseline/robust against trend in baseline?</th>
<th>Not overly affected by outliers/robust against outliers?</th>
<th>Can account for large variability/robust against high variability?</th>
<th>Not overly affected by number of data points in baseline or treatment phase?</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEM (Ma, 2006)</td>
<td>No; cannot model autocorrelation.</td>
<td>No.</td>
<td>No (Maggin et al., 2011).</td>
<td>Somewhat; less affected by within-case variability due to formula.</td>
<td>Not recommended if there is heteroscedasticity (Manolov et al., 2011).</td>
</tr>
<tr>
<td>ECL/ PEM-T (White &amp; Haring, 1980)</td>
<td>Yes; cannot model autocorrelation, but relatively robust (Tarlow, 2017).</td>
<td>Yes; unreliable trend control (Brossart et al., 2014). Trend is assumed to be linear and</td>
<td>No (Maggin et al., 2011; Rakap, 2015).</td>
<td>Somewhat; less affected by within-case variability due to formula.</td>
<td>Yes; mostly unaffected by number of data points in baseline or treatment phase (Pustejovsky, 2019).</td>
</tr>
<tr>
<td>Quantification technique</td>
<td>Can model/robust against autocorrelation?</td>
<td>Account for trend in baseline/robust against trend in baseline?</td>
<td>Not overly affected by outliers/robust against outliers?</td>
<td>Can account for large variability/robust against high variability?</td>
<td>Not overly affected by number of data points in baseline or treatment phase?</td>
</tr>
<tr>
<td>--------------------------</td>
<td>------------------------------------------</td>
<td>---------------------------------------------------------------</td>
<td>------------------------------------------------------</td>
<td>---------------------------------------------------------------</td>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>PZD (Scotti et al., 1991)</td>
<td>No; cannot model autocorrelation.</td>
<td>No.</td>
<td>No (Allison &amp; Gorman, 1993).</td>
<td>Somewhat; less affected by within-case variability due to formula.</td>
<td>---</td>
</tr>
<tr>
<td>NAP (Parker &amp; Vannest, 2009)</td>
<td>Yes; cannot model autocorrelation, but relatively robust (Manolov et al., 2011).</td>
<td>No.</td>
<td>Yes; relatively unaffected compared to other non-overlap indices (Barton et al., 2019).</td>
<td>Somewhat; less affected by within-case variability than other non-overlap indices. Unaffected by unequal within-case variability (Manolov et al., 2011).</td>
<td>Yes; can be used with shorter datasets (Parker, Vannest, Brown, et al., 2011). Unaffected by number of data points in baseline or treatment phase (Pustejovsky, 2019).</td>
</tr>
<tr>
<td>Tau-U (Parker et al., 2011a)</td>
<td>No; cannot model autocorrelation.</td>
<td>No.</td>
<td>Yes; relatively unaffected compared to other non-overlap measures (Barton et al., 2019).</td>
<td>Somewhat; less affected by within-case variability than other non-overlap indices. Small / medium</td>
<td>No; datasets with longer baseline phases and shorter treatment phases yield smaller results, on average.</td>
</tr>
<tr>
<td>Quantification technique</td>
<td>Can model/robust against autocorrelation?</td>
<td>Account for trend in baseline/robust against trend in baseline?</td>
<td>Not overly affected by outliers/robust against outliers?</td>
<td>Can account for large variability/robust against high variability?</td>
<td>Not overly affected by number of data points in baseline or treatment phase?</td>
</tr>
<tr>
<td>--------------------------</td>
<td>------------------------------------------</td>
<td>---------------------------------------------------------------</td>
<td>---------------------------------------------------</td>
<td>-------------------------------------------------</td>
<td>-------------------------------------------------------------</td>
</tr>
<tr>
<td>Tau-U Trend A (Parker et al., 2011a)</td>
<td>Autocorrelation has larger effect for upper range of values, and modest below this (Solomon, 2014).</td>
<td>Yes; inflated effects (Tarlow, 2017).</td>
<td>Yes; relatively unaffected compared to other non-overlap indices (Barton et al., 2019).</td>
<td>Somewhat; less affected by within-case variability than other non-overlap indices.</td>
<td>No; data sets with longer baseline phases and shorter treatment phases yielded smaller results, on average (Parker et al., 2011a).</td>
</tr>
</tbody>
</table>
**Quantification technique** | **Can model/robust against autocorrelation?** | **Account for trend in baseline/robust against trend in baseline?** | **Not overly affected by outliers/robust against outliers?** | **Can account for large variability/robust against high variability?** | **Not overly affected by number of data points in baseline or treatment phase?** |
--- | --- | --- | --- | --- | --- |
Baseline corrected Tau-U (Tarlow, 2017) | Yes; cannot model autocorrelation, but relatively robust (Tarlow, 2017). | Yes; can account for baseline trend. Does not perform well if less than 5 data points in baseline phase. | Yes; relatively unaffected compared to other non-overlap indices (Barton et al., 2019; Tarlow, 2017). | Somewhat; less affected by within-case variability than other non-overlap indices. | No; sometimes over-corrects trend when there are less than 5 data points (Tarlow, 2017). |
HPS-d (Hedges et al., 2012, 2013) | Yes. | No. | No (Maggin et al., 2011). | Somewhat; less affected by within-case variability than certain non-overlap measures. | No; power increases as number of measurement occasions increase. |

**Baseline corrected Tau-U (Tarlow, 2017)**

- **Can model/robust against autocorrelation?** Yes; cannot model autocorrelation, but relatively robust (Tarlow, 2017).
- **Account for trend in baseline/robust against trend in baseline?** Yes; can account for baseline trend. Does not perform well if less than 5 data points in baseline phase.
- **Not overly affected by outliers/robust against outliers?** Yes; relatively unaffected compared to other non-overlap indices (Barton et al., 2019; Tarlow, 2017).
- **Can account for large variability/robust against high variability?** Somewhat; less affected by within-case variability than other non-overlap indices.
- **Not overly affected by number of data points in baseline or treatment phase?** No; sometimes over-corrects trend when there are less than 5 data points (Tarlow, 2017).

**HPS-d (Hedges et al., 2012, 2013)**

- **Can model/robust against autocorrelation?** Yes.
- **Account for trend in baseline/robust against trend in baseline?** No.
- **Not overly affected by outliers/robust against outliers?** No (Maggin et al., 2011).
- **Can account for large variability/robust against high variability?** Somewhat; less affected by within-case variability than certain non-overlap measures.
- **Not overly affected by number of data points in baseline or treatment phase?** No; power increases as number of measurement occasions increase.
<table>
<thead>
<tr>
<th>Quantification technique</th>
<th>Can model/robust against autocorrelation?</th>
<th>Account for trend in baseline/robust against trend in baseline?</th>
<th>Not overly affected by outliers/robust against outliers?</th>
<th>Can account for large variability/robust against high variability?</th>
<th>Not overly affected by number of data points in baseline or treatment phase?</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC-SMD (Shadish et al., 2014)</td>
<td>Yes.</td>
<td>Yes.</td>
<td>No (Maggin et al., 2011). Outliers skew outcomes (Wolfe et al., 2019).</td>
<td>Can be modeled into the equation.</td>
<td>Somewhat; less affected than other non-overlap indices.</td>
</tr>
<tr>
<td>MPD (Manolov &amp; Solanas, 2013)</td>
<td>Yes; cannot model autocorrelation, but relatively robust (Tarlow, 2017)</td>
<td>Yes; trend is assumed to be linear and extends into treatment.</td>
<td>No (Maggin et al., 2011). Assumes no outliers (Solomon, 2014).</td>
<td>Somewhat; less affected by within-case variability than certain non-overlap measures.</td>
<td>No; results depend on phase length. When treatment phase longer than baseline phase, worse performance (Manolov &amp; Solanas, 2013).</td>
</tr>
<tr>
<td>MBLR (Campbell, 2004)</td>
<td>No; cannot model autocorrelation (Manolov, Guilera, Solanas, 2017).</td>
<td>No.</td>
<td>No.</td>
<td>Somewhat; less affected by within-case variability than certain non-overlap measures.</td>
<td>---</td>
</tr>
<tr>
<td>Quantification technique</td>
<td>Can model/robust against autocorrelation?</td>
<td>Account for trend in baseline/robust against trend in baseline?</td>
<td>Not overly affected by outliers/robust against trend in baseline?</td>
<td>Can account for large variability/robust against high variability?</td>
<td>Not overly affected by number of data points in baseline or treatment phase?</td>
</tr>
<tr>
<td>--------------------------</td>
<td>------------------------------------------</td>
<td>---------------------------------------------------------------</td>
<td>---------------------------------------------------------------</td>
<td>-------------------------------------------------------------------</td>
<td>------------------------------------------------------------------</td>
</tr>
<tr>
<td>GLS (Swaminathan et al., 2014)</td>
<td>Yes.</td>
<td>Yes; trend modeled more accurately than MPD (Solomon, 2014). Trend is assumed to be linear and extends into treatment.</td>
<td>No (Hagan &amp; Burke, 2007; Tarlow, 2017).</td>
<td>Somewhat; less affected by within-case variability than certain non-overlap measures.</td>
<td>No; recommend minimum of 20 measurement occasions per case and at least 5 observations per phase (Swaminathan et al., 2010). Results depend on phase length; when treatment phase longer than baseline phase, worse performance (Manolov &amp; Solanas, 2013).</td>
</tr>
<tr>
<td>Quantification technique</td>
<td>Can model/robust against autocorrelation?</td>
<td>Account for trend in baseline/robust against trend in baseline?</td>
<td>Not overly affected by outliers/robust against outliers?</td>
<td>Can account for large variability/robust against high variability?</td>
<td>Not overly affected by number of data points in baseline or treatment phase?</td>
</tr>
<tr>
<td>--------------------------</td>
<td>------------------------------------------</td>
<td>---------------------------------------------------------------</td>
<td>----------------------------------------------------------</td>
<td>---------------------------------------------------------------</td>
<td>---------------------------------------------------------------</td>
</tr>
<tr>
<td>HLM (Van den Noortgate &amp; Onghena, 2008)</td>
<td>Yes.</td>
<td>Yes.</td>
<td>No (Hagan &amp; Burke, 2007; Tarlow, 2017).</td>
<td>Somewhat; Less affected than certain non-overlap measures (e.g., Fingerhut et al., in press-a).</td>
<td>No; more reliable estimate with more measurement occasions (Ferron et al., 2014).</td>
</tr>
<tr>
<td>Log Response Ratio</td>
<td></td>
<td></td>
<td>Relatively unaffected by outliers if median is used for calculation instead of mean.</td>
<td>Somewhat; less affected by within-case variability than certain non-overlap measures.</td>
<td>Yes; bias-corrected LRR is insensitive to number of data points (Pustejovsky, 2019).</td>
</tr>
<tr>
<td>LRR (Pustejovsky, 2018)</td>
<td>No; cannot model autocorrelation.</td>
<td>No.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POGO (Ferron et al., 2020)</td>
<td>No; cannot model autocorrelation.</td>
<td>Can account for baseline trend; assumes that trend stabilizes over time.</td>
<td>Relatively unaffected by outliers if median is used for calculation instead of mean.</td>
<td>Somewhat; less affected by within-case variability than certain non-overlap measures.</td>
<td>---</td>
</tr>
</tbody>
</table>

*Note.* --- indicates no available information.
Desired Quantification Technique Features

Researchers may also prefer certain quantification techniques over others for quantification technique-specific reasons. This section will discuss these quantification technique-specific reasons.

Ability to Correctly Detect an Effect. Certain quantification techniques differ in ability to correctly detect an effect (e.g., Type I error), and quantification techniques differ in their biases (i.e., certain quantification techniques are more accurate than other quantification techniques). PND has large floor and ceiling effects (Parker & Vannest, 2009), while PAND is able to discriminate among the lowest 10% of effects, but not against the most successful 20% of interventions (indicating a ceiling effect; Parker & Vannest, 2009). PEM is criticized for having a severe ceiling effect and is described as being the “least precise” (Parker & Vannest, 2009), while Tarlow (2017) found PEM-T to have a severe ceiling effect as well, and that the distribution of estimates are negatively skewed. This means that the majority of estimates are towards 1.00. The distribution of estimates for NAP are also negatively skewed (Petersen-Brown et al., 2012). Tarlow (2017) found the Tau indices to have no floor or ceiling effects, although noting that Tau-U Trend A sometimes exceeds the 1.00 upper limit.

Other quantification techniques appear to yield estimates with lower standard errors (i.e., greater precision) than the non-overlap indices. Tarlow (2017) found MPD correctly estimates when there is no true effect, with no floor or ceiling effects, indicating a low Type I error rate. Regression-based quantification techniques have low Type I error rate and statistical power (e.g., HLM; Fingerhut et al., in press-a). PND, PAND and PEM have low statistical power, especially for studies with few cases (Petersen-Brown et al., 2012). Tau-U is one quantification technique with greater statistical power (Parker et al., 2011b). Ferron et al. (2014) found that when four participants are included in the study, reasonable power can be achieved (.80 or higher). Hedges
et al. (2012) found that power increases for HPS-d as the number of cases and/or measurement occasions increase.

**Standardized or Unstandardized.** Researchers may prefer to use a quantification technique that can express the results in standardized form, which is useful when comparing the results across studies or when conducting meta-analyses that synthesize the data. All SCED quantification techniques (e.g., PAND, SMD, PZD, etc.) can provide a standardized outcome. However some quantification techniques produce an unstandardized outcome, which must then be converted to a standardized outcome (e.g., LRR; Pustejovsky, 2018). Researchers may also prefer for estimates to be in unstandardized terms; unstandardized estimates may be easier to interpret if the scale is well-known and understood within the research field. Some quantification techniques that can provide an unstandardized outcome include POGO (Ferron et al., 2020) and regression (Van den Noortgate & Onghena, 2007).

Type I error, Type II error, power, and standardization are more complex terms that applied researchers may have difficulty understanding, and so while these terms are discussed here, the developed tool does not include these terms.

**Known Sampling Distribution.** Quantification techniques may be preferred for having a known sampling distribution and thus their ability to produce a $p$-value. This may be useful if the researcher wants to determine statistical significance. Certain quantification techniques can calculate a $p$-value from the known sampling distribution (e.g., regression-based approaches), while other quantification techniques can calculate a $p$-value from different non-parametric statistical tests (e.g., KRC statistical test for Tau indices or binomial statistical test for PEM-T).

**Ease of Use.** Researchers may also prefer quantification techniques that are easy to calculate. This may be the case if the researcher is conducting preliminary research, if the
researcher is inexperienced with certain statistical software or if the researcher does not have access to certain statistical software. However, relying solely on the ease of use can lead researchers to use inappropriate quantification techniques. PND is known as the easiest quantification technique to use, as it can easily be calculated in Excel or even with a ruler and pencil (Lenz, 2013). Although PND has been heavily criticized for its previously described flaws, it remains a popular quantification technique for calculating the effect for single-case designs (Jamdhisi et al., 2020), and so the overreliance on this technique presents many issues.

PAND, IRD, PEM, and PZD are other quantification techniques that are fairly easy to calculate by hand (Petersen-Brown et al., 2012; Brossart et al., 2014), while Tau indices are more difficult to calculate by hand, but can be calculated with a website calculator (e.g., http://ktarlow.com/stats/tau/ [Tarlow, 2016]; http://www.singlecaseresearch.org/ [Vannest et al., 2016]). Some quantification techniques may be more prone to errors because of the flexibility of the quantification techniques; for example, regression-based quantification techniques allow the researcher to adjust the model depending on the research questions, and this may lead to confusion as to how to adjust the model appropriately. Furthermore, regression-based quantification techniques cannot be calculated by hand, and researchers are required to make different decisions regarding the models to be used.

The different quantification techniques and their abilities to account for desired quantification technique features can be seen in Table 2.7. The categorizations of quantification techniques are based upon information provided by the quantification technique founders, unless otherwise specified.
Table 2.7

*Desired Quantification Technique Features and SCED Quantification Techniques*

<table>
<thead>
<tr>
<th>Quantification technique</th>
<th>Ability to detect an effect/ Accuracy of quantification technique</th>
<th>Can provide an unstandardized outcome?</th>
<th>p-value?</th>
<th>Easy to use?</th>
</tr>
</thead>
<tbody>
<tr>
<td>PND (Scruggs et al., 1987)</td>
<td>Large floor and ceiling effects (Chen et al., 2016; Parker &amp; Vannest, 2009).</td>
<td>N/A*</td>
<td>Yes.</td>
<td>Yes; commonly thought of as the easiest technique to use.</td>
</tr>
<tr>
<td>PAND (Parker et al., 2007)</td>
<td>Could discriminate among lowest 10% of effects (Parker, Vannest, &amp; Brown, 2011). Doesn’t discriminate as well against most successful 20% of interventions (Chen et al., 2016; Parker &amp; Vannest, 2009).</td>
<td>N/A*</td>
<td>Yes; with Phi.</td>
<td>Can be calculated by hand.</td>
</tr>
<tr>
<td>IRD (Parker et al., 2009)</td>
<td>Large ceiling effects, especially at the 60th and 80th percentile (Chen et al., 2016). Greater discriminability compared to other measures (Chen et al., 2016).</td>
<td>N/A*</td>
<td>Cannot calculate p-value; lacks known sampling distribution.</td>
<td>Can be calculated by hand.</td>
</tr>
<tr>
<td>Quantification technique</td>
<td>Ability to detect an effect/ Accuracy of quantification technique</td>
<td>Can provide an unstandardized outcome?</td>
<td>p-value?</td>
<td>Easy to use?</td>
</tr>
<tr>
<td>--------------------------</td>
<td>------------------------------------------------------------------</td>
<td>---------------------------------------</td>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>PEM (Ma, 2006)</td>
<td>No floor effect, but severe ceiling effect (Chen et al., 2016; Giannakakos &amp; Lanovaz, 2019; Parker &amp; Vannest, 2009). Least precise quantification technique; appears to have lowest power compared to other non-overlap indices (Brossart et al., 2014).</td>
<td>N/A*</td>
<td>Cannot calculate p-value; lacks known sampling distribution.</td>
<td>Can be calculated by hand.</td>
</tr>
<tr>
<td>ECL/ PEM-T (White &amp; Haring, 1980)</td>
<td>Estimates negatively skewed; Severe ceiling effect (Tarlow, 2017). Appears to have the lowest power among non-overlap methods (Brossart et al., 2014).</td>
<td>N/A*</td>
<td>Cannot calculate p-value; lacks known sampling distribution.</td>
<td>Can be calculated by hand.</td>
</tr>
<tr>
<td>PZD (Scotti et al., 1991)</td>
<td>Considered to be more stringent measure of treatment efficacy.</td>
<td>N/A*</td>
<td>Cannot calculate p-value; lacks known sampling distribution.</td>
<td>Can be calculated by hand</td>
</tr>
<tr>
<td>NAP (Parker &amp; Vannest, 2009)</td>
<td>Highly negatively skewed (Petersen-Brown et al., 2012), ceiling effects especially at the 60th and 80th percentile (Chen et al., 2016).</td>
<td>N/A*</td>
<td>Can calculate a p-value (under assumption of independent data). Confidence</td>
<td>With online calculator.</td>
</tr>
<tr>
<td>Quantification technique</td>
<td>Ability to detect an effect/ Accuracy of quantification technique</td>
<td>Can provide an unstandardized outcome?</td>
<td>p-value?</td>
<td>Easy to use?</td>
</tr>
<tr>
<td>--------------------------</td>
<td>---------------------------------------------------------------</td>
<td>--------------------------------------</td>
<td>--------</td>
<td>------------</td>
</tr>
<tr>
<td>Tau-U (Parker et al., 2011a)</td>
<td>Ceiling effects especially at the 60th and 80th percentile (Chen et al., 2016).</td>
<td>N/A*</td>
<td>Can calculate a p-value (under assumption of independent data). Confidence intervals can be obtained.</td>
<td>With online calculator.</td>
</tr>
<tr>
<td>Tau-U Trend A (Parker et al., 2011a)</td>
<td>Can yield results greater than 1, leading to interpretation errors. May interpret the results as being larger than they are, because of this (Fingerhut et al., in press-a). No floor of ceiling effects apparent (Tarlow, 2017).</td>
<td>N/A*</td>
<td>Cannot calculate p-value; lacks known sampling distribution.</td>
<td>With online calculator.</td>
</tr>
<tr>
<td>Baseline corrected Tau-U (Tarlow, 2017)</td>
<td>No floor of ceiling effects apparent (Tarlow, 2017)</td>
<td></td>
<td>Cannot calculate p-value; lacks known sampling distribution.</td>
<td>With online calculator.</td>
</tr>
<tr>
<td>HPS-d (Hedges et al., 2012, 2013)</td>
<td>---</td>
<td>Yes.</td>
<td>Can calculate a p-value and confidence</td>
<td>More difficult to calculate by hand; simpler to calculate</td>
</tr>
<tr>
<td>Quantification technique</td>
<td>Ability to detect an effect/ Accuracy of quantification technique</td>
<td>Can provide an unstandardized outcome?</td>
<td>p-value?</td>
<td>Easy to use?</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---------------------------------------------------------------</td>
<td>----------------------------------------</td>
<td>----------</td>
<td>-------------</td>
</tr>
<tr>
<td>BC-SMD (Shadish et al., 2014)</td>
<td>---</td>
<td>Yes.</td>
<td>Can calculate a p-value and confidence intervals can be obtained.</td>
<td>More difficult to calculate by hand. Can use website or software.</td>
</tr>
<tr>
<td>MPD (Manolov &amp; Solanas, 2013)</td>
<td>Effects positively skewed (Tarlow, 2017). Correctly estimated when there was no effect. No floor or ceiling effects (Tarlow, 2017).</td>
<td>Yes.</td>
<td>Cannot calculate p-value; lacks known sampling distribution.</td>
<td>Can be calculated by hand.</td>
</tr>
<tr>
<td>MBLR (Campbell, 2004)</td>
<td>---</td>
<td>Yes.</td>
<td>Cannot calculate p-value; lacks known sampling distribution.</td>
<td>Can be calculated by hand.</td>
</tr>
<tr>
<td>GLS (Swaminathan et al., 2014)</td>
<td>Somewhat sensitive (Solomon et al., 2015).</td>
<td>Yes.</td>
<td>Can calculate a p-value and confidence intervals can be obtained.</td>
<td>Website or software needed.</td>
</tr>
<tr>
<td>Quantification technique</td>
<td>Ability to detect an effect/ Accuracy of quantification technique</td>
<td>Can provide an unstandardized outcome?</td>
<td>p-value?</td>
<td>Easy to use?</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>-----------------------------------------------------------------</td>
<td>--------------------------------------</td>
<td>---------------------------------------------------------------------------</td>
<td>--------------------------------------</td>
</tr>
<tr>
<td>HLM (Van den Noortgate et al., 2008)</td>
<td>---</td>
<td>Yes.</td>
<td>Can calculate a p-value and confidence intervals can be obtained.</td>
<td>Website or software needed.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Log Response Ratio</td>
<td></td>
</tr>
<tr>
<td>LRR (Pustejovsky, 2018)</td>
<td>---</td>
<td>Yes.</td>
<td>Cannot calculate p-value; lacks known sampling distribution.</td>
<td>Can be calculated by hand.</td>
</tr>
<tr>
<td>POGO (Ferron et al., 2020)</td>
<td>---</td>
<td>Yes.</td>
<td>Cannot calculate p-value; lacks known sampling distribution.</td>
<td>Can be calculated by hand.</td>
</tr>
</tbody>
</table>

