Research in Developmental Disabilities 85 (2019) 31–41

Contents lists available at ScienceDirect

Research in Developmental Disabilities

journal homepage: www.elsevier.com/locate/redevdis

Comparison of visual analysis, non-overlap methods, and effect sizes in the evaluation of parent implemented functional assessment based interventions

Erin E. Barton[a,*], Hedda Meadan[b], Angel Fettig[c]

[a] Vanderbilt University, 230 Appleton Place, Peabody 228, Nashville, TN 37204, United States
[b] University of Illinois at Urbana-Champaign, United States
[c] University of Massachusetts Boston, United States

ARTICLE INFO

Keywords:
Single case research' visual analysis
Parent-implemented interventions

ABSTRACT

We used an existing body of research (i.e., parent implemented functional-assessment based interventions) to examine visual analysis features and processes and evaluate the reliability of two frequently used non-overlap indices (NAP & Tau-U) and a novel effect size index—the between-case standardized mean difference (BC-SMD). Results indicated that visual analysis terms and procedures were inconsistently used across studies. Further, there was limited agreement between the non-overlap indices and independent visual analysis. Results regarding the BC-SMD were inconclusive given only 5 of the 15 studies were eligible for analyses for different dependent variables. Our results suggest that visual analysis standards are needed by which single case researchers analyze and report their results. Further, additional research is needed refining SCR effect sizes, which can be used to describe the magnitude of change within and across SCR studies with functional relations.

What this paper adds?

We identified disagreements between visual analysis and commonly used overlap methods in estimating the magnitude of effects in single case studies. The availability of an effect size that is similar to the effect sizes used group design studies might facilitate the use of single case design research in meta-analytic reviews and be important for the field.

1. Introduction

In single case research (SCR), replication is used to establish plausible causal inference through repeated demonstrations of an effect that coincides with the manipulation of the independent variable. Data patterns are examined within and across adjacent conditions; when the data in one condition differs from what is predicted based on the preceding condition, behavior change is demonstrated. Three clear and consistent demonstrations of behavior change (i.e., basic effect) at three different points in time are required for evidence of functional relations and to establish experimental control (e.g., Horner et al., 2005; Kratochwill et al., 2013). Visual analysis of SCR data is the primary evaluative method to identify functional relations between the independent variable (i.e., the intervention) and the dependent variable (i.e., the outcomes) and identify functional relations (Horner et al., 2005; Kazdin,

[*] Corresponding author.
E-mail address: erin.e.barton@vanderbilt.edu (E.E. Barton).

https://doi.org/10.1016/j.ridd.2018.11.001
Received 8 August 2017; Received in revised form 2 October 2018; Accepted 1 November 2018
Available online 16 November 2018

2011).

Although there are no ubiquitous protocols for conducting visual analyses, SCR textbooks (Gast & Ledford, 2014; Kazdin, 2011), design standards (Kratochwill et al., 2013), and quality indicators (Horner et al., 2005) include general guidelines for visual analyses. These guidelines describe four characteristics of data patterns to examine within and across adjacent conditions: level, trend, stability, and variability. Further, three additional characteristics are examined across adjacent conditions: immediacy of the behavior change, overlap in data points, and consistency of the magnitude of behavior change or data patterns. What Works Clearinghouse (WWC; Institute for Education Sciences, 2017) provided evaluative criteria that can be used to conduct visual analyses. These are: (a) consistency of changes in level, trend, and variability within and across each condition and tier; (b) consistency of data and presence of overlap across conditions and tiers (i.e., vertical analysis); (c) projected patterns of the data; and (d) anomalies within the data (Kratochwill et al., 2013). Lane and Gast (2014) provided similar guidelines and advocated for assessing level, variability, trend, overlap, intercept gap, and consistency of data across conditions when conducting visual analysis in single case design studies. The dynamic, formative nature of visual analysis facilitates an iterative process to identify effective interventions related to meaningful outcomes. The researcher's close contact with the data through visual analysis allows him/her to make informed changes to the intervention or study design to ensure each participant benefits from the intervention (Barton et al., 2016; Gast & Spriggs, 2014; Wolery, 2013).

Visual analysis continues to be the primary method for evaluating SCR and identifying functional relations; however, there has been increased criticism of this evaluation method (Ninci, Vannest, Willson, & Zhang, 2015). Although much effort has been placed in creating trainings and protocols to enhance the reliability of visual analysis (e.g., Swoboda, Kratochwill, Horner, Levin, & Albin, 2012; Wolfe & Slocum, 2015), concerns remain that visual analysis procedures are not standardized and subjective judgments about behavior change and magnitude of effects are customary. This has resulted in higher than expected rates of disagreement (i.e., low reliability) when interpreting SCR data (Kazdin, 2011; Lieberman, Yoder, Reichow, & Wolery, 2010; Wolfe, Seaman, & Drasgow, 2016) even when accounting for previous experience or levels of expertise (Ottenbacher, 1993). Inconsistent visual analysis procedures also might limit the ability of researchers to quantify and aggregate findings across studies to evaluate evidence-based practices.