*Note.* --- indicate no available information. * indicates quantification technique produces an outcome in the original raw scale, which is a standardized scale.
Present Study

Much consideration needs to be put into picking an appropriate quantification technique for use with SCEDs. The complexity and difficulty of considering the different characteristics present in a SCED may be one reason why inappropriate quantification techniques are used. Fingerhut et al. (2020) found that quantification techniques are sometimes used for general reasons such as “to use an effect size measure” or because the quantification technique is easy to use. It is possible that the heavy reliance on visual analysis (noted in the beginning of this literature review) occurs because of the difficulty in picking an appropriate quantification technique for use. It may be overwhelming for researchers to consider all of the different data characteristics that are presented in the SCED data and graphs.

There are many tools and guidelines that have been developed to assist SCED researchers in a variety of ways. Many online calculators, R codes, and other applications have been developed to assist in calculations; these calculators are referenced in Tables 2.1-2.4 (e.g., Baseline Corrected Tau Calculator, Tarlow, 2016; Single Case Research calculator, Vannest et al., 2016). Other researchers have focused on improving the way that the quality and outcomes of SCEDs are assessed (e.g., Single-Case Analysis and Review Framework; Ledford et al., 2016) and reported (Single-Case Reporting Guideline in Behavioral Interventions; Tate et al., 2016). However, researchers may still benefit from using a tool that helps them systematically determine the best quantification technique to use depending on the research questions(s), SCED graph and/or SCED data. When left to determine the most appropriate quantification technique without support, this may be difficult for researchers to do due to the numerous considerations that must be made (see Tables 2.5-2.7). Thus, an Excel tool is presented in the next section of this paper to help address the issue of researchers picking an inappropriate SCED quantification
technique for use, as well as justifying the use of the quantification technique and calculating the estimate of the quantification technique.

The research questions and hypotheses for this study were as follows:

1. *Given an AB graph/data set and research question(s), is a newly developed tool for single-case quantification technique selection effective in assisting single-case researchers to choose an appropriate quantification technique to use?* I hypothesized that a tool for single-case quantification technique selection would cause a statistically significant positive change in single-case researcher’s ability to select an appropriate quantification technique.

2. *Given an AB graph/data set and research question(s), is a newly developed tool for single-case quantification technique selection effective in assisting single-case researchers to provide an appropriate justifications for using a quantification technique?* I hypothesized that a tool for single-case quantification technique selection would cause a statistically significant positive change in single-case researcher’s ability to provide an appropriate justifications for using a quantification technique.

3. *Given an AB graph/data set and a research question, when using a newly developed tool for single-case quantification technique selection, will single-case researchers be able to perform the calculation for the quantification technique that they determine is most appropriate?* I hypothesized that a tool for single-case quantification technique selection would cause a statistically significant positive change in single-case researcher’s ability to perform the calculation for the quantification technique that they determine is most appropriate.

4. *Will single-case researchers find a newly developed tool for single-case quantification technique selection to be socially valid?* I predicted that the tool for single-case
quantification technique selection would have high social validity (i.e., the majority of single-case researchers will report the tool to be at least “somewhat useful” [a rating of 3 or higher on a 4-point Likert scale]).
Chapter 3: Method

The development of a tool and the method for testing the effectiveness of the tool is outlined in this section. The first section, “Part 1: Development of Tool” outlines procedures followed to develop the tool to help users choose appropriate SCED quantification techniques, provide a justification for using the quantification techniques, and calculate an estimate correctly. The second section, “Part 2: Using the Tool as an Intervention” describes the procedures for testing the effectiveness of the tool.

Part 1: Development of Tool

The different components of the tool and how they were developed are described in this section. First, in one Excel document labeled Excel Tool, four sheets were created: Sheet 1 (Instructions), Sheet 2 (Tool), Sheet 3 (Metric Details), and Sheet 4 (Notes and References). Thus, the Excel Tool contained four parts, consisting of the four different sheets. Instructions were added to Sheet 1 (Instructions). Sheet 2 (Tool) included an input and output (i.e., the ranked quantification techniques), the development of which is explained in detail in the sub-section “Creating the Tool: Sheet 2 (Tool)”. Related information about each quantification technique was added to Sheet 3 (Metric Details). Sheet 4 (Notes and References) contained notes about the assumptions for using the tool, as well as references.

A second Excel document labeled Examples was also created. Example graphs/data sets and research questions were added to this document; the development of these graphs/data sets and research questions are described in detail in the sub-section “Creating the Examples Document”. The Excel Tool and Examples documents were developed to be used together as one intervention, and are referred to simply as “Excel tool” or “tool”.

69
Creating the Tool: Sheet 1 (Instructions)

Instructions for how to use the Excel Tool were added to the first sheet. Instructions explained to the user how to navigate the Excel Tool and use the three sheets, as well as how to use the Examples document. Instructions were provided in written form, as well as in video form.

Creating the Tool: Sheet 2 (Tool)

In order to create the input and output function for Sheet 2 (Tool) of the Excel Tool, the necessary information was first transferred into a table in a separate Excel file. The table in this separate document was a 16 x 9 matrix, with names of SCED quantification techniques in the sixteen rows and different components that may influence SCED quantification technique appropriateness in the nine columns. The sixteen different quantification techniques and the nine different components that were identified to be included in the tool were based upon previous research about SCED quantification techniques (e.g., Manolov & Moeyaert, 2017b; Manolov et al., 2021), and were the same quantification techniques and components that are displayed in Tables 2.5-2.7 of “Chapter 2: Literature Review”, with a few differences. The quantification technique BC-SMD and the analytical technique HLM were not included. These are used for between-case estimates and the focus of the current study is to evaluate the researchers’ ability to choose the most appropriate quantification technique for one case rather than across-cases. Furthermore, the components “autocorrelation”, “data type”, “standardized or unstandardized”, and “ability to accurately detect an effect” were not included. These components were not included because they are beyond the skill set of an applied researcher. These are more complex components, and including these may have ultimately skewed the data collected. For example, if the user did not know what autocorrelation is, this may have confused the user and affected the validity of the results from this study. These four components may be coded and added back into
a later improved tool pending results from this current study (i.e., positive feedback indicating that all included components were understood). Although these components were not included as part of the input and output function, information related to these components was still included in Sheet 3 (Metric Details), which is described in the next section.

Next, using the information provided within each Excel cell of the document, each quantification technique was coded as being able to address each of the nine components, somewhat addressing the component, not addressing the component, or no information available. For example, considering the component baseline trend, a quantification technique that “addresses the component” is a quantification technique that “can model/account for trend in baseline phase/the outcome is robust against baseline trend”. A quantification technique that “somewhat addresses the component” is a quantification technique that “can account for baseline trend, but has unreliable trend control”. A quantification technique that “does not address the characteristic” is a quantification technique that “cannot account for baseline trend/ the outcome is not robust against baseline trend”. Considering baseline corrected Tau-U, Tarlow (2017) found that baseline corrected Tau-U can account for baseline trend. Thus, for the component baseline trend, baseline corrected Tau-U was coded as “can address the component”.

This coded information was then used to create a logic code in Excel. The goal of Sheet 2 (Tool) of the Excel Tool was to allow the user to input “yes” to consider one or more of the nine components when determining the most appropriate quantification technique for use, and input “no” for the components that did not need to be accounted for. For example, if the component was “outliers” (i.e., “Are outliers present?”), the user could input “yes” if the user wanted to find a quantification technique that could account for outliers/is robust against outliers and “no” if the user did not need a quantification technique that accounted for outliers/is robust against outliers.
Thus, each quantification technique with each component was given a value between 1 and 20 for two different scenarios: When the component needed to be considered (user input = yes), or when the component did not need to be considered (user input = no). When the component should be considered, quantification techniques that could address the characteristic were coded as “5” for data characteristics/desired features, and “20” for if they could answer the research question. Research questions were weighted more heavily than data characteristics/desired features because it is impossible to use a quantification technique at all if it cannot answer the research question. Quantification techniques that somewhat addressed the data characteristics/desired feature were coded as “3”, and coded as “15” if they could somewhat answer the research question. Quantification techniques that did not address the data characteristics/desired feature and did not answer the research question, and quantification techniques where no information could be found, were coded as “1”. Considering baseline corrected Tau-U and a research question related to magnitude of change in trend, baseline corrected Tau-U was coded as “1”.

If the user input “no”, meaning that the user did not need a quantification technique that accounted for that component, the quantification technique and component was coded as “5”. The only condition where the cell was not coded as “5” was if “no” was inputted for the component “baseline trend” (i.e., “Can the quantification technique account for trend in baseline?”). Some quantification techniques that are baseline trend-specific do not work well if there is no baseline trend (e.g., baseline corrected Tau-U), and so these quantification techniques were coded as “1” for this component instead of being coded as “5”.

The complete logic code is displayed in Appendix 3.1.
A total score was calculated from the logic code, depending on what components were marked “yes” in the input (i.e., which components needed to be considered). The total score from the logic code was used to sort the list of quantification techniques so that the “most appropriate” quantification technique(s) were presented first (labeled I under Ranked Metrics in Figure 3.3). The information coded from Figure 3.1 was also displayed in the output. The input and output based upon the logic code is displayed in Figure 3.3, and this input and output was Sheet 2 (Tool) of the *Excel Tool* document.

**Figure 3.3**

*Input and Output Based on Logic Code*

<table>
<thead>
<tr>
<th>Input</th>
<th>Ranked Metrics (Output)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Possile research questions, data/characteristics, and overall metrics features</td>
<td>Star input</td>
</tr>
<tr>
<td>Response or no response to items in list</td>
<td>Improvement rate difference (PRI)</td>
</tr>
<tr>
<td>No</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td></td>
</tr>
</tbody>
</table>

**Creating the Tool: Sheet 3 (Metric Details)**

Information from Tables 2.1-2.4 was added to this sheet. This sheet included the original source of the quantification technique (including the founders of the quantification technique), information about free and open source online calculators, R-codes, or other sources that could be used to calculate the quantification technique, citations of examples of the quantification
technique used in the literature, and other information about the quantification technique that was not included elsewhere in Excel Tool.

**Creating the Tool: Sheet 4 (Notes and References)**

This sheet included assumptions for using the tool. It also explained the logic code used for Sheet 2 (Tool), and provided definitions for the data/graph characteristics. This sheet also included a complete list of references cited throughout the tool.

Screenshots of each of the four sheets are displayed in Appendix 3.2. The tool can also be viewed and downloaded at [https://osf.io/7usbj/](https://osf.io/7usbj/).

**Creating the Examples Document**

Written and video instructions were added to the top of the Examples document. Next, five AB graphs/data sets were generated using RStudio (RStudio Team, 2015) to be included in Examples. AB graphs were chosen for this study because these are the “building blocks” of more complex SCEDs (e.g., multiple-baseline designs, withdrawal designs, etc.). SCED quantification techniques that are used with AB designs can be extended for use with these more complex designs.

**Graph Components: Research Question(s).** Four different research questions could be displayed alongside each graph/data set; these research questions were based upon the research questions that can be seen in Table 2.5 of “Chapter 2: Literature Review”. These research questions were: a.) “What is the magnitude of the change in level?” b.) “What is the magnitude of the change in linear trend?” c.) “What is the magnitude of the change in variability” and d.) “What is the amount of non-overlapping data points?”

---

8 The possible SCED research questions are not limited to these four, but were limited to four for the purpose of this study.
**Graph Components: Data Characteristic(s).** Four different data characteristics could be manipulated and displayed in each graph/data set; these are the same data characteristics that are displayed in Table 2.6 of “Chapter 2: Literature Review”. These data characteristics were: a.) less than five data points in the baseline or treatment phase, b.) trend in baseline, c.) outliers, and d.) within-case variability. Autocorrelation was not considered when manipulating the graphs and data sets, as this is a more complex characteristic to measure and can be difficult for applied researchers to understand. Furthermore, it can be hard to display graphically (Matyas & Greenwood, 1990).

**Graph Components: Quantification Technique Desired Features(s).** One desired feature could be indicated as needed for each graph/data set. This desired quantification technique feature was a $p$-value to test for statistical significance, and was based upon the desired quantification technique features that can be seen in Table 2.7 of “Chapter 2: Literature Review”. The other desired quantification technique features (ability to detect an effect, provide an unstandardized outcome, and ease of use) were not included, as they are more difficult to express via graph, data set, or research question and/or are beyond the scope of understanding for an applied researcher.

**Example Answers.** On Sheet 2 of Excel Tool, underneath each example graph/data set and research question, example answers were created for the user to reference. The research question(s) and data characteristic(s) components were identified and recorded onto this sheet. Using the Excel Tool, the quantification techniques that were recommended as most appropriate for use were also recorded. Furthermore, example justifications for using the quantification technique...
technique and quantification technique calculations were created and provided. These justifications were created by the author of this study, and were written with information provided in Sheet 2 and Sheet 3 of the Excel Tool. Each constructed justification provided at least two reasons for using the quantification technique, as well as the calculation. An example of a justification that was created can be seen below:

The LRR is used because it can provide an unstandardized outcome, which may be easier for practitioners to interpret outcomes. LRR can express the magnitude of the effect; it also is related to the percentage change, which is often used and understood by practitioners (Pustejovsky, 2018). LRR was calculated using the SingleCaseES package. There is a 47% increase in scores.

A screenshot of the Examples document is displayed in Appendix 3.3. This document can also be viewed and downloaded at https://osf.io/7usbj/.

Application of Tool with Graph/Data Set

An example graph/data set and research question is displayed in Figure 3.4. The two components that are present in this graph/data set are a.) research question: magnitude of change in level and b.) data characteristic: less than five data points in the baseline phase or treatment phase. Thus, using the input and output (i.e., Sheet 2 of the Excel Tool) seen in Figure 3.3., all inputs would be set to “no” except for “magnitude of change in level” and “less than 5 data points in the baseline phase or treatment phase”. The input would be set to “yes” for these two components. The output shows that the recommended quantification technique to use is LRR. An example of a justification for using LRR technique that uses information provided from the tool could be as follows:

The LRR is used because it can provide an unstandardized outcome, which may be easier for practitioners to interpret outcomes. LRR can express the magnitude of the effect; it also is related to the percentage change, which is often used and understood by practitioners (Pustejovsky, 2018). LRR was calculated using the SingleCaseES package. There is a 95% increase in scores.
Research Question: What is the magnitude of change in level between A and B phase?

The development of the tool as outlined in this section signifies the first product of this study.

Part 2: Using the Tool as an Intervention

The next section describes how the effectiveness and the social validity of the tool was tested. The specific research questions to be answered, which were introduced in Chapter 2: Literature Review, were as follows: a.) Given an AB graph/data set and research question(s), is a newly developed tool for single-case quantification technique selection effective in assisting single-case researchers to choose an appropriate quantification technique to use? b.) Given an
AB graph/data set and research question(s), is a newly developed tool for single-case quantification technique selection effective in assisting single-case researchers to provide an appropriate justifications for using a quantification technique? c.) Given an AB graph/data set and a research question, when using a newly developed tool for single-case quantification technique selection, will single-case researchers be able to perform the calculation for the quantification technique that they determine is most appropriate? d.) Will single-case researchers find a newly developed tool for single-case quantification technique selection to be socially valid? The procedures outlined in the next sections aimed to answer these four research questions.

Institutional Review Board approval for this study was received from University at Albany, State University of New York.

Participants

Anyone who conducted SCED research was eligible to participate in this study. This was defined as someone who had interpreted the results of a SCED graph either with support (e.g., with help from an instructor or colleague) or independently within the last year. The participant needed to have conducted SCED analysis within the last year to ensure that those who participated in the study had prior knowledge regarding understanding SCED graphs and their features, and to help ensure that the participant was still an active SCED researcher. The participant also needed to be interested in quantifying the effect from a SCED graph; if the participant did not have an interest in quantifying the effect, the participant would likely continue to use visual analysis and not find the tool useful. Thus, two screening questions were used to determine if someone was eligible for the study: a.) Have you quantitatively analyzed a single-case graph either with support (e.g., with help from an instructor or colleague) or independently
within the last year? b.) Are you interested in quantifying the effect from a SCED graph? These screening questions were intended to include a range of participants, both those who were novices in quantifying the results from SCEDs (e.g., applied researchers who had limited experience quantifying results from a SCED, as well as graduate students) and those who were experts in quantifying the results from SCEDs (e.g., SCED methodologists). If any participant answered “no” to either screening question, the participant could still participate in the study, but their data was not included for the analyses.

Snowball sampling was used to recruit participants; potential participants were provided information about the study, and then asked to provide other potential participants with information about the study. Snowball sampling was used because the eligibility requirements for the study were specific, and so it might have been difficult to find potential participants. Initial participants were intended to be SCED researchers who had published SCED related articles. These researchers were found in several ways. First, potential researchers were found by searching “single-case” OR “single case” OR “single-subject” OR “single subject” on Web of Science starting from the year 2019, and sorting by Author/Creator. The top 10 authors related to these terms were emailed and asked to share the study. The study was also shared via a school psychology listserv. Other participants were recruited from the email list for participants who attended the Institute of Education Sciences Advanced Training Institute on single-case research methods. Lastly, the author of this dissertation shared the study with present and past colleagues. All researchers were able to participate in the study and/or send information to other potential participants, who could in turn share the information with potential participants, as well. The email script used to share the study with others is displayed in Appendix 3.4
A total of forty-nine participants started the study but did not finish, and a total of twenty-nine participants ultimately completed the study. A total of two participants indicated Bachelors as their highest degree obtained, sixteen participants indicated having a Master’s degree, and eleven participants indicated having a Ph.D. Participants identified themselves as being a part of a range of academic departments, including school psychology, educational psychology, and special education. The most popular academic field/department indicated by participants was special education, with a total of nine participants identifying with this field/department, followed by educational psychology ($n = 7$) and school psychology ($n = 6$).

A total of three participants were very confident in their ability to choose an appropriate quantification technique at the beginning of the study. These three participants had different education levels and were from different academic departments. A total of eighteen participants were somewhat confident, six participants were somewhat unconfident, and two participants were not at all confident. Similarly, two participants were very confident in their ability to justify using a quantification technique at the beginning of the study, seventeen participants were somewhat confident, eight participants were somewhat unconfident, and two participants were not at all confident. These two participants who were not at all confident were the same two participants who indicated they were not at all confident in their ability to choose an appropriate quantification technique. A total of ten participants had previously published SCED research in the past, fifteen participants had not yet published SCED research, but would like to in the future, and four participants had not and did not intend to publish SCED research.

One participant indicated not being interested in quantifying effects from a SCED. Three of the participants indicated that they had not quantitatively analyzed SCED graphs with support or independently in the past year. These four participants were not included for inferential
analysis due to their response to these questions, which were part of the inclusion criteria. Thus, a total of twenty-five participants were eligible for inferential analyses.

**Materials**

**Excel Tool.** The *Excel Tool* included the four sheets (Instructions, Tool, Metric Details, and Notes and References).

**Examples.** This separate document contained the five generated example AB graphs with data sets and research questions. It also contained sample quantification technique justifications for each example graphs/data sets and research questions, depending on the recommended quantification technique(s). This sheet also included a space for the user to write their own justification and calculation so that they could compare their response with the example justifications to determine if they were using the *Excel Tool* as intended.

**Pre/Posttest AB Graphs/Data Sets and Research Questions.** Five pre/posttest AB graphs/data sets and research questions were created to help determine the effectiveness of the intervention. These five graphs/data sets and research questions were created similarly to the graphs/data sets and research questions in *Examples*, but with a few restrictions.

The data for these graphs/data sets were randomly manipulated so that each of the five pre/posttest graph and data sets demonstrated a total of two components: one components related to data characteristic(s)/desired quantification technique feature and one related to the research question(s). All of the graphs that were generated were manipulated at random (different combinations could have been manipulated, which would have resulted in different graphs and data sets) and so would have the same level of difficulty. All graphs were created so that the size of the effect was held constant; the change in mean between the A and B phase for all graphs was between 4.5 and 5.5. An example of one pre/posttest graphs/data sets paired with the
research question, along with the two essential components that would need to be considered when picking an appropriate quantification technique, is displayed in Figure 3.4. The five AB graphs and research questions are displayed in Appendix 3.5.

**Measures**

A demographic survey and social validity survey were developed to help evaluate the effectiveness of the tool. The demographic survey (i.e., participant characteristics survey), which is displayed in Appendix 3.6, was used to collect descriptive data about the participants, which could be used to determine if there were differences in pretest and posttest scores depending on academic department or education level. The social validity survey, which is displayed in Appendix 3.7, was necessary to answer research question 4, regarding the perceived usefulness of the tool.

**Rubric with Indicators.** A rubric with three indicators was developed to objectively measure the appropriateness of the quantification technique chosen, the appropriateness of the justification provided, and the accuracy of the calculation. This rubric was used by the author of this study to grade participant responses. The rubric was used for both pretest and posttest.

The rubric that was developed to evaluate the user’s ability to choose an appropriate quantification technique, provide a justification for using the quantification technique, and perform the calculation accurately is displayed in Table 3.1. The rubric contained three indicators: 1. choice of quantification technique, 2. justification of quantification technique, and 3. calculation of quantification technique. These indicators and each of the levels are defined below.
### Table 3.1

**Rubric with Indicators**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Choice of quantification technique</strong></td>
<td>Quantification technique accounts for no essential graph elements</td>
<td>Quantification technique partially accounts for 1 essential graph element</td>
<td>Quantification technique accounts for 1 essential graph element</td>
<td>Quantification technique partially accounts for 2 essential graph elements</td>
<td>Quantification technique partially accounts for 1 essential graph element and accounts for 1 essential graph element</td>
<td>Quantification technique accounts for 2 essential graph elements</td>
</tr>
<tr>
<td><strong>Justification of quantification technique</strong></td>
<td>Justification includes no elements</td>
<td>Justification includes 1 element</td>
<td>Justification includes 2 elements</td>
<td>Justification includes 3 or more elements</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Calculation of quantification technique</strong></td>
<td>Calculation not attempted</td>
<td>Calculation attempted, with mistakes</td>
<td>Estimate is correctly calculated</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Each of the five pre/posttest AB graphs/data sets contained two components related to the indicators “choice of quantification technique”. For example, referring to Figure 3.4, the following are the two essential components for the indicator “choice of quantification technique”:

- Research question: Magnitude of change in level
- Data characteristic: Less than five data points in baseline phase or treatment phase

As discussed previously, each of the five pre/posttest graphs/data sets had two essential elements.

The elements used to evaluate the justification of the quantification technique are found below. These elements are founded upon the dimensions and facets identified by Manolov et al. (2021).

- Research question
- Quantification
- Number of observations
- Autocorrelation
- Outliers
- Missing data
- Baseline trend
- Variability
- Statistical properties
- Sampling distribution

Together, the components and elements for “choice of quantification technique” and “justification of the quantification technique” were used with the rubric in order to rate participants responses objectively.

The last indicator “calculation of the quantification technique” was used to evaluate the accuracy of the quantification technique calculation, if one was provided.
**Applied Example.** Referring again to Figure 3.4., below is an example response and how it would be graded with the rubric.

MPD can be used for this graph and data set. MPD is a quantification technique that can estimate the magnitude of change in level between the baseline and treatment phase. The MPD estimate is 30.12.

This response would be given a score of 5 for the indicator “choice of quantification technique”, as MPD partially accounts for one of the essential elements (the research question) and fully accounts for the other essential element (the unique data characteristic). MPD can answer research questions related to the magnitude of change in level. MPD can be used when there are less than five data points in the baseline phase, but results are dependent on phase length, especially when the treatment phase is longer than the baseline phase (Manolov & Solanas, 2013).

For “justification of the quantification technique”, this response would be given a score of 2 since it includes only one essential element in the justification (related to “research question”).

The calculation is incorrect, so for the indicator “calculation of quantification technique”, this response receives a score of 2.

**Procedures**

Before beginning the study, five experts in the field validated the tool. These experts were picked by searching “single-case” OR “single case” OR “single-subject” OR “single subject” on Web of Science starting from the year 2019, and sorting by Author/Creator. Five experts were picked from the top five authors shown from the search results. They were emailed and asked to give any feedback regarding the tool so that any necessary changes could be made before the study was sent to participants. One of the experts reported they were unable to provide
the feedback at the time of the validation process, so a sixth expert was identified through the search terms. This new sixth expert had already reviewed and interacted with the tool prior to the official validation process. The expert who had previously opted out of participating was later able to validate the tool, but was not able to submit the feedback prior to the start of the study. A total of five field experts provided feedback prior to the start of the study, and six experts submitted feedback altogether.

Four of the validators delivered feedback for the tool via writing (i.e., Word document) and two delivered feedback orally (i.e., Zoom meeting). The Zoom meetings were recorded and notes were taken by the author of this study based upon the oral feedback. The Excel Tool document and Examples document were adjusted according to the feedback. The main points provided through the feedback by the reviewers and the changes made to the tool to address the feedback were compiled into one document, which is displayed in Appendix 3.8.

After the tool had been pilot tested by the experts, participants were invited to participate via email. Participants were given access to a Qualtrics link where they were able to read the consent form, located on the first page. After digitally signing the consent form, participants answered the screening questions, which determined if the results for the participant would be included for analyses. Next, the participant were shown the demographic survey and asked to complete the survey. Participants were able to skip any questions they did not wish to answer. After completing the demographic survey, participants were shown the five pre/posttest AB graphs/data sets paired with the research questions. Participants were asked “Report the metric you would use to answer the research question(s). Provide a rationale for the metric(s) you would use. Optional: Complete the calculation.”
After submitting this, the participants moved onto the next page where a link to the tool (Excel Tool and Examples) was provided. At this time, the participants were no longer able to look back at the pretest graphs/data sets and research questions and/or change their submitted answers. The participants read the instructions: “The following link is to a downloadable Excel tool for picking and rationalizing appropriate SCED metrics. Please read the instructions found on Sheet 1 of the document Excel Tool before using the other sheets of the tool. You are also free to practice using the Excel Tool with the Examples document, or with any of your own graphs/data sets for as long as you would like. When you are ready you can move onto the next part of the study, found here, where you will be asked to use the tool to answer the same questions about the graphs/data sets and research questions that you were presented with in the beginning of this study.” After reading these instructions, the participants could access the Excel Tool and Examples documents. The participant could use the example graphs/data sets and research questions to practice using the tool for an unlimited amount of time.

When participants were ready, they returned to the Qualtrics link. Participants were shown the same five AB graphs they used during Step 5 and asked “Using the document Excel Tool, report the metric you would use to answer the research question. Provide a rationale for the metric(s) you would use. Optional: Complete the calculation.”). Once this was completed, participants moved onto the next page of Qualtrics and were asked to complete the demographic survey. The study completion time was recorded, and this marked the end of the study.

Statistical Analysis

The participant received a score using the rubric in Figure 3.6 for each of the three indicators, both before using the tool (pretest) and after accessing the tool (posttest). A score was given for each of the five graphs/data sets that were displayed to the participant, and a total score
across the five graphs was calculated. Thus, for indicator 1 the range of total possible scores for
the pre and posttest was between 5 – 28\(^{11}\), for indicator 2 it was between 5 – 20, and for indicator
3 it was between 5 – 15. Participants were ineligible to be included in the analysis if they scored
a “28” for indicator “quantification technique (s) chosen for use” and a “20” for “justification of
quantification technique (s)” during the pretest. This indicated that the participant had essentially
“baselined out” of the study, as they would have hit the ceiling effect of possible scores.
However, their demographic and social validity results could still be included. If a participant
chose, justified, or calculated the estimate of more than one quantification technique, only the
quantification technique with the highest score for choice, justification, and calculation of
estimate was included in the analysis.

Two-tailed paired t-tests were conducted to compare the pretest and post scores in choice,
justification, and calculation of estimate of quantification technique for each of the five
questions, as well as the accumulated score (i.e., the total score across the five questions). The
change in score for each participant was calculated, and the mean change in score was tested to
determine if it was statistically significantly different from zero. These results were
supplemented with the Wilcoxon signed-rank test, which is a non-parametric test, as well. These
analyses answered research questions 1, 2, and 3.

One-way ANOVAs were conducted to determine if there were statistically significant
differences in total (i.e., across the five questions) pretest scores and change in scores for choice,
justification, and calculation of quantification technique for participants depending on their
academic background (i.e., special education, educational psychology, school psychology, or
other). The non-parametric Kruskal-Wallis test was also conducted. These analyses determined if

\(^{11}\) The maximum score was 28 rather than 30 because for question/graph/dataset #4, the total possible score for
indicator 1 was a “4” rather than a “6”.

88
there were groups of individuals for which the tool was more or less effective (i.e., answered research questions 1, 2, and 3).

Two-tailed independent paired t-tests and the non-parametric Mann-Whitney test were conducted to determine if there were statistically significant differences in total pretest scores and change in scores for choice, justification, and calculation of estimate of quantification technique for participants depending on their education level (Master’s degree or Ph.D.). These analyses determined if there were groups of individuals for which the tool was more or less effective (i.e., answered research questions 1, 2, and 3).

Pearson’s correlational analysis and Spearman’s correlational analysis were conducted to determine if there was a relationship between change in scores and time spent completing the study.

Descriptive analyses (i.e., averages and frequency) were reported for social validity and demographic surveys to determine the acceptability of the tool amongst the users, and this information was used to answer research question 4.
Chapter 4: Results

The purpose of this study was to create and validate a user-friendly tool to help users choose and justify their use of SCED quantification techniques. The second purpose of this study was to test if such a tool was effective in improving how users choose and justify their use of SCED quantification techniques. This study also aimed to see if such a tool was effective in assisting users to calculate the SCED estimate (this was a secondary aim, as conducting the calculation was optional for participants). Five data sets were created and both the data sets and the graphical displays of the data sets were provided for participants to review both before and after using the developed tool (thus, the tool can be considered the “intervention”). Their responses were scored using a rubric that was developed to measure the appropriateness of choice and justification for using the quantification technique, as well as the calculation. Furthermore, this study aimed to determine if users would find the tool to be helpful, and if they would use it in the future.

In this section, the social validity results are reported first. Next, descriptive results for the pretest scores for choice and justification of quantification technique are reported. This is followed by descriptive statistics for the change in scores for choice and justification of quantification technique. Lastly, the descriptive statistics for the time to complete the study are reported. The descriptive statistics were also used to determine whether assumptions about the data were met and thus appropriate for inferential statistics. In Part 3 the inferential analyses are reported. First, inferential analyses were conducted to determine if the tool caused a statistically significant change in scores for choice and justification of quantification technique. Then, inferential analyses were conducted to determine if there were differences between academic departments (special education, educational psychology, school psychology, and others) in their pretest scores and change in scores for both choice of and justification of the quantification
technique. This is followed by inferential analyses to determine if there were differences between education levels (Master’s degree vs. Ph.D.) in pretest scores and change in scores for both choice and justification of the quantification technique. Lastly, inferential analyses determined if there was a relationship between time to complete the study and change in scores for choice of and justification of quantification technique. No participants conducted the optional calculation, and so descriptive and inferential results are not reported for calculation.

**Descriptive Statistics**

In this section, the descriptive statistics are presented, beginning with the social validity results. Next, the pretest scores for choice and justification of quantification technique, and the change in scores for choice and justification of quantification technique (both aggregated results across the five questions and results per individual question) are presented. IBM SPSS Statistics (2017) was used for all descriptive statistics.

**Social Validity**

Upon completing the study, seventeen of the twenty-five participants eligible for inferential analysis reported that they would be very likely to use the tool in the future, six participants reported that they would be somewhat likely, and two participants reported that they would be somewhat unlikely to use it in the future. A total of twenty-one participants rated the tool as being very helpful for choosing an appropriate quantification technique, three participants rated the tool as being somewhat helpful, and one participant rated the tool as somewhat unhelpful. A total of nineteen participants rated the tool as being very helpful for justifying the use of a quantification technique, four participants rated the tool as somewhat helpful, and two participants rated the tool as somewhat unhelpful.
Of the four participants who were not eligible for inferential analysis, two participants reported they would be very likely to use the tool in the future, one participant reported they would be somewhat likely, and one participant reported they would be somewhat unlikely to use the tool in the future. All four participants rated the tool as very helpful for choosing an appropriate quantification technique. Two participants rated the tool as very helpful for justifying the use of the quantification technique, one participant rated the tool as somewhat helpful, and one participant rated the tool as somewhat unhelpful. Fifteen participants left comments after completing the study and using the tool. The comments are displayed in Table 4.2.

**Table 4.2**

*Comments from Participants*

<table>
<thead>
<tr>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>
Comments

11 Makes decision making systematic and if everyone were to use same process there would be consistency in the literature, less errors. Definitely realized I knew less than I thought I did. Great tool!

12 I have a question. The first one and two ranking metrics are able to answer the research question and account for certain data features. why one is more recommended than others.

13 I am a doctoral student in school psych who is in the midst of proposing a SCD dissertation. I will definitely be referencing this tool when deciding which metric to choose for that project! This is excellent - thank you for creating such a helpful tool!

14 I am not terribly familiar with quantitative analyses but this tool is a very user-friendly introduction to the many options and provides insight into the rationale behind one versus the other without forcing a blind approach to the extensive literature on the topic.

15 This tool would be EXTREMELY useful in selecting single-case metrics. I found it easy to use, and made me a lot more confident in my decision-making about which metric to choose. During the pre-test, I was mostly guessing and it made me realize that I perhaps didn't know as much as I thought I did about how to quantitatively analyze the graphs. I am a visual person, and tend to justify my answers by looking at a graph rather than calculating a metric. This tool could help to quantify the justification.

16 This is such a great tool, and a great study! I do hope you will win some sort of grant so you can develop an app that can be used on PCs, mobile phones, and tablets. such a great great idea!

Note. Comment 6 was provided during the post-test, in one of the boxes meant for choosing and justifying a quantification technique, instead of at the end of the study.

Pretest Scores

The pretest score data was used to determine if it was more challenging to choose and justify quantification techniques depending on the research questions, data characteristics, and desired quantification technique features before accessing the tool (i.e., the study intervention)\(^{12}\). Pretest scores were also used for inferential analyses to see if there were differences in pretest scores (i.e., what participants knew before accessing the tool) for participants of different academic departments and education levels. As there were less than 30 participants, and non-

---

\(^{12}\) Pretest scores for calculation of estimate are not provided, as no participants completed the optional calculation.
normally distributed scores can indicate a non-normal sampling distribution, outliers and normality of scores are also reported to help determine the appropriateness of the parametric tests that were used for the inferential results. The total and individual question pretest scores are reported in Table 4.3. Although means are reported in the Table 4.3, the median is also reported in the following sections, as median is less influenced by outliers, and in instances when outliers are present median is therefore a better representative of scores. Posttest scores are also listed in the table; however, the data distribution for the posttest scores is not examined, as inferential analyses were only conducted on pretest scores and the change between pretest and posttest scores.
Table 4.3

Average Scores for Pretest and Posttest Questions 1-5.