Computational methods (e.g., overlap indices, statistical analyses) have been developed and are increasingly being applied to SCR to provide a quantitative summary of the effects, which can supplement visual analysis (Brossart, Vannest, Davis, & Patience, 2014; Parker, Vannest, & Davis, 2011). They also can be used to aggregate findings across studies, similar to meta-analytic techniques used in group design research (Barton, Pustejovsky, Maggin, & Reichow, 2017). Although their use is promising and recommended (Shadish, Hedges, Horner, & Odom, 2015), additional research is needed regarding their ubiquitous application (Odom, Barton, Reichow, Swaminathan, & Pustejovsky, 2018; Wolery, 2013) given none meet all requirements for application to SCR (Shadish, 2014). For example, Harrington and Velicer (2015) compared interrupted time series analysis to visual analysis for all single case studies published in the *Journal of Applied Behavior Analysis* in 2010. They found interrupted time series analysis and visual analysis had low level of agreement and recommended the two methods should be applied concurrently rather than exclusively. There also are a number of non-parametric effect sizes that do not rely on means, medians, or modes (which are detrimentally influenced by outliers), but rather consider the individual values of all data points in pairwise comparisons across phases (e.g., Nonoverlap of All Pairs [NAP], Tau-U). The effect sizes are relatively easy to calculate and can be used to describe graphed data *after* visual analysis is used identified functional relations (Parker, Vannest & Davis, 2011; Parker, Vannest, Davis, & Sauber, 2011 also see http://www.singlecaseresearch.org).

There also are different types of non-overlap indices that have been applied to SCR and used as within-case effect sizes. The non-overlap indices have various advantages and disadvantages (cf. Parker, Vannest & Davis, 2011; Parker, Vannest, Davis & Sauber, 2011). Two frequently used non-overlap indices are the Tau-U (Parker, Vannest, Davis & Sauber, 2011) and NAP (Parker & Vannest, 2009). Tau-U is a distribution free nonparametric method for measuring non-overlap data between adjacent phases. NAP is also a nonparametric technique to measure non-overlap of two adjacent phases; however, NAP does not evaluate data trend, which limits its utility. These two methods offer benefits over other overlap methods, because they use all data points in their calculations. Criticisms of these methods point to their inability to account for replication or magnitude, likely disagreement with visual analysis, and failure to correct or account for typical data patterns (e.g., trend) or serial dependency (Qi, Barton, Collier, Lin, & Montoya, 2015; Rakap, Snyder, & Pasia, 2014; Wolery, Busick, Reichow, & Barton, 2010). Further, non-overlap indices are strongly influenced by the direct observation procedural characteristics such as the number of observations in the baseline, the type of observation system, or intervention conditions and the length of observation sessions (Pustejovsky, 2016c, b) and the study designed used (Chen, Hyppa-Martin, Reichle, & Symons, 2016). These limitations make comparisons across cases or studies with different data patterns or procedural characteristics difficult, and increase the likelihood of adversely influencing judgments about data patterns.

In 2015, the Institute of Education Sciences (IES) commissioned a paper describing the benefits of between-case SCR analyses (Shadish et al., 2015). The authors indicated that the three approaches outlined in their paper (i.e., between-case *d*-statistics, multi-level modeling R-package, Bayesian multi-level approach) could be used to generate effect sizes and synthesize SCR. The authors argued these methods were better than non-parametric within case methods (i.e., overlap methods) and are increasingly used as effect sizes given they are in a metric that is familiar to educational researchers who work primarily with between-subjects designs. Further, they are on the same scale as standardized mean differences from between-groups designs and can be compared to the results of group designs studies (Pustejovsky, Hedges, & Shadish, 2014; Shadish, Hedges, & Pustejovsky, 2014). For example, the between-case standardized mean difference (BC-SMD) characterizes average treatment effects at the level of the study, rather than the individual case (Hedges, Pustejovsky, & Shadish, 2012, 2013). It is defined based on a hierarchical model for the data from a SCR which captures both within-case and between-case variation in the dependent variable. The BC-SMD has multiple advantages over non-overlap measures because it provides an average treatment effect at the level of the study, rather than for an individual case within a study and shares a common scale with the standardized mean difference parameter estimated by a between-groups experimental design (Pustejovsky et al., 2014; Shadish et al., 2015).