<table>
<thead>
<tr>
<th>Choice of Quantification Technique</th>
<th>Justification of Quantification Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question 1</td>
<td></td>
</tr>
<tr>
<td>Research question:</td>
<td></td>
</tr>
<tr>
<td>Magnitude of change in level</td>
<td></td>
</tr>
<tr>
<td>Data characteristic: Small</td>
<td></td>
</tr>
<tr>
<td>number of measurement occasions in baseline phase</td>
<td></td>
</tr>
<tr>
<td>Pretest</td>
<td>Posttest</td>
</tr>
<tr>
<td>$M = 2.44$</td>
<td>$M = 5.40$</td>
</tr>
<tr>
<td>$SD = 1.61$</td>
<td>$SD = 1.47$</td>
</tr>
<tr>
<td>$\Delta M = 2.96$</td>
<td>$SD = 1.88$</td>
</tr>
<tr>
<td>$M = 1.72$</td>
<td>$M = 2.48$</td>
</tr>
<tr>
<td>$SD = 0.84$</td>
<td>$SD = 0.82$</td>
</tr>
<tr>
<td>$\Delta M = 0.76$</td>
<td>$SD = 0.88$</td>
</tr>
<tr>
<td>Question 2</td>
<td></td>
</tr>
<tr>
<td>Research question:</td>
<td></td>
</tr>
<tr>
<td>Magnitude of change in trend</td>
<td></td>
</tr>
<tr>
<td>Data characteristic: baseline trend</td>
<td></td>
</tr>
<tr>
<td>Pretest</td>
<td>Posttest</td>
</tr>
<tr>
<td>$M = 4.08$</td>
<td>$M = 5.80$</td>
</tr>
<tr>
<td>$SD = 2.06$</td>
<td>$SD = 0.65$</td>
</tr>
<tr>
<td>$\Delta M = 1.72$</td>
<td>$SD = 0.40$</td>
</tr>
<tr>
<td>$M = 2.00$</td>
<td>$M = 2.88$</td>
</tr>
<tr>
<td>$SD = 0.71$</td>
<td>$SD = 0.93$</td>
</tr>
<tr>
<td>$\Delta M = 0.88$</td>
<td>$SD = 1.17$</td>
</tr>
<tr>
<td>Question 3</td>
<td></td>
</tr>
<tr>
<td>Research question:</td>
<td></td>
</tr>
<tr>
<td>Magnitude of change in variability</td>
<td></td>
</tr>
<tr>
<td>Data characteristic: Variability</td>
<td></td>
</tr>
<tr>
<td>Pretest</td>
<td>Posttest</td>
</tr>
<tr>
<td>$M = 2.56$</td>
<td>$M = 4.00$</td>
</tr>
<tr>
<td>$SD = 1.16$</td>
<td>$SD = 0.00$</td>
</tr>
<tr>
<td>$\Delta M = 1.44$</td>
<td>$SD = 1.16$</td>
</tr>
<tr>
<td>$M = 1.56$</td>
<td>$M = 2.36$</td>
</tr>
<tr>
<td>$SD = 0.58$</td>
<td>$SD = 0.86$</td>
</tr>
<tr>
<td>$\Delta M = 0.80$</td>
<td>$SD = 0.96$</td>
</tr>
<tr>
<td></td>
<td>Choice of Quantification Technique</td>
</tr>
<tr>
<td>------------------------------</td>
<td>------------------------------------</td>
</tr>
<tr>
<td></td>
<td>Pretest</td>
</tr>
<tr>
<td>Question 4</td>
<td></td>
</tr>
<tr>
<td>Research question:</td>
<td></td>
</tr>
<tr>
<td>Non-overlapping data</td>
<td></td>
</tr>
<tr>
<td>Data characteristic:</td>
<td></td>
</tr>
<tr>
<td>Outliers</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$M = 4.76$</td>
</tr>
<tr>
<td></td>
<td>$SD = 1.42$</td>
</tr>
<tr>
<td>Question 5</td>
<td></td>
</tr>
<tr>
<td>Research question:</td>
<td></td>
</tr>
<tr>
<td>Magnitude of change in</td>
<td></td>
</tr>
<tr>
<td>level</td>
<td></td>
</tr>
<tr>
<td>Desired quantification</td>
<td></td>
</tr>
<tr>
<td>technique feature:</td>
<td></td>
</tr>
<tr>
<td>$p$-value</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$M = 4.32$</td>
</tr>
<tr>
<td></td>
<td>$SD = 1.86$</td>
</tr>
<tr>
<td>Total</td>
<td>$M = 18.16$</td>
</tr>
<tr>
<td></td>
<td>$SD = 5.84$</td>
</tr>
</tbody>
</table>

*Note.* $n = 25$. Maximum possible score for each question for pretest and posttest for choice of quantification technique is 6.00, with exception of Question 3, for which it is 4.00. Maximum possible score for each question for pretest and posttest justification of quantification technique is 4.00.
**Total Pretest Scores.** Figure 4.1 shows the boxplots\(^\text{13}\) for total pretest scores (i.e., across the five questions) for choice and justification of quantification technique. For choice of quantification technique, the median total score during pretest was 18.00 (range = 8.00-26.00, \(M = 18.16, SD = 5.84\)). Shapiro-Wilk test showed evidence that the data were normally distributed, \(W(25) = 0.92, p = .054\), with skewness of -0.25 (\(SE = 0.46\)) and kurtosis of -1.26 (\(SE = 0.89\)). For justification of quantification technique, the median total score during pretest was 9.00 (range = 5.00-15.00, \(M = 9.16, SD = 2.49\)). Shapiro-Wilk test showed evidence that the data were normally distributed, \(W(25) = 0.98, p = .76\) with skewness of -0.20 (\(SE = 0.46\)) and kurtosis of 0.30 (\(SE = 0.89\)). The skewness and kurtosis values were small, and upon examining the boxplots in Figure 4.1, it appears there were no outliers and the distributions were approximately symmetrical and normal.

**Figure 4.1**

*Total Pretest Scores for Choice and Justification of Quantification Technique*

\(^{13}\) All boxplots were created using R version 3.6.2.
Figures 4.2 and 4.3 show the pretest scores for each of the individual questions, for choice and justification of quantification technique, respectively.

**Figure 4.2**

*Pretest Scores for Choice of Quantification Technique for Each Individual Question*

![Box plot showing pretest scores for choice of quantification technique for each individual question.](image)

**Figure 4.3**

*Pretest Scores for Justification of Quantification Technique for Each Individual Question*

![Box plot showing pretest scores for justification of quantification technique for each individual question.](image)
Question 1 (Research question: Magnitude of change in level. Data characteristic: Small number of data points in baseline or treatment phase.). For choice of quantification technique, the median score during pretest was 2.00 (range = 1.00-6.00, $M = 2.44$, $SD = 1.61$). Shapiro-Wilk test showed evidence that that the data were non-normally distributed, $W(25) = 0.79, p < .05$, with skewness of 1.16 ($SE = 0.64$) and kurtosis of 0.78 ($SE = 0.90$). For justification of quantification technique, the median score during pretest was 2.00 (range = 1.00-4.00, $M = 1.72$, $SD = 0.84$). Shapiro-Wilk test showed evidence that that the data were non-normally distributed, $W(25) = 0.79, p < .05$, with skewness of 1.10 ($SE = 0.46$) and kurtosis of 0.66 ($SE = 0.90$).

Question 2 (Research question: Magnitude of change in trend. Data characteristic: Baseline trend.). For choice of quantification technique, the median score during pretest was 5.00 (range = 1.00-6.00, $M = 4.08$, $SD = 2.06$). Shapiro-Wilk test showed evidence that that the data were non-normally distributed, $W(25) = 0.79, p < .05$, with skewness of -0.40 ($SE = 0.46$) and kurtosis of -1.67 ($SE = 0.90$). For justification of quantification technique, the median score during pretest was 2.00 (range = 1.00-3.00, $M = 2.00$, $SD = 0.71$). Shapiro-Wilk test showed evidence that that the data were non-normally distributed, $W(25) = 0.82, p < .05$, with skewness of 0.00 ($SE = .46$) and kurtosis of -0.85 ($SE = 0.90$).

Question 3 (Research question: Magnitude of change in variability. Data characteristic: Variability.). For choice of quantification technique, the median score during pretest was 2.00 (range = 1.00-4.00, $M = 2.56$, $SD = 1.16$). Shapiro-Wilk test showed evidence that that the data were non-normally distributed, $W(25) = 0.77, p < .05$, with skewness of 0.28 ($SE = 0.46$) and kurtosis of -1.52 ($SE = 0.90$). For justification of quantification technique, the median score during pretest was 2.00 (range = 1.00-3.00, $M = 1.56$, $SD = 0.58$). Shapiro-Wilk
test showed evidence that that the data were non-normally distributed, $W(25) = 0.73, p < .05$, with skewness of 0.43 ($SE = 0.46$) and kurtosis of -0.67 ($SE = 0.90$).

**Question 4 (Research question: Non-overlapping data. Data characteristic: Outliers).** For choice of quantification technique, the median score during pretest was 5.00 (range = 1.00-6.00, $M = 4.76$, $SD = 1.42$). Shapiro-Wilk test showed evidence that that the data were non-normally distributed, $W(25) = 0.76, p < .05$, with skewness of -1.62 ($SE = 0.46$) and kurtosis of 2.32 ($SE = 0.90$). For justification of quantification technique, the median score during pretest was 2.00 (range = 1.00-3.00, $M = 1.88$, $SD = 0.72$). Shapiro-Wilk test showed evidence that that the data were non-normally distributed, $W(25) = 0.81, p < .05$, with skewness of 0.19 ($SE = 0.46$) and kurtosis of -0.97 ($SE = 0.90$).

**Question 5 (Research question: Magnitude of change in trend. Desired quantification technique feature: p-value).** For choice of quantification technique, the median score during pretest was 6.00 (range = 1.00-6.00, $M = 4.32$, $SD = 1.86$). Shapiro-Wilk test showed evidence that that the data were non-normally distributed, $W(25) = 0.76, p < .05$, with skewness of -0.38 ($SE = 0.46$) and kurtosis of -1.52 ($SE = 0.90$). For justification of quantification technique, the median score during pretest was 2.00 (range = 1.00-3.00, $M = 2.00$, $SD = 0.71$). Shapiro-Wilk test showed evidence that that the data were non-normally distributed, $W(25) = 0.81, p < .05$, with skewness of 0.00 ($SE = 0.46$) and kurtosis of -0.85 ($SE = 0.90$).

**Change in Scores**

Change is scores are reported, as inferential analyses were conducted to determine whether the tool was effective for improving choice and justification of quantification technique, and if it was more effective for improving choice or justification of quantification technique for
individual questions\textsuperscript{14}. Inferential analyses were also conducted to determine if there were differences in change in scores across different academic departments and education levels. Outliers and normality of scores are reported here to help determine the appropriateness of the parametric tests that are used for the inferential results. Change in scores for choice and justification of quantification techniques are displayed in Table 4.3.

**Total Change in Scores.** The boxplots in Figure 4.4 shows the range and distribution of the change in total scores across the five questions for choice and justification of quantification technique. The median change in scores between pretest and posttest for choice of quantification technique was 7.00 (range = 1.00-20.00, \( M = 8.92, SD = 5.93 \)). Shapiro-Wilk test showed evidence that that the data were non-normally distributed, \( W(25) = 0.91, p < .05 \), with skewness of 0.42 (\( SE = 0.46 \)) and kurtosis of -1.24 (\( SE = 0.90 \)). The median change in scores between pretest and posttest for justification of quantification technique was 4.00 (range = -6.00-11.00, \( M = 3.96, SD = 3.67 \)). Shapiro-Wilk test showed evidence that that the data were normally distributed, \( W(25) = 0.98, p = .62 \), with a skewness of -0.58 (\( SE = 0.46 \)) and kurtosis of 1.15 (\( SE = 0.90 \)). Upon examining the boxplots in Figure 4.3, it can be concluded that there were no outliers for change in scores for choice of quantification technique, and one outlier for change in scores for justification of quantification technique. The distributions were approximately symmetrical, particularly for justification of quantification technique.

\textsuperscript{14} Change in scores for calculation of estimate are not provided, as no participants completed the optional calculation.
Figures 4.5 and 4.6 show the change in scores for each individual question, for choice and justification of quantification technique, respectively.
Question 1 (Research question: Magnitude of change in level. Data characteristic: Small number of data points in baseline or treatment phase.). The median change in score for choice of quantification technique was 3.00 (range = -1.00-5.00, $M = 2.96$, $SD = 1.88$). Shapiro-Wilk test showed evidence that the data were non-normally distributed, $W(25) =$
0.86, \( p < .05 \), with skewness of -0.71 (SE = 0.46) and kurtosis of -0.51 (SE = 0.90). The median change in scores for justification of quantification technique was 1.00 (range = -1.00-2.00, \( M = 0.76 \), \( SD = 0.88 \)). Referring to Figure 4.6, no outliers were present. Shapiro-Wilk test showed evidence that that the data were non-normally distributed, \( W(25) = 0.88, p < .05 \), with skewness of -0.29 (SE = 0.46) and kurtosis of -0.43 (SE = 0.90).

**Question 2 (Research question: Magnitude of change in trend. Data characteristic: Baseline trend.).** The median change in score for choice of quantification technique was 1.00 (range = 0.00-5.00, \( M = 1.72 \), \( SD = 0.40 \)). No outliers were present. Shapiro-Wilk test showed evidence that that the data were non-normally distributed, \( W(25) = 0.77, p < .05 \), with skewness of 0.62 (SE = 0.46) and kurtosis of -1.34 (SE = 0.90). The median change in score for justification of quantification technique was 1.00 (range = -2.00-3.00, \( M = 0.88 \), \( SD = 1.17 \)). Shapiro-Wilk test showed evidence that that the data were non-normally distributed, \( W(25) = 0.89, p < .05 \), with skewness of -0.09 (SE = 0.46) and kurtosis of 0.89 (SE = 0.90). Two outliers were present.

**Question 3 (Research question: Magnitude of change in variability. Data characteristic: Variability.).** The median change in score for choice of quantification technique is 2.00 (range = 0.00-3.00, \( M = 1.44 \), \( SD = 1.16 \)). Shapiro-Wilk test showed evidence that that the data were non-normally distributed, \( W(25) = 0.77, p < .05 \), with skewness of -0.28 (SE = 0.46) and kurtosis of -1.52 (SE = 0.90). The median change in score for justification of quantification technique was 1.00 (range = -1.00-2.00, \( M = 0.78 \), \( SD = 0.96 \)). Shapiro-Wilk test showed evidence that that the data were non-normally distributed, \( W(25) = 0.86, p < .05 \), with skewness of -0.50 (SE = 0.46) and kurtosis of -0.49 (SE = 0.90). No outliers were present.
**Question 4 (Research question: Non-overlapping data. Data characteristic: Outliers).** The median change in score for choice of quantification technique was 1.00 (range = -1.00-5.00, $M = 1.12$, $SD = 1.51$). There were two outliers. Shapiro-Wilk test showed evidence that that the data were non-normally distributed, $W(25) = 0.81, p < .05$, with skewness of 1.44 ($SE = 0.46$) and kurtosis of 1.89 ($SE = 0.90$). The median change in score for justification of quantification technique was 1.00 (range = -2.00-2.00, $M = 0.68$, $SD = 1.14$). There was one outlier. Shapiro-Wilk test showed evidence that that the data were non-normally distributed, $W(25) = 0.86, p < .05$, with skewness of -0.80 ($SE = 0.46$) and kurtosis of -0.22 ($SE = 0.90$).

**Question 5 (Research question: Magnitude of change in trend. Desired quantification technique feature: p-value.).** The median change in score for choice of quantification technique was 0.00 (range = 0.00-5.00, $M = 1.68$, $SD = 1.86$). There were no outliers. Shapiro-Wilk test showed evidence that that the data were non-normally distributed, $W(25) = 0.76, p < .05$, with skewness of 0.38 ($SE = 0.46$) and kurtosis of -1.52 ($SE = 0.90$). The median change in score for justification of quantification technique was 1.00 (range = -1.00-3.00, $M = 0.84$, $SD = 0.90$). Shapiro-Wilk test showed evidence that that the data were non-normally distributed, $W(25) = 0.90, p < .05$, with skewness of 0.34 ($SE = 0.46$) and kurtosis of 0.24 ($SE = 0.90$).

**Summary.** For choice of quantification technique, the question with the greatest change between pretest and posttest was Question 1 ($\Delta M = 2.96$). During pretest, participants scored an average of 2.44 ($M = 2.44$, $SD = 1.61$); most participants chose a quantification technique that would partially account for one essential graph element. During posttest for Question 1, participants scored an average of 5.40 ($M = 5.40$, $SD = 1.47$), demonstrating that the average participant chose a quantification technique that partially accounted for one essential element and
accounted for one essential element. Question 4 was the question with the smallest change in
scores between the pretest and posttest for quantification technique choice ($\Delta M = 1.12$).
Participants on average scored 4.76 ($M = 4.76$, $SD = 1.42$) during the pretest for Question 4,
meaning that the average participant chose a quantification technique that partially accounted for
1 essential graph element and accounted for 1 essential graph element. These high pretest scores
prevented the participants from making large gains during posttest. During posttest for Question 4,
participants scored an average of 5.88 ($M = 5.88$, $SD = 0.44$) for quantification technique
choice; the average participant chose a quantification technique that could account for two
essential graph elements. For each of the five questions, there was an improvement between
pretest and posttest for quantification technique choice, with some questions demonstrating a
larger improvement than others.

For justification of quantification technique, the question with largest average change in
scores between pretest and posttest was Question 2 ($\Delta M = 0.88$). During pretest, participants
scored an average of 2.00 ($M = 2.00$, $SD = 0.71$); the average participant provided one
appropriate justification for using the quantification technique. During posttest for the same
question, participants scored an average of 2.88 ($M = 2.88$, $SD = 0.93$), demonstrating that the
average participant provided two appropriate justifications for using the quantification technique
during posttest. Question 4 was the question with the smallest change in scores between the
pretest and posttest for quantification technique choice ($\Delta M = 0.68$). Participants on average
scored 1.88 ($M = 1.88$, $SD = 0.73$) during the pretest for Question 4, meaning that the average
participant provided one appropriate justification. During posttest for Question 4, participants
scored an average of 2.56 ($M = 2.56$, $SD = 0.87$) for quantification technique choice; the average
participant chose a quantification technique that could account for two essential graph elements.
For each of the five questions, there was an improvement between pretest and posttest for quantification technique justification, with some questions demonstrating a larger improvement than others (see Table 4.3).

**Time to Complete Study**

The boxplot in Figure 4.7 shows the range of time taken to complete the study. The median time taken to complete the study was 48.00 minutes (range = 14.95 - 9,063.00, $M = 484.00$, $SD = 1,813.00$). Shapiro-Wilks test showed non-normal distribution, $W(25) = 0.27$, $p < .05$, with skewness of 4.80 ($SE = 0.46$) and kurtosis of 23.44 ($SE = 0.90$). It is evident that there are outliers. Five outliers were identified through examination of the interquartile range (i.e., more than 1.5 times the interquartile range above the third quartile or below the first quartile). The median time taken to complete the study with the five outliers removed was 40.73 minutes (range = 14.95 – 62.27, $M = 40.48$, $SD = 13.49$). Figure 4.8 shows the boxplot and range with these five outliers removed for easier interpretation of the data.

**Figure 4.7**

*Total Time to Complete Study*
Part 3: Inferential Analyses

In this section, the results from the inferential statistical analyses are discussed. A two-tailed paired t-test and the non-parametric equivalent Wilcoxon signed-rank test were conducted to determine if the change in scores between pretest and posttest for choice and justification of quantification technique was statistically significant. Next, a one-way ANOVA and Kruskal-Wallis test were conducted to determine if there were statistically significant differences in pretest scores and change in scores for participants depending on their academic background (i.e., special education, educational psychology, school psychology, or other). A two-tailed independent paired t-test and Mann-Whitney test were conducted to determine if there were statistically significant differences in pretest scores and change in scores for participants depending on their education level (Master’s degree or Ph.D.). Lastly, a correlation analysis was conducted to determine if there was a relationship between change in scores for choice and justification of quantification technique and time to complete the study. A total of 25 participants

Figure 4.8

Total Time to Complete Study with Five Outliers Removed
were eligible for inferential analyses. IBM SPSS Statistics (2017) was used to analyze all inferential results.

**Change in Scores**

Six two-tailed paired t-tests were conducted to determine if the change in scores for a.) choosing and b.) justifying the quantification technique before and after using the tool was statistically significant for total change in scores and for each of the individual questions. The paired t-test is a parametric test, and makes the assumption of normality. A total of 30 participants are typically needed to make the assumption that the sampling distribution is normally distributed. If there are less than 30 participants and the pretest scores and change in scores are non-normal, then the sampling distribution will be non-normal. As shown in the previous section, many of the pretest scores and change in scores were non-normal. The total pretest scores for choice of quantification technique, total pretest scores for justification of quantification technique, and the total change in scores for justification of technique were the only data for which normality assumption was met. For this reason, all inferential results are supplemented with non-parametric tests, which do not make asymptotical assumptions. Bonferroni correction was applied to account for Type I error (.05 / 6 = .0083).

**Total Change in Scores.** The total average change in scores for choosing an appropriate quantification technique before \((M = 18.16, SD = 5.84)\) and after \((M = 27.08, SD = 1.63)\) using the tool was \(8.92 (SD = 5.93), t(24) = 7.52, p < .0083, d = 1.50, 95\% \text{ CI} [6.47, 11.37]\), indicating a statistically significant change in total scores (i.e., across the five questions) after accessing the tool. After accessing the tool, participants’ scores were 1.50 standardized units higher than their scores before accessing the tool. The non-parametric Wilcoxon signed-rank test was also conducted \((Z = 4.28, p < .0083)\), and is in line with results of the parametric test. The total
change in scores for justifying the use of a quantification technique before \( (M = 9.16, SD = 2.49) \) and after \( (M = 13.12, SD = 3.52) \) using the tool was 3.96 \( (SD = 3.67) \), \( t(24) = 5.40, p < .0083, d = 1.08, 95\% CI [2.45, 5.47] \), indicating a statistically significant change in scores between the pretest and posttest. After accessing the tool, participants’ scores were 1.08 standardized units higher than their scores before accessing the tool. The Wilcoxon signed-rank test was also conducted, and the results were the similar \( (Z = 3.69, p < .0083) \).

**Question 1 (Research question: Magnitude of change in level. Data characteristic: Small number of data points in baseline or treatment phase.).** The average change in scores for choosing an appropriate quantification technique before and after using the tool was 2.96 \( (SD = 1.88) \), \( t(24) = 7.87, p < .0083, d = 1.57, 95\% CI [2.18, 3.74] \), indicating a statistically significant change in score after accessing the tool. After accessing the tool, participants’ scores were 1.57 standardized units higher than their scores before accessing the tool. The non-parametric Wilcoxon signed-rank test also indicated a statistically significant change in score \( (Z = 4.02, p < .0083) \). The average change in scores for justifying the use of a quantification technique before and after using the tool was 0.76 \( (SD = 0.88) \), \( t(24) = 4.32, p < .0083, d = 0.86, 95\% CI [0.40, 1.12] \), indicating a statistically significant change in scores between the pretest and posttest. After accessing the tool, participants’ scores were 0.86 standardized units higher than their scores before accessing the tool. The Wilcoxon signed-rank test also indicated a statistically significant change in score \( (Z = 3.27, p < .0083) \).

**Question 2 (Research question: Magnitude of change in trend. Data characteristic: Baseline trend.).** The average change in scores for choosing an appropriate quantification technique before and after using the tool was 1.72 \( (SD = 2.01) \), \( t(24) = 4.28, p < .0083, d = 0.86, 95\% CI [0.89, 2.55] \), indicating a statistically significant change in score after accessing the tool.
After accessing the tool, participants’ scores were 0.86 standardized units higher than their scores before accessing the tool. The non-parametric Wilcoxon signed-rank test also indicated a statistically significant change in scores ($Z = 3.20$, $p < .0083$). The average change in scores for justifying the use of a quantification technique before and after using the tool was 0.88 ($SD = 1.17$), $t(24) = 3.77$, $p < .0083$, $d = 0.76$, 95% CI [0.40, 1.36], indicating a statistically significant change in scores between the pre and posttest. After accessing the tool, participants’ scores were 0.76 standardized units higher than their scores before accessing the tool. The Wilcoxon signed-rank test also indicated a statistically significant change in score ($Z = 3.06$, $p < .0083$).

**Question 3** (Research question: Magnitude of change in variability. Data characteristic: Variability.). The average change in scores for choosing an appropriate quantification technique before and after using the tool was 1.44 ($SD = 1.16$), $t(24) = 6.22$, $p < .0083$, $d = 1.24$, 95% CI [0.96, 1.92], indicating a statistically significant change in score after accessing the tool. After accessing the tool, participants’ scores were 1.24 standardized units higher than their scores before accessing the tool. The non-parametric Wilcoxon signed-rank test also indicated a statistically significant change in score ($Z = 3.70$, $p < .0083$). The average change in scores for justifying the use of a quantification technique before and after using the tool was 0.80 ($SD = 0.957$), $t(24) = 4.18$, $p < .0083$, $d = 0.96$, 95% CI [0.40, 1.20], indicating a statistically significant change in scores between the pre and posttest. After accessing the tool, participants’ scores were 0.96 standardized units higher than their scores before accessing the tool. The Wilcoxon signed-rank test also indicated a statistically significant change in score ($Z = 3.22$, $p < .0083$).

**Question 4** (Research question: Non-overlapping data. Data characteristic: Outliers.). The average change in scores for choosing an appropriate quantification technique...
before and after using the tool was 1.12 (SD = 1.51), \( t(24) = 3.71, p < .0083, d = 0.74, 95\% \) CI [0.50, 1.74], indicating a statistically significant change in score after accessing the tool. After accessing the tool, participants’ scores were 1.12 standardized units higher than their scores before accessing the tool. The non-parametric Wilcoxon signed-rank test also indicated a statistically significant change in score (\( Z = 3.33, p < .0083 \)). The average change in scores for justifying the use of a quantification technique before and after using the tool was .68 (SD = 1.14), \( t(24) = 2.97, p < .0083, d = 0.59, 95\% \) CI [0.21, 1.15], indicating a statistically significant change in scores between the pre and posttest. After accessing the tool, participants’ scores were 0.59 standardized units higher than their scores before accessing the tool. The Wilcoxon signed-rank test also indicated a statistically significant change in scores (\( Z = 2.56, p < .0083 \)).

**Question 5 (Research question: Magnitude of change in trend. Desired quantification technique feature: \( p \)-value).** The average change in scores for choosing an appropriate quantification technique before and after using the tool was 1.68 (SD = 1.86), \( t(24) = 4.51, p < .0083, d = 0.90, 95\% \) CI [0.91, 1.21], indicating a statistically significant change in score after accessing the tool. After accessing the tool, participants’ scores were 0.90 standardized units higher than their scores before accessing the tool. The non-parametric Wilcoxon signed-rank test also indicated a statistically significant change in score (\( Z = 3.17, p < .0083 \)). The average change in scores for justifying the use of a quantification technique before and after using the tool was 0.84 (SD = 0.90), \( t(24) = 4.68, p < .0083, d = 0.94, 95\% \) CI [0.47, 1.21], indicating a statistically significant change in scores between the pre and posttest. After accessing the tool, participants’ scores were 0.94 standardized units higher than their scores before accessing the tool. The Wilcoxon signed-rank test also indicated a statistically significant change in score (\( Z = 3.46, p < .0083 \)).
Summary. There was a statistically significant change in total scores between pretest and posttest for both choice and justification of quantification technique. There was a statistically significant change in scores between pretest and posttest for each of the individual questions, as well. There was a very large effect of the tool on change in scores for choice of quantification technique, particularly for Question 1 ($d = 1.24$) and Question 3 ($d = 1.57$). There was a large effect of the tool on change in scores for justification of quantification technique, particularly for Question 3 ($d = 0.96$) and Question 5 ($d = 0.94$). The results are displayed in Table 4.4.

Table 4.4

Paired t-test Results

<table>
<thead>
<tr>
<th>Choice of Quantification Technique</th>
<th>Justification of Quantification Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$</td>
</tr>
<tr>
<td>Question 1</td>
<td>2.96</td>
</tr>
<tr>
<td>Question 2</td>
<td>1.72</td>
</tr>
<tr>
<td>Question 3</td>
<td>1.44</td>
</tr>
<tr>
<td>Question 4</td>
<td>1.12</td>
</tr>
<tr>
<td>Question 5</td>
<td>1.68</td>
</tr>
<tr>
<td>Total</td>
<td>8.92</td>
</tr>
</tbody>
</table>

Note. *$p < .0083$

Academic Department Differences

Four one-way ANOVAs were conducted to determine if there was a statistically significant difference in a.) total pretest scores (i.e., across the five questions) and b.) total change in scores for choice and justification of quantification technique, depending on the participant’s academic department (special education, educational psychology, school psychology, and others). For the purpose of inferential analyses, participants who reported being
a part of both special education and applied behavior analysis were coded as “special education” 
\( (n = 2) \). Bonferroni correction was applied to account for Type I error \((.05 / 4 = .0125)\).

G*Power version 3.1.9.1 was used to calculate power. Results of the ANOVAs are displayed in Table 4.5.

**Table 4.5**

**ANOVA results for Academic Department**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Academic Department</th>
<th></th>
<th></th>
<th></th>
<th>F(3, 21)</th>
<th>( \eta^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Special Education</td>
<td>Educational Psychology</td>
<td>School Psychology</td>
<td>Other</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Total pretest score</td>
<td>Choice</td>
<td>14.00</td>
<td>4.69</td>
<td>21.14</td>
<td>5.90</td>
<td>17.70</td>
</tr>
<tr>
<td>Justification</td>
<td>8.50</td>
<td>1.76</td>
<td>9.86</td>
<td>3.08</td>
<td>8.50</td>
<td>1.97</td>
</tr>
<tr>
<td>Justification</td>
<td>4.67</td>
<td>2.88</td>
<td>3.71</td>
<td>4.23</td>
<td>4.33</td>
<td>2.50</td>
</tr>
</tbody>
</table>

**Pretest Scores.** The ANOVA showed no statistically significant effect of academic department on the total pretest scores (i.e., across the five questions) for choice of quantification technique \([F(3, 21) = 1.96, \ p = .15, \ \eta^2 = .22]\), nor for justification of quantification technique \([F(3, 21) = 0.51, \ p = .68, \ \eta^2 = .07]\). The power to detect the effect of academic department on choice of quantification technique pretest scores was determined to be .04, and for justification it was determined to be .01. This indicates that there was not enough power to detect an effect of academic department on pretest scores. A total of 22% and 7% of the variance in pretest scores for choice and justification of quantification technique may be attributed to academic department, respectively. The non-parametric Kruskal-Wallis test was also conducted and
showed no effect of academic department on total pretest scores for choice of quantification technique \([H(3) = 5.29, p = .15]\), nor for justification of quantification technique \([H(3) = 1.90, p = .59]\).

**Change in Scores.** The ANOVA showed no statistically significant effect of academic department on total change in scores for choice of quantification technique \([F(3, 21) = 2.14, p = .13, \eta^2 = .23]\), nor for justification of quantification technique \([F(3, 21) = 0.18, p = .91]\). The power to detect the effect of academic department on choice of quantification technique change in scores was determined to be .04, and for justification it was determined to be .01. This indicated that there was not enough power to detect a difference between academic departments for quantification technique choice nor justification change in scores. A total of 23% and 2% of the variance in change in scores for choice and justification of quantification technique can be attributed to academic department, respectively. The non-parametric Kruskal-Wallis test was also conducted and showed no effect of academic department on change in scores for choice of quantification technique \([H(3) = 5.84, p = .12]\), nor for justification of quantification technique \([H(3) = 0.44, p = .93, \eta^2 = .02]\).

**Education Level Differences**

Four independent two sample t-tests were conducted to determine if there was a difference in a.) total pretest scores (i.e., across the five questions) and b.) total change in scores for choice and justification of quantification technique depending on education level (Master’s Degree vs. Ph.D.). Bonferroni correction was applied to account for Type I error \(.05 / 4 = .0125\). G*Power version 3.1.9.1 (Faul et al., 2009) was used to calculate power. Results of the independent t-tests are displayed in Table 4.6.
Table 4.6

*Independent t-test Results for Education Level*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Education Level</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Master’s Degree</td>
<td>Ph.D.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$M$</td>
<td>$SD$</td>
<td>$M$</td>
<td>$SD$</td>
<td>$T(21)$</td>
<td>$p$</td>
<td>Cohen’s $d$</td>
<td></td>
</tr>
<tr>
<td>Total Pretest Score</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Choice</td>
<td>18.50</td>
<td>5.49</td>
<td>17.19</td>
<td>6.55</td>
<td>0.51</td>
<td>.37</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>Justification</td>
<td>9.08</td>
<td>1.56</td>
<td>8.82</td>
<td>3.31</td>
<td>0.25</td>
<td>.81</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>Total Change in</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scores</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Choice</td>
<td>8.42</td>
<td>5.60</td>
<td>10.00</td>
<td>6.69</td>
<td>-0.61</td>
<td>.54</td>
<td>-0.26</td>
<td></td>
</tr>
<tr>
<td>Justification</td>
<td>4.83</td>
<td>2.82</td>
<td>3.64</td>
<td>3.41</td>
<td>0.92</td>
<td>.39</td>
<td>0.38</td>
<td></td>
</tr>
</tbody>
</table>

**Pretest Scores.** The two-tailed independent two sample t-test showed total pretest scores for quantification technique choice were not statistically significantly different for those with a Master’s Degree ($M = 18.50, SD = 5.49), t(21) = 0.51, p = .37, d = 0.22, than for those with a Ph.D. ($M = 17.19, SD = 6.55). The power to detect a difference in education levels for choice of quantification technique was found to be .04. This indicated that there was not enough power to detect education level differences in pretest scores for choice of quantification technique. During pretest, participants with a Master’s degree had scores for choice of technique that were 0.22 standardized units higher than the scores of participants with a Ph.D. A Mann-Whitney test indicated no difference in total pretest scores for quantification technique choice between those with a Master’s degree than for those with a Ph.D. ($U = 60.50, p = .73$). The independent two sample t-test showed total pretest scores for quantification technique justification were not statistically significantly different for those with a Master’s Degree ($M = 9.08, SD = 1.56), t(21) = 0.25, p = .81, d = 0.10, than for those with a Ph.D. ($M = 8.82, SD = 3.31). The power to detect a difference in education levels for justification of quantification technique was found to be .02.
This indicated that there was not enough power to detect education level differences in pretest scores for justification of quantification technique. During pretest, participants with a Master’s degree had scores that were 0.10 standardized units higher than the scores of participants with a Ph.D. A Mann-Whitney test also indicated no difference in total pretest scores for quantification technique justification between those with a Master’s degree than for those with a Ph.D. ($U = 58.50, p = .64$).

**Change in Scores.** The independent two sample t-test showed total change in score for quantification technique choice were not statistically significantly different for those with a Master’s Degree ($M = 8.42, SD = 5.60$), $t(21) = -0.61, p = .54, d = -0.26$, than for those with a Ph.D. ($M = 10.00, SD = 6.69$). The power to detect a difference in education level for choice of quantification technique was found to be .03. This indicated that there was not enough power to detect education level differences in change in scores for choice of quantification technique. Participants with a Master’s degree had a total change in scores for choice of technique that were 0.26 standardized units lower than the scores of participants with a Ph.D. A Mann-Whitney test also indicated no difference in total change in scores for quantification technique choice between those with a Master’s degree than for those with a Ph.D. ($U = 57.50, p = .60$). The independent two sample t-test showed total change in scores for quantification technique justification were not statistically significantly different for those with a Master’s Degree ($M = 4.83, SD = 2.82$), $t(21) = 0.92, p = .39, d = 0.38$), than for those with a Ph.D. ($M = 3.64 SD = 3.41$). The power to detect a difference in education level for justification of quantification technique was found to be .05. This indicated that there was not enough power to detect education level differences in change in scores for justification of quantification technique. Participants with a Master’s degree had a total change in scores for justification of technique that were 0.38 standardized units lower.
than the scores of participants with a Ph.D. A Mann-Whitney test also indicated no difference in total change in scores for quantification technique justification between those with a Master’s degree than for those with a Ph.D. \((U = 47.00, p = .24)\)

**Time to Complete Study**

Correlational analyses were conducted to determine the relationship between total time to complete the study and change in scores across the five questions. Results showed no statistically significant relationship between time to complete the study and change in scores for choosing a quantification technique, \(r = .31, p = .13\). Kendall’s rank correlation showed similar results \(\tau = .23, p = .11\). There was also no statistically significant relationship between time to complete study and change in scores for justifying the use of the quantification technique, \(r = .22, p = .30\). Kendall’s rank correlation showed similar results \(\tau = -.06, p = .69\). Taken together, it could be concluded that there was no relationship between the time it took to complete the study and the change in scores between pretest and posttest.
Chapter 5: Discussion

In this section, the results of this dissertation are discussed in detail, with a focus on the effectiveness of the Excel tool for choosing and justifying quantification techniques. The perceived effectiveness of the tool is also discussed. Furthermore, both the statistical and practical differences between participants (i.e., as a function of pretest scores, education level, and academic department) are discussed. This is followed by a discussion of the implications of the study, the limitations of both the study and the tool itself, and areas for future research.

Effectiveness of Excel Tool for Choosing Quantification Techniques

The tool is highly effective in assisting participants to choose an appropriate quantification technique for various research questions, data characteristics, and desired quantification technique features. This effectiveness is demonstrated through statistical significance, but also through practical significance. Practical significance can be established by comparing pretest and posttest scores. For example, referring to Question 1\textsuperscript{15}, during pretest, the majority of participants either chose a quantification technique that accounted for no essential data characteristics, and scored a 1.00 using the rubric shown in Table 3.1, or chose a quantification technique that accounted for one essential data characteristic and scored a 3.00. A participant who scored a 1.00 for Question 1 might have chosen a quantification technique such as PAND, which has low power with a small number of data points in the baseline phase (Petersen-Brown et al., 2012) and cannot answer research questions related to change in magnitude because it is a non-overlap quantification technique. This means that for Question 1, 40% \((n =10)\) of participants chose a quantification technique like PAND, which was

\textsuperscript{15}Research question: Magnitude of change in level. Data characteristic: Small number of data points in baseline or treatment phase.
inappropriate for use due to the research question and the data characteristics. After accessing the tool, most participants improved in their ability to choose an appropriate quantification technique for Question 1. During posttest, 80% \((n = 20)\) of participants chose a quantification technique that accounted for the two essential data characteristics (scored a 6.00), such as log response ratio, which can both answer the research question and account for small number of measurement occasions in the baseline phase.

A similar pattern of improvement is found for the other four individual questions, as well. Question 5\(^{16}\) was the question with the least change in scores between pretest and posttest; the median change in score was 0.00 \((\Delta M = 1.68)\), indicating no change. One possible reason that the median change in score for choice of quantification technique was 0.00 could be attributed to the research question and desired quantification technique feature for this particular question. Question 5 asked the participant about the magnitude of change in level and required the participants to find a quantification technique that could produce a \(p\)-value. Thus, one acceptable quantification technique for this particular question would be regression, as this could provide an estimate reflecting the magnitude of change in level, and could also produce a \(p\)-value. It is possible that participants were not certain regression was an appropriate SCED quantification technique, but were familiar with regression. Regression is used to predict an outcome score based on independent variables, and so participants might have known that a regression equation can also be used to quantify the change in level. They might have also already known that regression can produce a \(p\)-value. Eleven (44\%) of participants chose regression during pretest for Question 5; indeed one justification given for using regression was “I'm not sure on this, but I

\(^{16}\)Research question: Magnitude of change in level. Desired quantification technique feature: \(p\)-value.
think GLS may be the only approach listed with an accompanying significance test”. So although
the median change for Question 5 was 0.00, it is likely that some did not change their choice of
quantification technique after using the tool, but now were able to explain with certainty why the
quantification technique was appropriate for Question 5. Indeed, the median change in score for
justification of the quantification technique for Question 5 was 1.00, showing that participants
were able to provide one additional reason for using the quantification technique after accessing
the tool.

Participants also scored highly during pretest for choice of quantification technique for
Question 417, with a median score of 5.00. As a result of this high pretest score, the median
change in score between pretest and posttest for choice of quantification technique was only 1.00
($\Delta M = 1.12$). This lower change in score between pretest and posttest can be attributed to the
high pretest scores; the high pretest scores did not leave participants much room to improve
between pretest and posttest. It is possible that participants scored high during pretest for this
particular question because the research question was about amount of non-overlapping data and
the unique data characteristic was outliers. Due to the popularity of non-overlap techniques for
analyzing SCED data (Fingerhut et al., 2020), it is likely that most participants were already
familiar with non-overlap techniques, and the verbiage of the research question clearly indicated
that a non-overlap technique was needed. Therefore, it was likely easier to choose an appropriate
quantification technique for this particular question. Furthermore, most non-overlap
quantification techniques are relatively unaffected by outliers (with the exception of PND), so if
the participant chose any non-overlap measure besides PND, they would have scored at least a
5.00 for choice of quantification technique for this question.

17Research question: Non-overlapping data. Data characteristic: Outliers.
This may indicate yet another reason why the median change in score for Question 5 was 0.00. Questions 1 and 5 had similar research questions. Question 1 research question was “Does the intervention have an effect on scores, with effect defined as magnitude of change in level?”, while Question 5 research question was “What is the magnitude of change in level between baseline and intervention? \( p < .05 \)”. Both questions asked about the magnitude of change in level. For Question 1, a total of 16 (64%) participants chose a non-overlap technique to answer the research question, although non-overlap indices do not reflect the magnitude of the change in level. When participants were introduced to Question 4, the wording of the question (“How much data overlap or non-overlap is there between phases A and B?”) may have indicated to participants that these are the type of research questions for which non-overlap indices are appropriate. Thus, when participants were introduced to the last question in the study, Question 5, the participants may have then realized that non-overlap indices are not appropriate for questions about magnitude of change, and knew to no longer choose a non-overlap technique. Participants may have learned a little about the appropriateness of SCED quantification techniques simply from participating in the study.