These effect sizes should only be applied after a functional relation is confirmed using visual analysis; however, they *have been used* to confirm or support functional relation across published SCR (cf. Bowman-Perrott et al., 2013; Dart, Collins, Klingbeil, & McKinley, 2014; Qi et al., 2015; Rakap, 2015; Vannest & Ninci, 2015; Whalon, Conroy, Martinez, & Werch, 2015). This is problematic given using visual analysis and effect sizes to confirm functional relations inaccurately assumes visual analysis and effect sizes serve the same purpose. Effect sizes likely have critical roles in SCR, which are (a) descriptive verification of magnitude of effect and (b) quantification of effect across a common scale to facilitate syntheses; effect sizes are not useful for establishing the presence of a functional relation (Parker & Vannest, 2012; Pustejovsky, 2018; Wolery et al., 2010; Zimmerman et al., 2018).

Given the field's continued emphasis on synthesizing research to identify effective practices, an in-depth examination of visual analysis procedures, non-overlap indices, and statistical analyses is warranted. We (Fettig & Barton, 2014) conducted a systematic review of parent implemented functional assessment (FA)-based interventions to examine the implementation features and methodological rigor of published studies. Given all 13 studies identified for the review used SCR studies, this literature seemed particularly well suited to an examination of visual analysis features and non-overlap indices. FA-based interventions focus on understanding that challenging behaviors serve a specific purpose or function for the individual child (e.g., gaining attention, escaping a demand, getting access to a tangible item or activity). Within the FA-based interventions, assessments are conducted to identify the function, and interventions are developed and implemented to teach the child more appropriate and effective skills that serve the same function and could replace the challenging behaviors. Research shows that FA-based interventions have distinct benefits when compared to non-function based interventions (Ingram, Lewis-Palmer, & Sugai, 2005), and they are effective in reducing challenging behaviors and increasing pro-social behaviors in toddlers and preschoolers (Conroy, Dunlap, Clarke, & Alter, 2005; Duda, Clarke, Fox, & Dunlap, 2008; Dunlap & Fox, 2011; Fettig & Ostrosky, 2011). Research also suggests that parents can be supported to implement FA-based interventions (Fettig, Schultz, & Sreckovic, 2015; Kurtz, Boelter, Jarmolowicz, Chin, & Hagopian, 2011). We used this literature to examine (a) visual analysis features and procedures described in FA-based interventions implemented by parents, (b) compare the independent visual analyses to what was reported by authors, and (c) compare independent visual analyses to two non-overlap methods and a method of statistical analysis. The research questions were: (a) Do authors describe their single-case data in terms of trend, level, variability, average/mean, overlap, immediacy of change, and consistency? (b) Do independent visual analyses of published data agree with author(s) reported results? and (c) How does independent visual analysis compare to two non-overlap methods (i.e., Tau-U & Nonoverlap of All Pairs [NAP]) and BC-SMD calculations?

2. Method

The search terms, search procedures, and article inclusion criteria were used in the exact manner as in the original review (Fettig and Barton, 2014, n = 13). FA-based parent intervention studies were identified using several combinations of the following two categories of terms: *functional assessment*, *functional communication training* AND *challenging behaviors*, *parent training*, *parent intervention*. Searches were conducted in PsycINFO and ERIC databases with limits set for peer-reviewed journal articles, those written in English, and those reporting studies involving human participants. Ancestral searches were conducted using the identified studies. Studies were included if they: (a) had a functional assessment (FA) component, (b) focused on FA-based parent interventions to reduce challenging behaviors, (c) included and reported direct observations of child behaviors as part of the dependent variable, (d) had parents as the primary participants, (e) had children ages 8 years or younger with or without disabilities and with challenging behaviors, and (f) employed an experimental SCR design. An updated search was conducted with limits set for the years to identify only studies published after 2012 (the upper year limit of the previous review; Fettig and Barton, 2014). A hand search of the reference lists of identified studies was also conducted. Two additional studies were identified (Fettig et al., 2015; Robertson, Wehby, & King, 2013); a total of 15 articles, published between 1997 and 2015, were identified for the current review.

2.1. Coding visual analysis procedures and terms

Four visual analysis coding procedures were utilized. First, we (the three authors) collectively created a descriptive coding system for extracting information reported in each study regarding study methods and visual analysis terms and procedures. We practiced using the coding system with non-study articles and refined the system. The first author coded information regarding study design and measurement system using the WWC Pilot Single Case Design Standards as interpreted from the WWC Procedures and Standards Handbook Version 4.0. (IES, 2017; Kratochwill et al., 2013). We then applied the single case design standards exactly as specified in the handbook.

We independently used the system to identify child and parent graphed variables. A total of 45 child and 7 parent variables were measured and graphed across the 15 studies. We then extracted information from all 52 variables related to the visual analysis terms (i.e., mean, level, trend/slope, variability, stability, immediacy of change, overlap/non-overlap, behavior change, similar pattern, and consistency) and procedures (i.e., include data pattern in research questions, evaluate data within and between conditions, evaluate data across tiers, provide a reference) reported by authors to describe the results of the study (i.e., in the results or discussion sections). Descriptions of all terms are included in Table 1. Also, iterations of the terms were counted when they were used to describe data patterns (e.g., immediate change was accepted for immediacy of change; consistent for consistency, variable for variability, same for similar).

Table 1
Visual Analysis Terms and Descriptions.

Term	Description
Level	For one datum point, it represents the value of the data on the dependent measure. The relative magnitude of the data within a condition can be identified using the y-axis for scale. It is often reported as high, medium, or low and assessed within and compared across conditions.
Mean	The calculated average of the data values within a condition.
Trend/slope	The overall direction and slope of the data within a condition. It is often reported as accelerating (increasing) or decelerating (decreasing) and assessed within and compared across conditions.
Variability	The fluctuation from one datum point to the next and the range of values of data points with a condition. It is the opposite of stability or the lack of predictability in adjacent data points. It is often reported as high, medium, or low and assessed within and compared across conditions.
Stability	The predictability and consistency of data values within a condition. It is the opposite of variability or the lack of fluctuations in adjacent data points. It can be influenced by the scale of the y-axis given the range of the scale will impact judgments.
Immediacy of change	The magnitude of change (level, trend, or variability) between the final 3-5 data points in one condition and the first 3-5 data points in the subsequent condition.
Overlap	The values of the data in one condition are in the same range of values of the data in the subsequent condition. It is often reported as the proportion of data from one condition that coincides with an adjacent condition.
Non-overlap	The values of the data in one condition are different than the range of values of the data in the subsequent adjacent condition.
Behavior change	The data patterns (level, trend, or variability) in one condition are different from the data patterns in the subsequent, adjacent condition for the same variable.
Same/Similar Pattern	The data patterns in one condition are compared to adjacent conditions and similar conditions. Patterns of data are characterized by level, trend, and variability.
Consistency	The data patterns in one condition (level, trend, or variability) are similar to the data patterns in adjacent conditions and similar conditions.

Second, we independently reviewed the 52 different graphed dependent variables and evaluated the presence of a functional relation. We used three criteria to identify functional relations: (a) the use of a methodologically sound single case research design, (b) clear demonstration that the independent variable produced consistent changes in the dependent variable changed (i.e., three behavior changes at three temporally related points in time), and (c) identification and minimization of threats to internal validity. The four-step visual analysis process outlined by Kratochwill et al. (2013) was used: (a) documentation of a stable baseline, (b) examining within condition patterns, (c) comparing adjacent conditions to assess behavioral change, and (d) analyzing data across conditions and tiers to document at least three demonstrations and three different points in time to determine if a functional relation exists. We then documented study authors' conclusions of functional relations between the independent variable and each target behavior as reported in the results or discussions sections of each study. We also judged the social significance of the functional relation (when present) and coded if the decision regarding functional relations matched results reported by the study author(s).

Lastly, we also independently judged the magnitude of behavior change for the graphs with a functional relation. Although there are no known guidelines for visually evaluating magnitude, we used the following scale: *large effect, medium effect, small effect,* and *no effect*. We discussed all disagreements until consensus was achieved for all terms and definitions; however, overall agreement prior to discussions was 94.5% for all variables except magnitude ratings. We separated the magnitude ratings out given it was the only variable utilizing a rating scale. Across the 3 raters, we had 75% agreement across magnitude ratings and all disagreements were within one rating; we were able to come to consensus with disagreements.

2.2. Calculation of non-overlap indices

2.2.1. Data extraction

To obtain numerical values from graphed data, we trained a graduate student to digitize each datum point using the Plot Digitizer™ software program. The 15 single case studies included in this review had a total of 52 different graphed dependent variables (for both child and parent behaviors). All 52 data series were digitized using the following steps. First, a portable document format (PDF) file for each article was opened in the Plot Digitizer™ program. Second, the axes for each graph were identified and scaled by clicking-on the origin and specifying their maximum values. Third, each datum point in each relevant data series was selected (i.e., clicked on) and exact coordinates were recorded. Fourth, the values were exported into a Microsoft Excel spreadsheet. A second trained graduate student independently extracted and digitized data from 25 (48%) of the 52 variables. An agreement was noted when data were within 2% of each other. The total percent agreement was 99%.

2.2.2. Non-overlap calculations

The Excel files with extracted data were used to calculate Tau-U and NAP. Tau-U and NAP calculations were made for the primary adjacent conditions (i.e., baseline and intervention). Online Tau-U and NAP calculators (i.e., http://www.singlecaseresearch.org/) were used for all calculations. Summary scores for each primary adjacent condition were recorded into a separate Excel spreadsheet. Given decisions about functional relations are made based on comparisons within and across adjacent conditions, we calculated and compared adjacent baseline and intervention conditions. We used the criterion suggested by Parker and Vannest (2009) and averaged across tiers or adjacent conditions for each study design as per their recommendations. Specifically, Parker and Vannest (2009) characterized values between 0 and +/−0.65 as "weak," values between +/−.66 and +/−.92 as "medium," and values between

+/−.93 and +/−.1.00 as “large” (p. 364).