The effectiveness of the tool in improving users’ choice of SCED quantification techniques was also demonstrated through the change in variability of scores between pretest and posttest. Across all five questions, there was a reduction in the variability of scores between pretest and posttest. Participants had a large range of scores for their choice of quantification technique before accessing the tool; for example, for Question 3\(^{18}\), the range of scores during pretest was between 1-4, with a median score of 2.00. This shows that some participants were able to choose the most appropriate quantification technique during pretest (keeping in mind that

the maximum possible score for Question 3 was 4.00), while others were not able to choose an appropriate quantification technique during pretest. However, after accessing the tool, there were no variability in posttest scores, meaning that the maximum score of 4.00 was consistently obtained. A similar pattern is seen for the other four questions, again demonstrating the effectiveness of the tool for choice of quantification technique.

**Effectiveness of Excel Tool for Justifying Quantification Techniques**

The tool was also effective in improving the way that users justify their use of quantification techniques. The effectiveness is demonstrated across data sets with varying research questions, data characteristics, and desired quantification technique features, and was evident through statistical significance but also through practical significance (i.e., examination of individual pretest and posttest scores).

The median change in score for justification of the quantification technique for each of the five questions was 1.00. This means that most participants were able to provide one additional justification after accessing the tool. For example, for Question 1, 12 (48%) participants provided no appropriate justification for using the quantification technique (scored a 1.00 using the rubric shown in Table 3.1) and nine (36%) participants provided one appropriate justification for using the quantification technique (scored a 2.00). During posttest, six (24%) participants provided one appropriate justification and fourteen participants (56%) provided two appropriate justifications. Similar results were found for the other four questions.

Although some individual questions (e.g., Questions 4 and 5) had no change or small change between pretest and posttest for choice of quantification technique, the majority of participants still improved in their ability to justify use of the quantification techniques between pretest and posttest. For example, for Question 5, the majority of participants chose an
appropriate quantification technique both before and after accessing the tool; however, the majority of participants improved in ability to justify their choice of the quantification technique. This demonstrates that most participants gained a better understanding of why the quantification technique was appropriate for use after accessing the tool.

**Perceived Effectiveness of the Excel Tool**

The results from the social validity questions were in line with the results demonstrating the effectiveness of the tool. The majority of participants \((n = 25, 86\%)\) rated the tool as highly helpful for choosing an appropriate quantification technique to use. The majority of participants \((n = 21, 72\%)\) also rated the tool as highly helpful for justifying use of the quantification technique. It is possible slightly more participants found the tool to be useful for choosing a quantification technique than for justifying the use of a quantification technique due to the way information was compiled in the tool. Information about each of the quantification techniques was spread throughout the tool (e.g., on the Tool tab and also the Metric Details tab), requiring the participant to sift through the information provided to determine an appropriate justification. On the other hand, to choose an appropriate quantification technique, the user only needed to input the information into the tool and view the sorted quantification techniques directly on the page of the Tool tab. This could be one reason users perceived the tool to be slightly more helpful for choosing a quantification technique than for justifying the quantification technique. However, the change in scores between the pretest and posttest for both quantification technique choice and justification were statistically and practically significant, demonstrating that the tool was indeed effective for both helping participants choose and justify a quantification technique regardless of the participants’ own perception of the tool’s effectiveness.
Nineteen (66%) participants reported that they would be very likely to use the tool in the future. This was slightly less than the number of participants who reported the tool as being highly effective. One reason that only 66% of participants reported they would be very likely to use the tool in the future could be because the tool was more useful in certain fields. For example, participants in the field of applied behavior analysis might have found the tool to be very helpful for choosing and justifying quantification techniques, but might not be very likely to use the tool because, as one participant in the applied behavior analysis field reported, “A really neat tool, but just doesn’t have a place in my practice”. Fields such as applied behavior analysis rely heavily on visual analysis rather than quantitative analysis, and so certain participants might not be as likely to use the tool due to their line of work. Still, the majority of participants reported that they would be at least somewhat likely to use the tool in the future (with only three participants reporting they would be somewhat unlikely to use the tool in the future), demonstrating that the majority of participants found the tool to be practical.

**Relation between Findings and Pretest Scores, Academic Department, and Education Level**

*Pretest Scores*

Although extent of participants’ prior SCED experience was not captured in this study, it is likely that the participants who scored higher during pretest had greater prior SCED experience than the participants who scored lower during pretest. The results of this study show that even those who scored high during pretest were still able to benefit from the tool. No participants scored the maximum score (i.e., scored a 6.00 by choosing a quantification technique that could account for two essential elements for Questions 1, 2, 4 and 5, or scored a 4.00 by choosing a quantification technique that could partially account for two essential elements for Question 3) for all five questions during pretest. Participants who scored high in
ability to choose an appropriate quantification technique during pretest still improved in ability
to choose an appropriate quantification technique after using the tool. For example, the top
eight\(^{19}\) (32\%) of participants who scored highest during the pretest with a total score (i.e., across
the five questions) of at least 23 and at most 26 out of 28 still improved after accessing the tool
by achieving a score of 28/28 for quantification technique choice at the end of the study. This
shows that even those who likely had prior SCED knowledge were able to improve in their
choice of SCED quantification technique after accessing the tool.

There are similar findings for justification of the quantification technique, with a few
exceptions. For example, for the top eight (32\%) participants who scored highest during the
pretest (with a total score of at least 11 and at most 15 out of 20), six were able to improve in
their justification of using the quantification technique after accessing the tool. This shows that
the tool was still effective for improving participants’ justification of quantification techniques
regardless of any past SCED experience/knowledge. However, there were three participants
(Participants 1, 18, and 29), including two who scored high during pretest in ability to justify the
quantification technique, who did slightly worse in their ability to justify the use of the
quantification technique between pretest and posttest. This decline in score could be attributed to
fatigue from the study; the recorded time to complete the study for Participants 1, 18, and 29,
was 5 hours and 46 minutes, 47 minutes, and 42 minutes, respectively. After accessing the tool,
these participants may have become tired from completing the study and so they did not spend as
much time justifying their choice of the quantification technique during the posttest.

The issue of fatigue is especially apparent for Participant 1, who wrote “yes” for each
justification; it seems that the participant was no longer interested in participating in the study

---

\(^{19}\) The top eight participants are reported rather than top five because there were ties among participants up until the
ninth participant.
based on this response. Participants 18 and 29 scored one point worse between pretest and posttest. Upon closer examination of their responses, both participants did worse on Question 4, for which the research question asked about the amount of non-overlap and the unique data/graph was an outlier. During pretest, Participant 18 chose PAND and wrote “There is obvious change of level. Overlapping points are observed clearly.” This justification was scored a 2.00 using the rubric, as one appropriate justification was provided (their reference to the overlapping data points). During posttest, this same participant chose NAP and wrote “Clearly observed immediate treatment effect visually”. This response was scored a 1.00 using the rubric, as it does not include an appropriate justification for using NAP. During pretest, Participant 29 chose PAND and wrote “I choose PAND here even though a couple of the quantification techniques can be used. There are two reasons, (1) it takes into account all data points; and (2) with certain manipulations it can be transformed into an effect size (phi2 or something along these lines)”. This response was scored a 3.00 using the rubric, as PAND does account for all data points and it can be converted into Phi. During posttest, this participant chose IRD and wrote “This quantification technique is not as affected by the outliers as some others. PAND is the second desirable choice though”. This response was scored a 2.00 using the rubric because IRD is indeed not highly affected by outliers. Another possible reason besides fatigue that Participant 29 did slightly worse between pretest and posttest in justification of the quantification technique could be that this particular participant already had high pretest score, and was not able to benefit as much as other participants did from the tool. This participant scored highest among all participants during pretest for justification of the quantification technique. Participant 29 rated the tool as “somewhat helpful” rather than “very helpful” for justifying SCED quantification techniques. It is possible that this rating reflects what this participant’s score reflects as well;
there was not “room” to improve in ability to justify choice of the quantification technique due to them already possessing the knowledge needed to provide appropriate quantification technique justifications. Furthermore, this participant had only small improvement between pretest and posttest for choice of quantification technique (a change in score of 4.00). Paired with study fatigue, as it is more cognitively demanding to justify use of a quantification technique than it is to choose a quantification technique, this could explain their slight decline in scores before and after using the tool.

**Education Level**

The results of the study showed that the tool was effective across different levels of education. It is unsurprising that education level was not found to be related to the effectiveness of the tool; those with various education levels could have various SCED experience. More participants with Ph.Ds. \((n = 7, 64\%)\) reported that they had published SCED research in the past than did participants with Master Degrees \((n = 3, 19\%)\). However, nine \((56\%)\) participants with Master Degrees reported that they would like to publish SCED research in the future. Furthermore, those with Master Degrees could have been taking a SCED course for a Ph.D. program at the time of the study, and so their SCED knowledge would be fresh in their minds during the time of the study. One professor who disseminated this study shared it with the students in their graduate level SCED course. Although there is no way to tell how many students from this class completed the study, this shows that some of the people with Master Degrees would indeed have knowledge of SCEDs that may have been on par with the knowledge of participants with Ph.Ds.
Results of this study showed no statistically significant difference in change in scores for choice or justification of the quantification technique across academic departments. However, the effect size for the difference in change in scores for choice of quantification technique between academic departments was relatively high ($\eta^2 = .23$), indicating that the lack of statistical significance could possibly be attributed to insufficient power as a result of a small sample size. Although this was not formally tested using a statistical significance test, as the overall ANOVA indicated that there were no differences between the groups, there appears to be the greatest difference in change in scores between those from special education department and those from educational psychology department. Participants who identified with the academic department special education had a median change in scores for choice of quantification technique of 14.50 ($\Delta M = 13.67$). Participants from educational psychology had a median change of 4.00 ($\Delta M = 6.14$). During pretest, those from special education department had a median score of 12.50 ($M = 14.00$) across the five questions for choice of quantification technique. Those from educational psychology scored a median of 23.00 ($M = 21.14$) during pretest. The tool was likely effective for both groups, as they had similar posttest scores ($Mdn = 28.00$ for both special education and educational psychology, $M = 27.67$ for special education, and $M = 27.28$ for educational psychology), but participants from special education were simply able to make greater gains between pretest and posttest due to their lower pretest scores. It is possible that participants from special education departments scored lower during pretest because special education historically uses visual analysis more than SCED quantification techniques. These participants may not have had as much exposure to SCED quantification techniques before beginning the study, while those from the field of educational psychology may have previously
learned about and used these quantification techniques in research. Regardless of these
differences during pretest, there was virtually no difference in posttest scores, showing that the
tool was helpful for both groups of participants.

**Implications of Current Study**

**Current Use of SCED Quantification Techniques**

The findings from this study have significant implications for the field of SCED. First,
the pretest results demonstrate that the average SCED researcher, with varying education levels
and from various academic departments, needs support in choosing and justifying SCED
quantification techniques. This may be especially true for those who are in the special education
field, as these participants scored lower on average during the pretest for this study. For example,
during pretest for Questions 1 and 3, 16 (64%) participants chose non-overlap techniques for use,
although non-overlap techniques are unable to provide an estimate that reflects the magnitude of
the change in level, nor the magnitude of change in variability. Thus, such quantification
techniques cannot provide an estimate that answers either research question. These results reflect
the misuse of non-overlap quantification techniques in particular, as similar pattern of mistakes
are not evident for Question 4. The research question for Question 4 was about the amount of
overlap/data separation between phases; only one participant chose a quantification technique
that could not appropriately answer this research question (i.e., the participant did not choose a
non-overlap quantification technique). SCED users may be used to using non-overlap
quantification techniques, due to their popularity (Fingerhut et al., 2020), but do not realize that
non-overlap quantification techniques cannot represent the magnitude of change/size of the
effect. Although terms such as “metric” and “effect size” are sometimes used to describe non-
overlap indices, it is recommended that the field of SCED no longer uses such terms in order to
avoid confusion regarding what these indices actually measure. The overuse of non-overlap quantification techniques in this study may highlight the importance of changing SCED terminology. No longer using terms such as “metric” to refer to non-overlap measures may better help SCED users carefully consider the quantification techniques they use, and better understand what exactly the estimates represent.

Similarly, participants of various education levels and academic departments chose quantification techniques that were inappropriate for use with the present graph/data characteristics. During pretest for Question 2, four (16%) participants chose quantification techniques that were highly affected by baseline trend, even though trend was present in the baseline phase. Similarly, some participants chose quantification techniques that could not completely account for outliers for Question 4 (n = 14, 56%). Taken together, this inability to choose an appropriate quantification technique is especially concerning because certain quantification techniques may not yield reliable results if the quantification techniques are inappropriate for use with the data. For example, Fingerhut et al. (in press-a) found that Tau estimates depend upon data characteristics such as baseline trend, within-case variability, and number of data points in the baseline phase. Pustejovsky (2019) found that procedural features such as the length of the observation sessions and the recording system used to collect measurements of behavior can influence the estimates of certain SCED quantification techniques. Other researchers have acknowledged previous research showing that certain quantification techniques are more or less appropriate depending on the research questions and data characteristics. Vannest et al. (2018) explicitly calls for researchers to choose SCED quantification techniques carefully in the context of their own research questions and data.

---

20 Research question: Magnitude of change in trend. Unique data characteristic: Baseline trend.
characteristics. Thus, it is concerning that pretest results highlight the use of inappropriate SCED quantification techniques.

Results of this study also show that SCED users from various education levels and academic departments need more support in justifying their use of the quantification techniques. This is a logical finding, as people cannot justify their use of quantification techniques if they are unable to choose quantification techniques that are wholly appropriate for use in the first place. While it is encouraging that the average participant could provide one justification during pretest, the results of this study show that SCED users need more support in justifying their use of quantification techniques.

Failure to take the research questions and data characteristics into account when choosing a quantification technique can ultimately result in the dissemination of findings that do not accurately represent actual intervention effects. This is especially concerning due to the wide use of evidence-based practices in certain fields. For example, since the enactment of laws such as Every Study Succeeds Act in 2015, there has been a greater push to use evidence-based practices in the field of special education. Therefore, it has become more imperative that estimates from studies accurately reflect the effectiveness of interventions. SCEDs are used in the field of applied behavior analysis to determine the effectiveness of interventions, or to record functional behavior assessment data, for example. If teachers, behaviorists, and school psychologists rely on SCED research to determine which interventions to use with students, they may be using interventions that are not actually effective if the results are unreliable. Meta-analyses use and rely on primary level summary findings; if inappropriate quantification techniques are used in these primary studies, then biased or misleading results may be obtained at the meta-analytic level as well. Thus, it is imperative that appropriate SCED quantification techniques are used,
and this study highlights that those conducting SCED research may currently be using inappropriate quantification techniques to analyze their data.

**Changing the Way People Analyze SCED Data**

However, the results of this study are encouraging in that this study provides evidence that teachers, practitioners, and researchers are able to improve the way that they understand and analyze SCED data. A formal, synchronous training is not needed to improve the way that people analyze data. This study took a median of 45 minutes to complete, demonstrating that those who want to learn to use the tool outside of the formal study would likely be able to do so in even less time.

This tool has several practical uses. Journal editors could offer the tool to those who submit SCED manuscripts to their journals as a tool to help the researchers choose more appropriate SCED quantification techniques. One participant in this study commented “I have reviewed several single-case experiments that incorporate three or four ESs into their results and I've always suggested that there should be more thoughtfulness and intentionality behind ES selection.” It would be useful if those who review SCED studies, such as the person who wrote this comment, could reference the authors of studies undergoing peer review to the SCED tool so that they can improve their research. The tool can be referenced in this manner throughout the peer review process that many journals require. Such review process might be especially helpful for encouraging researchers to choose quantification techniques more carefully.

This SCED tool could be introduced in college level SCED courses. As the median time to complete the study was 45 minutes, this tool could easily be incorporated into a one or two hour college course. Those who are first learning about SCEDs and different SCED quantification techniques would benefit from this tool. For example, those who are training to
become Board Certified Behavior Analysts are required to take a SCED course during their course sequence; using the tool can expose them to SCED quantification techniques, especially since the field of applied behavior analysis relies heavily on visual analysis instead of quantification techniques.

This tool can be introduced and disseminated at special education, applied behavior analysis, or school psychology conferences (e.g., Association for Behavior Analysis Internationals, American Educational Research Association, etc.). The Institute of Education Sciences hosts an advanced training session each summer for single-case research methods; this is another environment in which the tool can be introduced and disseminated.

This tool could be used for a priori decisions, as recommended by Manolov et al. (2021). Researchers can predict what their data characteristics may look like (such as the number of measurement occasions in the baseline and intervention phase), and then use the tool to help them determine which quantification technique would be most appropriate for use.

Limitations and Future Research

There are several limitations of both this study and the tool that should be acknowledged. This next section will address these limitations, and provide direction of future research to address these limitations.

Limitations of the Study

This study had a small number of participants. A total of 29 people participated in this study, and only 25 were eligible for inclusion for the inferential analyses. Generalizations beyond this sample are limited. For example, although results indicated no differences in change in scores between participants with varying education levels or academic departments, only eleven participants had their Ph.D. and ten indicated having a Masters Degree. Similarly, only
nine participants identified as being a part of the special education department, seven as part of educational psychology department, and six as part of school psychology department. These participants may not be representative of their typical education level or academic department. It is possible that if there are less than 30 participants and if the scores are non-normally distributed, then the sampling distribution will also be non-normal. As a result, the results must be interpreted with caution. To combat this issue, both parametric and non-parametric calculations were conducted. Furthermore, the results of the study are interpreted in context, and both the statistical and practical significance are considered. However, it would be beneficial if this study were repeated with a larger sample to ensure that results are indeed representative and reliable.

It is possible that only those who were previously interested in learning about quantification techniques participated in the study, and so the results may be biased. Several different recruitment methods were used to find participants (i.e., reaching out to SCED field experts, past members of SCED trainings, listservs, etc.), but ultimately those who were interested in learning more about quantification techniques may have been more likely to participate and complete the study.

This study had a high attrition rate. A total of 49 people started the study, but did not complete the study. Of these, 15 people reached the tool and then did not continue the study. It is possible that these 15 participants were no longer motivated to complete the study once they gained access to the tool. Other participants stopped after reaching the first question of the study. These participants may have read the first question of the pretest, and felt intimidated if they did not know the answer, ultimately deciding not to participate in the study. Alternatively, it is possible that other participants felt the study was too easy and did not want to spend time
participating in the study. Some participants may have stopped participating in the study because of the time it took to answer the questions, as the median time to complete the study was 45 minutes. Each of the five pretest and posttest questions required participants to critically think, and this may have caused fatigue. It is possible more participants may have completed the study if there had been an incentive for participation (e.g., names entered into a pool eligible to win a gift card). If this study were conducted again, this could help combat attrition. Furthermore, the study could be implemented at a different time of year, as recruitment occurred at the end of the Fall academic semester, during Christmas/New Years break, when people were likely taking time off from work.

No participants completed the optional calculation (i.e., calculated the estimate), so it was not possible to determine the effect of the tool on ability to perform accurate calculations and interpret these calculations. There are several possible reasons why participants did not complete the quantification technique calculations. As previously discussed, this was a lengthy study, and it is likely that the length contributed to participants’ willingness to complete the calculations. The instructions for the study said that completing the calculations were optional, and so participants likely focused on choosing and justifying the quantification techniques rather than spending time conducting the calculations. Another possible reason participants did not complete the calculations is because they did not know how, even after accessing the tool. The calculators and R codes for each of the quantification techniques were located on the Metric Details tab of the tool. It is possible that participants who did not explore all tabs of the tool missed the section where the calculation formulas and tools are listed, and so they did not conduct the calculations for this reason. However, no comments were provided regarding confusion as to where to find the calculation tools, so it is more likely that participants were fatigued from the study and
simply did not want to complete the optional calculations. It would be beneficial for a future study to focus solely on the effectiveness of the tool in assisting users to conduct calculations correctly, and whether the tool helps users report the resulting estimates correctly.

The time recorded through Qualtrics was not an entirely valid measure for representing the total time spent completing the study. Five participants spent significantly longer completing the study than other participants. For example, Participant 9 spent a total of 151 hours and 3 minutes completing the study; it is likely that this participant had the study open on their computer but were not spending the entire 151 hours reviewing or using the tool. Participants were also able to exit out of the study and complete it at a later time. Other participants may have reached the midpoint of the study, downloaded the tool, and then exited out of the study while they spent time familiarizing themselves with the tool. If this occurred, the time spent towards completing the study would not have been accurately documented. These confounding variables are possible reasons why time was not shown to be related to change in scores between the pretest and posttest.