2.2.3. Single-case effect sizes (BC-SMD)

The BC-SMD effect size is premised on a statistical model with the following assumptions: (a) the baseline has no trend and is stable; (b) the intervention produces an immediate change in the dependent variable; (c) the intervention effect is constant across cases; (d) the outcome is normally distributed about case- and phase-specific mean levels; and (e) deviations from mean levels follow a first-order autoregressive process. The BC-SMD effect size estimate allows serial dependence—of a certain type—rather than assuming independence of the dependent variable(s). However, a primary limitation the BC-SMD is that it can only be calculated for studies that use an across participant multiple baseline or multiple probe or withdrawal (i.e., A-B-A-B) design replicated across at least three participants. In our study, analyses of the BC-SMD were therefore limited to five studies that used an acceptable design that fit the model. All five studies used a multiple baseline across (at least three) participants SCR design. The app—scdhlm—was used for all BC-SMD difference calculations (i.e., downloaded from: https://jepusto.shinyapps.io/scdhlm/; Pustejovsky, 2016b).

2.2.4. Reliability calculations

We calculated inter-rater agreement, kappa, and the Pearson correlation coefficient for all visual analysis, overlap, and effect size ratings using Microsoft Excel.

3. Results

3.1. Descriptive summary

Studies included in this review were most likely to use a time-lagged SCR designs; 13 of the 15 studies used a multiple baseline design and 2 used a withdrawal design (see Table 2). Twelve of the 15 studies used a partial interval recording system, four used event recording, and one used a latency measure to measure child behaviors; one study used two types of measurement systems (Dunlap & Fox, 1999). Parent behaviors were measured using accuracy systems (percentage correct) or partial interval recording. One study met WWC single case design standards, five met with reservations, and the remaining nine studies did not meet standards. Seven of the nine studies that did not meet design standards, did not meet the design standard related to data points per condition, which requires 3 or more data points per condition. Two studies did not meet design standards related to interobserver agreement calculations.

3.2. Visual analysis

3.2.1. Visual analysis terms

Our first research question focused on the visual analysis characteristics examined and procedures used. Across the 15 studies, 7 parent variables and 45 child dependent variables were examined. As shown in Table 3, study authors were most likely to discuss

Table 2
Measurement and Design Information Across Studies.

Reference	No. Children (Parents)	SCR Design	Measurement for Child Behavior	Measurement for Parent Behavior	WWC Evaluation (DS[a])
Duda et al. (2008)	3 (1)	MBxR	PIR	*	MWR
Dunlap and Fox (1999)	6	MBxS, P**	ER, PIR	*	DNM (DPC)
Dunlap et al. (2006)	2 (2)	MBxR	PIR	PIR	MWR
Fettig and Ostrosky (2011)	2 (2)	MBxP	PIR	% Correct	DNM (DPC)
Fettig et al. (2015)	3 (3)	MBxP**	PIR	% Correct	MWR
Frea and Hepburn (1999)	2	MBxP	ER	*	DNM (DPC)
Galensky et al. (2001)	3	MBxP	PIR	*	DNM (DPC)
Harding et al. (2009)	1	Withdrawal	PIR	*	MWR
Koegel et al. (1998)	3	MBxP**	PIR	*	DNM (IOA, DPC)
Lucyshyn et al. (2007)	1	MBxS	ER, Lat	*	M
Marcus et al. (2001)	3 (3)	MBxP**	PIR	% Correct	DNM (IOA)
Moes and Frea (2002)	3	MBxP**	PIR	*	DNM (DPC)
Robertson et al. (2013)	2	Withdrawal	ER	*	DNM (DPC)
Vaughn et al. (1997)	1	MBxS	PIR	*	DNM (DPC)
Vaughn et al. (2002)	1 (1)	MBxR	PIR	PIR	MWR

Note. MBxR: multiple baseline across routines; MBxS: multiple baseline across settings; MBxP: multiple baseline across participants; PIR: partial interval recording; ER: Event Recording; Lat: Latency; DS: design standard; M: meets standards; MWR: meets with reservations; DNM: does not meet.

[a] Results based on the Works Clearinghouse (WWC) Pilot Single Case Design Standards (Appendix E) as interpreted from the WWC Procedures and Standards Handbook Version 3.0. The () indicates the specific standard(s) that was not met. DPC = data points per condition; IOA = Interobserver Agreement.

* Study authors did not report or measure parent behaviors.

** Study authors employed a multiple baseline across participants design with at least 3 participants; thus, was included in the BC-SMD analyses.