**Limitations of the Tool**

Although the results of the study are encouraging regarding the effectiveness of the tool, the tool itself has some limitations that should be acknowledged. While the tool contains various methodological research and information for the user to reference, there are circumstances when using the tool alone without referencing the original manuscripts can lead to a limited or inappropriate use of the quantification technique. For example, the tool simply states that generalized least squares regression is affected by outliers (Parker & Hagan-Burke, 2007; Tarlow, 2017). However, the person conducting the analysis could systematically find the outliers, run the analyses both with the outliers and without the outliers, and report these results.
Thus, the person could still effectively use regression when outliers are present. Other quantification techniques such as POGO, which use mean calculations, can use the median instead of the mean so that the estimate is less affected by outliers. Thus, for the most appropriate and effective use of the quantification techniques, the researchers need an in-depth understanding of the quantification techniques and should not rely solely on the information provided in the tool. A caveat was added into the tool on the introduction page, advising users to refer to the original papers and methodological papers, but some users may choose not to do so. Thus, some users may use the tool and end up with a limited understanding of the quantification techniques and how they can be used.

The tool was not designed to provide assistance in interpretations of the underlying scales of the different quantification techniques. Any calculations conducted and the resulting estimates must be interpreted carefully, and assistance in understanding the clinical significance of the estimates is not something that is provided through the use of this tool. Users should interpret any calculation estimates carefully and remember that many of the quantification techniques have unknown underlying scales. Thus, each quantification technique estimate should be interpreted differently (i.e., a .80 for Tau-U is different from a .80 log response ratio estimate). Those who use the tool are encouraged to reference recent and past literature regarding the underlying scales of quantification techniques (e.g., Fingerhut et al., in press-a).

The tool can list several quantification techniques as equally appropriate for use, and the user then needs to consider different factors in order to make an informed decision about which quantification technique is most appropriate for use. The logic code for this tool was only developed based upon four research questions, four data characteristics, and one desired quantification technique feature. If the user has a research question, data characteristic, or desired
quantification technique feature that was not incorporated into the logic code, it can be difficult to determine which quantification technique is most appropriate. For example, while information about autocorrelation for each of the quantification techniques can be found under the Metric Details tab, the user must manually consider this information to make an informed decision about quantification technique appropriateness, since autocorrelation was not a part of the logic code for the tool. The purpose of this study was to first establish the effectiveness of a simplified tool; now that the tool has been established as effective, future developments of the tool can include new research questions, data characteristics, and desired quantification technique features. For example, one of the expert reviewers of the tool recommended for different statistical interpretations besides $p$-value to be added. It was important to first establish that the tool is useful and effective, and now that the results of the study are promising, the tool can be further edited and developed.

The tool incorporates the most commonly used quantification techniques based upon the review of the SCED literature, as well as the input from experts. It is possible that the tool is missing some quantification techniques, especially as the field of SCED continues to grow. However, the tool is posted through OSF https://osf.io/7usbj/, and on the Notes and References page an email address is provided for users to send feedback (SCEDmetrictool@gmail.com). Thus, input of the broader research community is invited and can be used to further improve the tool.

This tool is meant to be used for within-case estimates. While some of the quantification techniques listed in the tool can be used for across-case estimates, there may be better options when conducting across-case estimates. For example, the hierarchical linear modeling technique is not listed in the tool, but can be beneficial for across-case estimates because it takes into
account observations within participants and participants within studies, and the estimates can vary across participants and studies (Van den Noortgate & Onghena, 2003). A caveat was added onto the Notes & References tab of the tool, reminding users that the tool is meant for within-case calculations, but some users may not see this warning and may inappropriately use certain quantification techniques for across-case estimates. The tool can be updated in the future to incorporate these across-case estimate techniques.

Furthermore, as several of the expert reviewers pointed out, the tool is only appropriate for AB-based research designs, as the information listed in the tool assumes that the user is applying the quantification techniques to an AB design. The AB design is the building block of other design types such as multiple-baseline design and reversal designs, so this was an appropriate and logical starting point for testing the tool. However, some of the quantification techniques may be more or less appropriate for other types of designs. Future research can be done to update the tool to consider other types of designs as well.

The tool is currently only available in the form of the downloadable Excel document. Those who do not have Excel, or later versions of Excel, cannot use the tool on their computer. It would be helpful to develop the tool into a website or R package for users to access and use. The tool could be converted into an R code or online calculator that can be easily accessed online and does not need to be downloaded for use.

**Future Research: Summary**

This study could be replicated with a larger sample. Prior SCED experience of participants could be a demographic question presented to participants, so that inferential results can be conducted to determine if there is a statistically significant relationship between length of SCED experience in years and the effectiveness of the tool. A more reliable measure for time to
complete the study could be used to further determine the relationship between time to complete the study and change in scores between pretest and posttest. It would be helpful to conduct the study during a time of year when people are not off from work (e.g., during the academic semester). Furthermore, future studies can focus on the effect of the tool on participants’ ability to calculate the estimate correctly and interpret the estimate as well.

There are several ways in which the tool can be updated and made more complex, and it would be helpful to conduct a formal study to determine if the tool is still helpful after it is made more complex. For example, more research questions (e.g., research questions about gradual effects or delayed effects) and data characteristics (e.g., autocorrelation) can be added into the tool and logic code. The tool can also be edited so that it can be used for between-case estimates and with various research designs. It also can be converted into an R code or website that can be more easily accessed. After the tool is updated and changed, it would be helpful to conduct another study to determine if it is still effective and helpful.

**Conclusion**

This dissertation study had two aims: a.) to create a tool to help users choose and justify their use of SCED quantification techniques and b.) to test the effectiveness of such a tool. Results of the study show the tool improves the way that users choose and justify their use of SCED quantification techniques. The majority of participants found the tool to be helpful for both choosing and justifying SCED quantification techniques, and would be likely to use the tool in the future. These are promising findings and signify the importance of such a tool in fields that utilize SCEDs for research. The tool is already having an effect on the field of SCED, as several of the expert reviewers of the tool requested to use and incorporate it into their graduate level SCED courses. The tool is open sourced and freely accessible; the link to download the tool is
https://osf.io/7usbj/. The field of SCED will continue to grow and improve as SCED research and SCED tools, such as this tool, are disseminated into the field.
References


---

21 Fingerhut and Xu are jointly first authors.

https://doi.org/10.1177%2F0198742918809341


https://doi.org/10.3102/10769986006002107


https://doi.org/10.1002/jrsm.1052


Manolov, R. (2013). *MPD.R* [R Syntax]

https://www.dropbox.com/s/nky75oh40f1gbwh/MPD.R?dl=0

Manolov, R. (2015a). *GLS.R* [R Syntax]

https://www.dropbox.com/s/dni9qq5pqiqi3pc23/GLS.R?dl=0

https://www.dropbox.com/s/wt1qu6g7j2ln764/MBLR.R?dl=0.


https://www.dropbox.com/s/k57dj32gyit934g/PZD.R?dl=0%20


https://doi.org/10.1177/0145445516664307


https://doi.org/10.1016/j.beth.2016.04.008


http://www.ub.edu/gcai/soft/Software_resources_for_applied_researchers_April2015.pdf


https://link.springer.com/content/pdf/10.3758/BRM.41.4.1262.pdf
Manolov, R., & Solanas, A. (2013). A comparison of mean phase difference and generalized
https://doi.org/10.1016/j.jsp.2012.12.005

https://doi.org/10.1016/j.beth.2010.12.003

Manolov, R., & Vannest, K. J. (2019). A visual aid and objective rule encompassing the data
features of visual analysis. *Behavior Modification, 1*-32.
https://doi.org/10.1177/0145445519854323

macro. http://faculty.ucmerced.edu/wshadish/software/software-meta-analysis-single-
case-design/dhps-version-march-7-2015

riders, and to what extent? Evaluating the effectiveness of a New Zealand therapeutic
riding intervention using a single-case experimental design. *Annals of Physical and
Rehabilitation Medicine, 61*, 522-523. https://doi.org/10.1016/j.rehab.2018.05.1216

research designs for evidence-based practice. *Research in Autism Spectrum Disorders, 6*,

of single-case charts: An addendum. *Occupational Therapy Journal of Research, 10*, 208-
220. https://doi.org/10.1177/0145445515581327


https://doi.org/10.1177/00224669070400040101

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

https://doi.org/10.1177/001440290907500201

https://doi.org/10.1016/j.beth.2010.08.006

https://doi.org/10.1177/0145445511399147


https://jepusto.shinyapps.io/scdhlm/

https://doi.org/10.1016/j.jsp.2018.02.003

https://doi.org/10.1037/met0000179


http://www.rstudio.com


https://doi.org/10.1002/jcad.12038


https://doi.org/10.1016/j.ridd.2018.04.029

https://ies.ed.gov/ncee/wwc/handbooks


### Appendix 3.1

#### Logic Code

<table>
<thead>
<tr>
<th>Form Questions</th>
<th>Answer(s)</th>
<th>Percentage of data exceeding the median trend (PEM-T)/Extended celeration line (ECL)</th>
<th>Improvement rate difference (IRD)</th>
<th>No overlap of all pairs (NAP)</th>
<th>Percentage of all non-overlapping data (PAND)</th>
<th>Percentage of data points exceeding the median of the baseline phase (PEM)</th>
<th>Percentage of zero data (PZD)</th>
<th>Tau-U</th>
<th>Baseline Corrected Tau-U (Tarlow, 2017)</th>
<th>Tau-U Trend A (Parker, Vannest, Brown, et al., 2011)</th>
<th>“Hedges g for single-case”</th>
<th>Log response ratio (LRR)</th>
<th>Mean phase difference (MPD)</th>
<th>Generalized least squares regression (GLS)</th>
<th>Percent of goal obtained (POGO)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overlap/ Data separation between phases</strong></td>
<td>Yes</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Magnitude of change in level</strong></td>
<td>Yes</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

1. **Overlap/ Data separation between phases** ("What is the percentage of non-overlap between phases?"): Yes/No
2. **Magnitude of change in level** ("What is the magnitude of the change in level between baseline and treatment phase?"): Yes/No
<table>
<thead>
<tr>
<th>Data/ Graph Characteristic</th>
<th>Magnitude of change in slope (&quot;What is the magnitude of change in slope between baseline and treatment?&quot;)</th>
<th>Magnitude of change in variability (&quot;What is the magnitude of change in variability between the baseline and treatment phase?&quot;)</th>
<th>Small number of data points in the baseline and/or treatment phase (e.g., less than 5 data points in the baseline phase or treatment phase)</th>
</tr>
</thead>
<tbody>
<tr>
<td>between baseline and treatment phase?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Magnitude of change in slope (&quot;What is the magnitude of change in slope between baseline and treatment?&quot;)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Magnitude of change in variability (&quot;What is the magnitude of change in variability between the baseline and treatment phase?&quot;)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Small number of data points in the baseline and/or treatment phase (e.g., less than 5 data points in the baseline phase or treatment phase)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

- **Yes**: 1
- **No**: 1
- **1**: 1
- **15**: 1
- **20**: 1
- **5**: 1
<table>
<thead>
<tr>
<th>Desired Metric Feature</th>
<th>Able to calculate a p-value and/or provide confidence intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>No</td>
</tr>
</tbody>
</table>

| and/or treatment phase (e.g., less than 5 data points in the baseline phase or treatment phase) | Yes | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 5 | 3 | 1 | 1 | 5 | 5 | 1 | 5 |
|----------------------------------------------------------------------------------------------|-----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| Baseline trend (e.g., visually clear baseline trend, statistically significant baseline trend) | No  | 1 | 5 | 5 | 5 | 5 | 5 | 5 | 1 | 1 | 5 | 5 | 5 | 5 | 5 | 5 |
| Outliers present (e.g., data point that lies more than 1.5 times the interquartile range)   | Yes | 1 | 3 | 3 | 1 | 1 | 1 | 3 | 3 | 3 | 1 | 1 | 1 | 3 | 3 | 3 |
| Outliers present (e.g., data point that lies more than 1.5 times the interquartile range)   | No  | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| Large variability (e.g., 80% of data within stability envelope)                              | Yes | 3 | 1 | 3 | 1 | 1 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| Large variability (e.g., 80% of data within stability envelope)                              | No  | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
Appendix 3.2

Screenshots of the Excel Tool

Instructions (Sheet 1)

Instructions

Link for video instructions (users are encouraged to watch the instructional video before using this tool) https://www.loom.com/share/b941b3c24352492b9c0f9b6550a2ea3c

The purpose of this tool is to 1.) help users determine the appropriate single-case metric to use for their data set depending on their research question(s) and data/graph characteristic(s), 2.) help users justify their use of the metric by providing the user with information (e.g., original sources, findings from methodology papers, etc.) about why different metrics are appropriate or not, and 3.) provide users with tools to perform correct metric calculations.

The Tool tab shows research questions, data/graph characteristics, and desired metric features that affect which single-case metrics are appropriate for use. The user checks “Yes” or “No” depending on what needs to be accounted for when picking a metric. For example, if there is baseline trend in the graph, the user would select “Yes” for “Baseline trend” (located in the grey “Data/ Graph Characteristic(s)” section); because baseline trend is present, the user needs a metric that can account for baseline trend. The metrics recommended for use (see: Ranked Metrics [Output]) are automatically updated as the user changes the Input, and ranked from most recommended to least recommended. The user can click on any of the metric names, located at the top, or more simply the Metric Details tab at the bottom, and read more about the metric (e.g., metric founders, how the metric is calculated, tools to calculate the metric, and examples of the metric in applied literature).

Please note that this tool is meant to be used as a guide only. Readers are highly encouraged to reference the original papers of the metrics, as well as methodological research, to further determine if a metric is/isn’t appropriate for use with their data. See Notes and References tab for more details.
<table>
<thead>
<tr>
<th>Input</th>
<th>Ranked Metrics [Output]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Change the User Input to &quot;Yes&quot; or &quot;No&quot; depending on your own research questions, data/graph characteristics, and desired metric feature(s)</td>
<td>Metrics listed in order of appropriateness, depending on User Input. Most appropriate metrics are ranked &quot;1&quot;, followed by the 2nd most appropriate metrics &quot;2&quot;, etc. Any metrics found within the same ranking (e.g., &quot;1&quot;) are equally appropriate, and are listed in alphabetical order with non-overlapping indices listed first, standardized mean difference second, regression third, and log-response ratio metrics fourth.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Possible research questions, data/graph characteristics, and desired metric features</th>
<th>User Input</th>
<th>Mean phase difference (MPD)</th>
<th>Generalized least squares regression (GLS)</th>
<th>Percent of goal obtained (POGO)</th>
<th>&quot;Hedges g for single-case&quot;</th>
<th>Mean baseline reduction (MBLR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research Question(s)</td>
<td>Multiple research questions can be selected</td>
<td>Yes</td>
<td>Cannot answer the research question.</td>
<td>Cannot answer the research question.</td>
<td>Cannot answer the research question.</td>
<td>Cannot answer the research question.</td>
</tr>
<tr>
<td>Magnitude of change in level (What is the magnitude of the change in level between baseline and treatment phase?)</td>
<td>No</td>
<td>Can answer the research question (overall difference: as cited in Manolios &amp; Meizoor, 2017).</td>
<td>Can answer the research question.</td>
<td>Can answer the research question.</td>
<td>Can answer the research question.</td>
<td>Can answer the research question.</td>
</tr>
<tr>
<td>Magnitude of change in slope (What is the magnitude of change in slope between baseline and treatment phase?)</td>
<td>No</td>
<td>Can answer the research question (overall difference: as cited in Manolios &amp; Meizoor, 2017).</td>
<td>Preferable to MPD for estimating slope change.</td>
<td>Efficient estimation.</td>
<td>Can answer the research question.</td>
<td>Can answer the research question.</td>
</tr>
<tr>
<td>Magnitude of change in variability (What is the magnitude of change in variability between the baseline and treatment phase?)</td>
<td>No</td>
<td>Cannot provide effect size parameter, but can measure the heterogeneity of variance between phases.</td>
<td>Cannot provide effect size parameter, but can measure the heterogeneity of variance between phases.</td>
<td>Cannot answer the research question.</td>
<td>Cannot answer the research question.</td>
<td>Cannot answer the research question.</td>
</tr>
<tr>
<td>Number of data points in the baseline and/or treatment phase (e.g., less than 2 data points in the baseline phase or treatment phase)</td>
<td>No</td>
<td>Results depend on phase length. When treatment phase is longer than baseline phase, worse performance.</td>
<td>Results depend on phase length. When treatment phase is shorter than baseline phase, worse performance.</td>
<td>Manolios &amp; Solanas, 2013.</td>
<td>Manolios &amp; Solanas, 2013.</td>
<td>—</td>
</tr>
<tr>
<td>Power increases as number of measurement occasions increase.</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Input</td>
<td>Ranked Metrics (Output)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Possible research questions, data/graph characteristics, and desired metric features</td>
<td>Metrics listed in order of appropriateness, depending on User Input. Most appropriate metrics are ranked “1”, followed by the 2nd most appropriate metrics (”2”), etc. Any metrics found within the same ranking (e.g., “1”) are equally appropriate, and are listed in alphabetical order with non-overlap indices listed first, standardized mean difference second, regression third, and log-response ratio metrics fourth.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User Input</th>
<th>Mean phase difference (MPD)</th>
<th>Generalized least squares regression (QLS)</th>
<th>Percent of goal obtained (POGO)</th>
<th>“Hedges g for single-case”/HPS-d</th>
<th>Mean baseline reduction (MBLR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Can account for baseline trend. Trends assumed to be linear and extend into treatment.</td>
<td>Can account for baseline trend. Trends modeled more accurately than MPD (Slonim, 2014). Trend is assumed to be linear and extends into treatment.</td>
<td>Can account for baseline trend. Assumes that trend stabilizes over time.</td>
<td>Cannot account for baseline trend.</td>
<td>Cannot account for baseline trend.</td>
</tr>
<tr>
<td>No</td>
<td>Affected by outliers (Maggin et al., 2013).</td>
<td>Affected by outliers (Slonim, 2014).</td>
<td>Relatively unaffected by outliers if median is used for calculation instead of mean.</td>
<td>Affected by outliers (Maggin et al., 2013).</td>
<td>Affected by outliers (Kloke et al., 2019).</td>
</tr>
<tr>
<td>No</td>
<td>Cannot calculate p-value, lacks known sampling distribution.</td>
<td>Can calculate a p-value.</td>
<td>Cannot calculate p-value, lacks known sampling distribution.</td>
<td>Can calculate a p-value and confidence intervals can be obtained.</td>
<td>Cannot calculate p-value, lacks known sampling distribution.</td>
</tr>
</tbody>
</table>

*Multiple defined metric features can be selected.*
## Metric Details (Sheet 3)

### Ranked Metrics Details

<table>
<thead>
<tr>
<th>Rank</th>
<th>Metric Details</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mean phase difference (MPD)</td>
<td>Generalized least squares regression (GLS)</td>
</tr>
<tr>
<td>2</td>
<td>&quot;Hedges g for single-case&quot; / HPS-d</td>
<td>Mean baseline reduction (MBR)</td>
</tr>
</tbody>
</table>

#### Original reference

<table>
<thead>
<tr>
<th>Rank</th>
<th>Metric Details</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mean phase difference (MPD)</td>
<td>Matlovic &amp; Deferasi (2012)</td>
</tr>
<tr>
<td>2</td>
<td>&quot;Hedges g for single-case&quot; / HPS-d</td>
<td>Hedges et al. (2002, 2013)</td>
</tr>
</tbody>
</table>

#### How it is calculated

**1.** Calculate the trend in the baseline phase and project into the intervention phase.

**2.** Compare projected baseline phase trend with actual intervention data and calculate the average difference.

**1.** Determine the model. An example of a basic model is: \( y_t = b_0 + b_1 t \), with \( y \) representing the outcome scores, \( b_0 \) indicating the baseline intercept, and \( b_1 \) indicating the change in level between baseline and intervention. This model can be expanded (e.g., autocorrelation can be modeled into the equation).