Table 3
Analysis Terms and Procedures Used Across Figures.

Term	# Figures (%) (n = 52)
Total Use of Analysis Terms Across Figures	
Level	24 (46%)
Mean/Median/Average	32 (62%)
Trend/Slope	10 (19%)
Stability	12 (23%)
Variability	4 (8%)
Immediacy of change	8 (15%)
Overlap/non-overlap	0
Behavior change	3 (6%)
Same pattern	2 (4%)
Consistency	9 (17%)
Visual Analysis Methodology Across Figures	
Agreement with authors	20 (39%)
Include expected data patterns in hypotheses or research questions	0
Evaluate data patterns within conditions	47 (90%)
Evaluate data patterns between conditions	38 (73%)
Conduct vertical analysis (n = 44)*	16 (36%)
Provide a reference for visual analysis procedures	0

Note. *Forty-four of the 52 figures used tiered experimental designs that required vertical analysis.

results using mean/median/average. Level, trend (or slope), and stability were the most frequently used visual analysis terms. Given level, trend, slope, variability, and stability might be differentially used depending on data patterns in an individual study, we also calculated the numbers of studies using at least one of these terms to evaluate dependent variables; at least one of these five terms was used to describe 31 variables (60%). When authors used just one visual analysis term to describe results, it was level in 12 of 20 instances. None of the study authors used the terms overlap or non-overlap to describe data patterns across conditions. The terms consistency, immediacy of change, behavior change, and same/similar pattern were used to describe data patterns for fewer than 10 variables.

3.2.2. Visual analysis procedures

Authors reported evaluating data within conditions for 90% of the variables and between conditions for 73% of the variables (see Table 3). Forty-four of the variables were conducted as part of a tiered design; yet authors described conducting vertical analyses for only 16 (36%) of the 44 variables. None of the authors included the expected, directional data patterns in their research questions or hypotheses. Further, none of the authors included a reference with their visual analysis procedures.

3.3. Visual analysis agreement

Our second research question examined whether independent visual analyses of published data agreed with reported results. Of the 52 variables coded across the 15 studies, original authors reported that there was a functional relation for 44 variables. Our independent visual analyses yielded agreement for 20 variables (39%; 12 demonstrated functional relations). Interestingly, all of the disagreements (n = 32) were when the study author(s) indicated there was a functional relation, whereas independent visual analysts concluded there was not a functional relation. Finally, we also evaluated the magnitude of behavior change for all functional relations. Of the 12 variables with functional relations, 4 were identified through consensus coding as having large effects, 3 with medium effects, and 5 with small effects.

3.4. Visual analysis, overlap methods, and Single-case effect sizes (BC-SMD)

The third research question examined agreement between visual analysis and two non-overlap indices (i.e., Tau-U &NAP). NAP and Tau-U were only calculated for the 11 studies that had a sufficient number of opportunities to demonstrate a functional relation (i.e., 3 opportunities), which included 30 variables. As shown in Table 4, NAP scores ranged from 0.01 to 1.00 for the 30 variables, and Tau-U ranged from -0.99 to 1.00. NAP identified medium or large effects for 10 of the 30 variables, Tau-U identified medium or large effects for 26 of the 30 variables, and the visual analysts identified medium or large effects for 7 of the 30 variables. The Pearson correlation coefficient for Tau-U and NAP indicated a moderate positive relation ($r = 0.35$, $n = 30$).

As shown in Table 4, agreement among the three measures (i.e., visual analysis, Tau-U, and NAP) for the 30 variables which had sufficient number of opportunities to demonstrate a functional relation was low ($n = 4$, 13%). We then compared across studies with documented functional relations based on visual analysis ($n = 12$). Visual analysis, Tau-U, and NAP agreed for 4 variables (33%) across 3 studies. Visual analysis and Tau-U agreed for 5 (42%) of the 12 variables with a functional relation ($r = 0.30$, $n = 12$). Visual analysis and NAP agreed for 7 (58%) of the 12 variables with a functional relation ($r = 50$, $n = 12$). Tau-U and NAP agreed for 5 (42%) of the variables with a functional relation ($r = 0.52$, $n = 12$). The agreements were low with two exceptions: when the visual

Table 4
Visual Analysis, Tau-U, NAP, BC-SMD Designs With Sufficient Opportunities to Demonstrate a Functional Relation.

Reference	Figure, DV	Visual Analysis		Tau-U (average)	NAP (average)	BC-SMD	
		FR	Magnitude[*]			Est. (SE)	95% CI
Koegel, Stiebel, and Koegel (1998)[a]	Figure 1, CB	NO		Medium (-0.88)	Weak (0.06)	−0.6 (0.33)	[−1.27, −0.08]
Dunlap and Fox (1999)[a]	Figure 2, CB	NO		Large (-0.93)	Weak (0.03)	−2.2 (0.44)	[−3.10, −1.37]
Marcus, Swanson, and Vollmer (2001)[a]	Figure 2, PCI	NO		Medium (0.86)	Large (0.93)	2.8 (0.58)	[1.77, 3.98]
	Figure 2, PII	NO		Large (0.90)	Large (0.95)		
	Figure 3, CB	NO		Weak (-0.53)	Weak (0.24)		
	Figure 3, AB	YES	Medium	Large (0.97)	Large (0.99)		
Moes and Frea (2002)[a]	Figure 2, CB	NO		Large (-0.94)	Weak (0.03)	−2.5 (0.51)	[−3.48, −1.50]
	Figure 2, AB	NO		Large (0.96)	Large (0.98)	2.2 (0.55)	[1.18, 3.29]
Vaughn, Wilson, and Dunlap (2002)[a]	Figure 1, AB	NO		Medium (0.89)	Large (0.94)		
	Figure 1, CB	NO		Medium (-0.89)	Weak (0.06)		
	Figure 2, PCI	NO		Medium (0.89)	Large (0.94)		
Dunlap, Ester, Langhans, and Fox (2006)[b]	Figure 1, CB	YES	Medium	Medium (-0.81)	Weak (0.09)		
	Figure 2, CB	YES	Large	Medium (-0.90)	Weak (0.05)		
Lucyshyn et al. (2007)[c]	Figure 2, CB	YES	Large	Medium (-0.89)	Weak (0.06)		
	Figure 3, CB[**]	YES	Large	Large (0.96)	Large (0.98)		
Duda et al. (2008)[b]	Figure 1, CB	YES	Medium	Large (-0.99)	Weak (0.01)		
	Figure 2, CB	NO		Medium (-0.72)	Weak (0.14)		
	Figure 2, AB	NO		Medium (0.74)	Medium (0.87)		
	Figure 3, CB	NO		Medium (-0.78)	Weak (0.11)		
	Figure 3, AB	NO		Medium (0.86)	Large (0.93)		
Harding, Wacker, Berg, Lee, and Dolezal (2009)[b]	Figure 3, T1, CB	YES	Small	Weak (0.65)[***]	Weak (0.42)		
	Figure 3, T2, AB	YES	Small	Large (0.97)[***]	Weak (0.57)		
	Figure 3, T2, AB2	NO		Weak (0.27)[***]	Weak (0.55)		
	Figure 3, T3, AB3	YES	Small	Weak (0.56)[***]	Weak (0.59)		
Robertson et al. (2013)[b]	Figure 3, AB	NO		Medium (0.85)[***]	Weak (0.53)		
	Figure 3, CB	NO		Medium (0.78)[***]	Weak (0.48)		
	Figure 4, AB	NO		Medium (0.66)[***]	Weak (0.35)		
	Figure 4, CB	YES	Small	Large (0.93)[***]	Weak (0.49)		
Fettig et al. (2015)[b]	Figure 1, CB	YES	Small	Medium (-0.87)	Weak (0.06)	−1.4 (0.48)	[−2.40, −0.54]
	Figure 1, PCI	YES	Large	Large (1.00)	Large (1.00)	3.2 (0.69)	[2.00, 4.63]

FR: Functional relation; CB: child challenging behavior; PCI: parent correct implementation; PII: parent incorrect implementation; AB: child appropriate behavior; BC: behavior change.

[*] Magnitude was only evaluated via visual analysis for functional relations.

[**] Lucyshyn et al. (2007) measured latency to CB in Figure 3.

[***] Averaged the absolute value to aggregate Tau-U per demonstration.

[a] ***Did not meet*** Works Clearinghouse (WWC) Pilot Single Case Design Standards (Appendix E) as interpreted from the WWC Procedures and Standards Handbook Version 3.0.

[b] ***Met*** Works Clearinghouse (WWC) Pilot Single Case Design Standards (Appendix E) ***with reservations*** as interpreted from the WWC Procedures and Standards Handbook Version 3.0.

[c] ***Met*** Works Clearinghouse (WWC) Pilot Single Case Design Standards (Appendix E) ***without reservations*** as interpreted from the WWC Procedures and Standards Handbook Version 3.0.

analysis rating for magnitude was medium or large, the TAU value was large and when the visual analysis rating for magnitude was small, the NAP value was weak. There were no other clear patterns observed. When visual analysts concluded there was not a functional relation ($n = 18$), Tau-U indicated a large effect for 4 variables (22%), a medium effect for 12 variables (67%), and weak effects for 2 variables (17%). Conversely, NAP indicated a large effect for 6 of these variables (33%), a medium effect for 1 variable (6%), and weak effects for 11 variables (61%). Table 5 lists complete reliability calculations across original study authors, independent visual analyses, Tau-U, NAP, and BC-SMD.