**2.** Use the calculator or software tool to run the analysis.

**For desired behavior increase:**

1. Subtract the expected level of behavior without the intervention from the obtained level of behavior.
2. Divide the outcome from step 1 by the goal level of behavior minus the expected level of behavior without the intervention.
3. Multiply by 100.

**For desired behavior decrease:**

1. Subtract the obtained level of behavior from the expected level of behavior without the intervention.
2. Divide the outcome from step 1 by the expected level of behavior without intervention minus the goal level of behavior.
3. Multiply by 100.

#### Calculators/Sources/Tools

<table>
<thead>
<tr>
<th>Rank</th>
<th>Calculator</th>
<th>Source</th>
</tr>
</thead>
</table>
| 1    | Matlovic (2015a) | [Website](https://matlovnic.github.io/)
| 2    | Matlovic (2015b) | [Website](https://matlovnic.github.io/MPDRegression/) |

Other:

- SPS Macro (Matrov & Shadish, 2015) [Website](https://www.spsmacro.com)
- Meta-Analyses of Single-Case Design [Website](https://www.spsmacro.com)
<table>
<thead>
<tr>
<th>Rank</th>
<th>Metric Details</th>
<th>Other Information</th>
<th>Examples in the Literature</th>
</tr>
</thead>
</table>
| 1    | Mean phase difference (MPD) | Robust to autocorrelation (Tofte, 2017). Recommended for practitioners due to ease of interpretability (Tofte, 2017). Effects positively skewed (Tofte, 2017). Correctly estimated when there was no effect. No floor or ceiling effects (Tofte, 2017). Can be standardized by dividing by the baseline or pooled standard deviation. | Kappath et al. (2019)  
http://dx.doi.org/10.1037/a0000351  
**Hierarchical linear modeling analysis technique, extension of GLM** |
https://doi.org/10.1177/0749835217729500  
**Between-case standardized mean difference** |
Notes and References (Sheet 4)

Metric notes:
This tool focuses on GLS, but ordinary least squares regression (OLS; e.g., Huitema & McKean, 1998) can also be used whenever GLS can be used. OLS cannot account for autocorrelation. Furthermore, this tool does not include piecewise regression (Center et al., 1985-1986) or slope and level change (Solanas et al., 2010); readers are advised to reference the original sources to learn more about these regression-based techniques.

Citations found within tool:
All information included within the “Tool” and “Metric Details” tab can be found in the original reference(s) for that metric, unless otherwise specified.

Citing this tool:

For more details regarding any of the above information, questions, or general comments, please send emails to scedmetricctool@gmail.com.

References


Appendix 3.3

Screenshots of Examples Document

Instructions

Link for video instructions (users are encouraged to watch instructional video)

https://www.loom.com/share/83739923c48904a6b2316dec9d3a8f56

Below are 5 AB graphed data sets, along with the research question(s) found below.

Clicking “Show the answer” reveals the research question, data characteristics and desired metric features that should be marked “Yes” in the tool input. This also shows the metrics that are recommended for use, along with an example justification for using that metric.

The user is encouraged to write their own answers in the User response section, and then compare their answers with the Example of correct answer. This is meant to help improve appropriate metric selection, appropriate metric justification, and correct calculations.

Users are advised that it is unlikely all researchers would reach a consensus regarding the most appropriate metric or exact formulation of an appropriate justification. This is meant to guide usage of single-case metrics and not serve as showing a “correct answer”.

Example 1

<table>
<thead>
<tr>
<th>Phase</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>8</td>
</tr>
<tr>
<td>A</td>
<td>9</td>
</tr>
<tr>
<td>B</td>
<td>5</td>
</tr>
<tr>
<td>B</td>
<td>6</td>
</tr>
<tr>
<td>B</td>
<td>12</td>
</tr>
<tr>
<td>B</td>
<td>6</td>
</tr>
<tr>
<td>B</td>
<td>9</td>
</tr>
<tr>
<td>B</td>
<td>12</td>
</tr>
<tr>
<td>B</td>
<td>9</td>
</tr>
</tbody>
</table>

What is the magnitude of change in level between phase A and B?

Considerations for picking the appropriate metric, rationales and calculations

<table>
<thead>
<tr>
<th>Consideration</th>
<th>User response</th>
<th>Example of correct answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research question:</td>
<td></td>
<td>Magnitude of change in level</td>
</tr>
<tr>
<td>Data characteristics:</td>
<td></td>
<td>Less than 5 data points in baseline, large variability</td>
</tr>
<tr>
<td>Desired metric features:</td>
<td></td>
<td>N/A</td>
</tr>
<tr>
<td>Metrics for possible use:</td>
<td></td>
<td>Log Response Ratio (LRR)</td>
</tr>
<tr>
<td>Example justification and calculation:</td>
<td></td>
<td>LRR is used because it can provide an unstandardized outcome, which may be easier for practitioners to interpret outcomes. LRR can express the magnitude of the effect. It also is related to the percentage change, which is often used and understood by practitioners (Pustejovsky, 2018). LRR was calculated using the SingleCaseES package (Pustejovsky &amp; Swan, 2019). There is a 47% increase in scores.</td>
</tr>
</tbody>
</table>
Appendix 3.4

Recruitment Email

Dear______

My name is Joelle Fingerhut, and I am a PhD student in the department of Counseling and Educational Psychology at the University at Albany, SUNY. As part of my dissertation, I am testing the effectiveness of an Excel tool, which aims to help researchers pick and rationalize their use of techniques to analyze single-case design research data. I am looking for participants to test this Excel tool, including graduate students who analyze and conduct single-case research. I would appreciate if you could share the following link (which includes all materials including the electronic consent form) with potential participants, letting them know they are not obligated to participate.

Thank you for the consideration. Please let me know if you have further questions.

Best,

Joelle Fingerhut
Appendix 3.5

Pretest and Posttest Graphs

1.

Research Question: Does the intervention have an effect on scores, with effect defined as magnitude of change in level?

**Essential Graph Components:**

- Research question: magnitude of change in level
- Data characteristics: small number of measurement occasions in baseline phase or treatment phase
Research Question: What is the magnitude of the change in slope?

*Essential Graph Components:*
- Research question: magnitude of change in slope
- Data characteristics: baseline trend
Research Question: What is the magnitude of the change in variability between phases A and B?

**Essential Graph Components:**
- Research question: magnitude of change in variability
- Data characteristics: variability
Research Question: How much data overlap or nonoverlap is there between phases A and B?

**Essential Graph Components:**
- Research Question: non-overlapping data
- Data characteristics: outliers
Research Question: What is the magnitude of change in level between baseline and intervention? (p<.05)

_Essential Graph Components:_
- Research question: magnitude of change in level
- Desired metric feature: \( p \)-value
Appendix 3.6

Demographic Survey

1. Are you interested in quantifying effects from a SCED?
   a. No, I am not interested in quantifying effects from a SCED (e.g., I use visual analysis only)
   b. Yes, I am interested in quantifying effects from a SCED

2. Have you quantitatively analyzed a single-case graph either with support (e.g., with help from an instructor or colleague) or independently within the last year?
   a. Yes
   b. No

3. What is your highest level of education obtained?
   a. High School/GED
   b. Associates Degree
   c. Bachelors Degree
   d. Masters Degree
   e. PhD

4. What is the name of the academic department that you best associate yourself with? (e.g., statistics, special education, educational psychology, etc.).

5. How confident are you in your own ability to pick an appropriate single-case statistical technique?
   a. Not confident at all
   b. Somewhat unconfident
   c. Somewhat confident
   d. Very confident

6. How confident are you in your own ability to justify your use of statistical analysis techniques?
   a. Not confident at all
   b. Somewhat unconfident
   c. Somewhat confident
   d. Very confident

7. Which statement is most accurate?
   a. I have published SCED-related research
   b. I have not yet published SCED-related research, but would like to do so in the future
   c. I do not intend to publish SCED-related research
Appendix 3.7

Social Validity Survey

*Please circle the answer that you believe is most accurate.*

1. How likely would you be to use this Excel tool in the future for your own research?
   a. Not at all likely,
   b. Somewhat unlikely
   c. Somewhat likely
   d. Very likely

2. How helpful do you rate the Excel tool to be in picking an appropriate SCED metric?
   a. Not at all helpful
   b. Somewhat unhelpful
   c. Somewhat helpful
   d. Very helpful

3. How helpful do you rate the Excel tool to be in providing a rationale for using a SCED metric?
   a. Not at all helpful
   b. Somewhat unhelpful
   c. Somewhat helpful
   d. Very helpful

Other additional comments:
# Appendix 3.8

## Expert Reviewer Comments

<table>
<thead>
<tr>
<th><strong>Recommendations from Reviewer 1</strong></th>
<th><strong>Response</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Add explanation about the division between sections 1 and 2</td>
<td>Further explanation added: “Any metrics found within the same ranking (e.g., &quot;1&quot;) are equally appropriate, and are listed in alphabetical order with non-overlap indices listed first, standardized mean difference second, regression third, and log-response ratio metrics fourth.”</td>
</tr>
<tr>
<td>Unclear what general effect is</td>
<td>Removed this research question from the tool</td>
</tr>
<tr>
<td>Advises to incorporate a priori use throughout</td>
<td>Do not want to overcomplicate the tool; will leave out for now and can add back into an updated version</td>
</tr>
<tr>
<td>Clarify definition of variability</td>
<td>Clarified: “Large variability (e.g., 80% of data within stability envelope)”</td>
</tr>
<tr>
<td>-add information about confidence intervals</td>
<td>Added information about confidence intervals for each of the metrics under “desired metric features”</td>
</tr>
<tr>
<td>-refer users to BC-SMD</td>
<td>Added fourth tab, “Notes and References”, with the following information: “Although many of the metrics listed in the tool can be used across-cases, this tool is meant to be used for within-case estimates. This is why HPS-d metric (Hedges et al., 2012, 2013) is included, but not the between-case standardized mean difference (Pustejovsky et al., 2014). Furthermore, generalized least squares regression (GLS; Swaninathan, Roger, Horner, 2014) is included in the tool instead of the hierarchical linear modeling across-case estimation technique (HLM; Van den Noortgate &amp; Onghena, 2003a, 2003b). Users are highly encouraged to reference the original papers to determine if the metric is appropriate for estimation across cases.”</td>
</tr>
<tr>
<td>Clarify notation of Tau metric</td>
<td>Change notation of Taus throughout (e.g., “baseline corrected Tau-U” for Tarlow (2017) variant.</td>
</tr>
<tr>
<td>Clarify classification of HPS-d as able to answer question about change of variability</td>
<td>Clarified with statement: Cannot provide effect size parameter, but can measure the heterogeneity of variance between phases.</td>
</tr>
<tr>
<td>Recommendations from Reviewer 1</td>
<td>Response</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>----------</td>
</tr>
<tr>
<td>Update information about how GLS is calculated</td>
<td>More details added to GLS: “An example of a basic model is: $y_t = \beta_0 + \beta_1 t$, with $y_t$ representing the outcome (scores), $\beta_0$ indicating the baseline intercept, and $\beta_1 t$ indicating the change in level between baseline and intervention. This model can be expanded (e.g., autocorrelation can be modeled into the equation).”</td>
</tr>
<tr>
<td>Recommendations from Reviewer 1</td>
<td>Response</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>----------</td>
</tr>
</tbody>
</table>
| Review the following literature for more information about nonoverlap methods:  
<table>
<thead>
<tr>
<th><strong>Recommendations from Reviewer 1</strong></th>
<th><strong>Response</strong></th>
</tr>
</thead>
</table>
| Speech-Language Pathology, 27(1S), 495-503.  
[https://doi.org/10.1044/2017_AJSLP-16-0197](https://doi.org/10.1044/2017_AJSLP-16-0197) |  
Psychological Methods, 24(2), 217-235. [https://doi.org/10.1037/met0000179](https://doi.org/10.1037/met0000179) |
|  |  
|  |  
|  |  
|  |  
Clarify that no two researchers would reach a consensus for the examples provided on Examples document

Sentence added on Examples document: “Users are advised that it is unlikely all researchers would reach a consensus regarding the most appropriate metric or exact formulation of an appropriate justification. This is meant to guide usage of single-case metrics and not serve as showing a "correct answer".”
<table>
<thead>
<tr>
<th>Recommendations from Reviewer 1</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1- wouldn’t use GLS</td>
<td>Coding of metrics reviewed and edited</td>
</tr>
<tr>
<td>Example 2- would use piecewise regression</td>
<td>Sentence added that piecewise is preferable, though not covered in the tool: “<strong>Although not listed in the Excel Tool, a piecewise regression model (Center et al., 1985-1986) is appropriate for use as it can provide a separate quantification of the immediate change in level and change in trend</strong>”</td>
</tr>
<tr>
<td>Example 3- would use MLBR</td>
<td>All examples edited, as research question about general effect and gradual effect were removed from Excel tool</td>
</tr>
<tr>
<td>Example 4- would use piecewise regression</td>
<td>Sentence added that piecewise is preferable, though not covered in the tool: “<strong>Although not listed in the Excel Tool, a piecewise regression model (Center et al., 1985-1986) is appropriate for use as it can provide a separate quantification of the immediate change in level and change in trend</strong>”</td>
</tr>
<tr>
<td>Example 5: would need to pick another measure if non-overlap picked</td>
<td>All examples edited, as research question about general effect and gradual effect were removed from Excel tool</td>
</tr>
<tr>
<td>Need to separate data characteristics and desired metric features into two dimensions on Examples file</td>
<td>“data characteristics” and “desired metric features” split into two dimensions</td>
</tr>
<tr>
<td><strong>Recommendations from Reviewer 2</strong></td>
<td><strong>Response</strong></td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Clear definition of trend needed: linear trend only</td>
<td>Clarified: “trend is assumed to be linear...”</td>
</tr>
<tr>
<td>Clarify question about change in variability</td>
<td>Clarified with statement: “Cannot provide effect size parameter, but can measure the heterogeneity of variance between phases.”</td>
</tr>
<tr>
<td>Describe tool carefully; not meant to be a complete answer for picking a metric</td>
<td>Statement added throughout: “Please note that this tool is meant to be used as a guide only. Readers are highly encouraged to reference the original papers of the metrics, as well as methodological research, to further determine if a metric is/isn’t appropriate for use with their data. See Notes and References tab for more details.”</td>
</tr>
<tr>
<td>Unsure what definition of general effect is</td>
<td>Research question deleted from tool</td>
</tr>
<tr>
<td>Unsure what definition of gradual effect is</td>
<td>Research question deleted from tool</td>
</tr>
<tr>
<td>Recommendations from Reviewer 3</td>
<td>Response</td>
</tr>
<tr>
<td>--------------------------------------------------------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Confused about recommendation from Example 3</td>
<td>Research question adjusted, as general effect and gradual effect research question removed from tool</td>
</tr>
<tr>
<td>Example 4- MPD would not be appropriate</td>
<td>Coding of metrics reviewed and edited</td>
</tr>
<tr>
<td>Example 5- all techniques listed as appropriate</td>
<td>Coding of metrics reviewed and edited</td>
</tr>
<tr>
<td>Issues when trying own research questions</td>
<td>Coding of metrics reviewed and edited</td>
</tr>
<tr>
<td>Some examples in the literature that are provided are better than others</td>
<td>Reviewed all of the examples and added new ones where appropriate</td>
</tr>
<tr>
<td>Recommendations from Reviewer 4</td>
<td>Response</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>----------</td>
</tr>
<tr>
<td>Confusion over “general effect”</td>
<td>Research question removed from tool</td>
</tr>
<tr>
<td>Can all metrics be used across design types?</td>
<td>Clarified that meant to be used with AB and building blocks of AB: “This tool is to be used with AB-based research designs (multiple-baseline design, withdrawal design, alternating treatment design, etc.).”</td>
</tr>
<tr>
<td>Hide metrics ranked outside of 1</td>
<td>Explained to reviewer that tool shouldn’t be used as a definite decision, so appropriate to leave all choices</td>
</tr>
<tr>
<td>Use more complex design examples in Examples document</td>
<td>Can add at later date, but to test the tool we will use AB only</td>
</tr>
<tr>
<td><strong>Recommendations from Reviewer 5</strong></td>
<td><strong>Response</strong></td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Make clear that video instructions contain more details than the text instructions</td>
<td>Clarified: “Link for video instructions (users are encouraged to watch the instructional video before using this tool)”</td>
</tr>
<tr>
<td>Clarify why GLS is recommended by not OLS</td>
<td>Clarified with following statement: “This tool focuses on GLS, but ordinary least squares regression (OLS; e.g., Huitema &amp; McKean, 1998) can also be used whenever GLS can be used. OLS cannot account for autocorrelation. Furthermore, this tool does not include piecewise regression (Center et al., 1985-1986) or slope and level change (Solanas et al., 2010); readers are advised to reference the original sources to learn more about these regression-based techniques.”</td>
</tr>
</tbody>
</table>
Specify how the data characteristics are defined

Statements added to Notes and References tab: “Users are advised that each of the categories within data/graph characteristics can be defined in different ways. Users are advised to reference the original reference papers and/or methodological papers cited throughout the tool to determine if the metric is appropriate for use.

Small number of data points in baseline/treatment phase: This is defined this as “5 data points or less in the baseline or treatment phase”, as the recommendation from What Works Clearinghouse (2020) is a minimum of 5 data points per phase for reversal, multiple-baseline and alternating treatment designs.

Baseline trend: This is defined this as a visually clear baseline trend and/or statistically significant baseline trend. Users are advised to reference the original papers and/or methodological papers to determine how trend is defined to determine if the metric is appropriate for use (e.g., Tarlow [2017] recommends using Baseline Corrected Tau-U if monotonic baseline trend is statistically significant [see pg. 437]).

Outliers present: This is defined as a data point that lies more than 1.5 times the interquartile range. It could also be defined as any data points that are a certain number of standard deviations away from the mean (e.g., 3 standard deviations). Readers are advised to reference the original papers and/or methodological papers to determine how outliers are defined and whether the metric is appropriate for use.

Large variability: This is defined as 80% of data within a stability envelope. Variability does not need to be homogeneous across phases. Users are advised to reference the original papers and/or methodological papers to determine how variability is defined and whether the metric is appropriate for use.”
<table>
<thead>
<tr>
<th>Remove research question about general effect and gradual effect</th>
<th>Both questions removed from Excel tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clarify GLS explanation</td>
<td>More details added to GLS:</td>
</tr>
<tr>
<td></td>
<td>“An example of a basic model is:</td>
</tr>
<tr>
<td></td>
<td>( y_t = \beta_0 + \beta_1 t ), with ( y_t ) representing the outcome (scores), ( \beta_0 ) indicating the baseline intercept, and ( \beta_1 t ) indicating the change in level between baseline and intervention. This model can be expanded (e.g., autocorrelation can be modeled into the equation).</td>
</tr>
<tr>
<td>Recommendations from Reviewer 6</td>
<td>Response</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>----------</td>
</tr>
<tr>
<td>Clarify that tool is for within-case measures</td>
<td>Clarified as follows: “Although many of the metrics listed in the tool can be used across-cases, this tool is meant to be used for within-case estimates. This is why HPS-d metric (Hedges et al., 2012, 2013) is included, but not the between-case standardized mean difference (Pustejovsky et al., 2014). Furthermore, generalized least squares regression (GLS; Swaninathan, Roger, Horner, 2014) is included in the tool instead of the hierarchical linear modeling across-case estimation technique (HLM; Van den Noortgate &amp; Onghena, 2003a, 2003b). Users are highly encouraged to reference the original papers to determine if the metric is appropriate for estimation across cases.”</td>
</tr>
<tr>
<td>Statistical inferences that are included in tool are narrow, need to expand beyond p-value</td>
<td>Confidence interval information added under “desired metric features”. Other statistical inferences to be added into tool after it has been tested.</td>
</tr>
</tbody>
</table>
**Common recommendations across experts:**

- Explain tool and how it can be used carefully (Reviewers 2 and 5)
- Explain that tool is meant for within-case estimates (Reviewers 1 and 6)
- Explain definition of data characteristics (Reviewers 1, 2, and 5)
- Confusion over research question “general effect” (Reviewers 1, 2, 4, and 5)
- Confusion over research question “gradual effect” (Reviewers 2, 4 and 5)