The BC-SMD were calculated for seven outcomes across five studies that fit the designs for which BC-SMD methodology currently

Table 5
Reliability Calculations.

	Percentage Agreement	Weighted Percentage Agreement	Kappa	Weighted Kappa
Independent Visual Analysis Across Three Raters	75% (9/12)	100% (12/12)	–	–
Study Author versus Independent Visual Analysis	39% (20/52)	–	0.10	–
Tau-U versus Independent Visual Analysis	42% (5/12)	83% (10/12)	0.14	0.21
NAP versus Independent Visual Analysis	58% (7/12)	83% (10/12)	0.32	0.40
Tau-U versus NAP	42% (5/12)	75% (9/12)	0.11	0.29
BC-SMD versus Independent Visual Analysis	29% (2/7)	–	–	–

Note. Agreements/Disagreements are provided in parentheses for percentage agreement calculations.

exists (i.e., A-B-A-B or multiple baseline with three cases). In this case, all five studies were used multiple baseline across participants designs. Outcomes across the studies varied from child challenging behaviors (4), child appropriate behaviors (1), and parent implementation (2). For the BC-SMD calculations, we used the REML estimations, which is recommended for multiple baseline designs. After inspecting our data across all five studies we assumed no trends. At the case level, as shown in Table 4, the confidence intervals for all BC-SMD were significantly different from zero indicating treatment effects. However, given visual analyses concluded functional relations did not exist for the majority of these variables (i.e., 5 of 7 variables), additional analyses including average effect size estimates were not conducted. Further, none of the eligible studies met WWC single case design standards and only one met with reservations.

4. Discussion

The purposes of the current study were (a) to examine visual analysis features and procedures described in the FA-based intervention research literature, (b) to conduct independent visual analyses, and (c) to compare the independent visual analyses to: what was reported by study author(s), two non-overlap methods, and BC-SMD calculations.

4.1. Visual analysis summary

Overall, authors of the identified studies frequently used mean/median/average to describe their results; level, trend (or slope), and stability were the most frequently used visual analysis terms. The terms consistency, immediacy of change, behavior change, and same/similar pattern were infrequently used, and none of the authors of the identified studies used the terms overlap or non-overlap to describe data patterns across conditions. The results support and extend previous research. For example, previous research indicated authors in SCR rarely report the analytic methods or procedures used to conduct visual analysis (Shadish & Sullivan, 2011; Smith, 2012). However, our study is the first to examine the terms used to describe SCR results, and found authors are not consistent using visual analysis terms. Further, our independent visual analyses of published data only yielded 38% of agreement with the results reported by the authors of the identified studies, which might have been impacted by the lack of transparent visual analysis procedures or consistent terminology. All of the disagreements were when the study author(s) indicated there was a functional relation and our independent analyses concluded there was not a functional relation.

4.2. Effect size measures and visual analysis comparison

Overall, for functional relations and magnitude ratings the observed reliability between all indicators (i.e., original study authors, independent visual analyses, NAP, Tau-U, and BC-SMD) ranged from 29% to 75% using the percentage agreement method, from 0.10 to 0.32 for kappa, and from 0.21 to 0.40 for weighted kappa (see Table 5 for all calculations). NAP, Tau-U, and independent visual analyses had high rate of disagreement (*range* = 42%–58% for percentage agreement; 0.10–0.32 for kappa; 0.21–0.40 for weighted kappa). There was inconsistency and low reliability across visual analysts' judgements and the non-overlap and effect size metrics. This is important and suggests additional research is needed to endorse the continued use of non-overlap measures as descriptive verification to complement visual analysis.

The magnitude ratings did not correlate with TAU or NAP, which is not surprising given the absolute value of TAU and NAP are impacted by the number of data points that overlap across conditions, not the magnitude of the overlap across conditions. Inconsistencies in magnitude agreements also could be due to the fairly subjective nature of visual analyses. However, the NAP and Tau-U pair-wise comparisons result in decisions of yes, tied, or no (overlapping or not) rather than the magnitude of differences between data points, which results in a magnitude of overlap rather than magnitude of behavior change across conditions. Interestingly, studies with large effects and studies that met WWC standards with or without reservations had a high likelihood of visual analysis (i.e., functional relation) and Tau-U agreement. This suggests the lack of overlap—which result in large Tau-U scores—is a strong predictor of large effect ratings by visual analysts, which should be examined in future research. For example, researchers should continue to examine and compare visual analysis judgements to non-overlap and effect size calculations using a larger and more variable set of research. Interestingly, correlation coefficients indicated moderate to strong positive relationships; however, given the limited sample size these calculations should be interpreted with caution.

Our findings also suggest that SCR studies with high levels of rigor based on contemporary SCR design standards might be more likely to arrive to comparable conclusions between visual analysis and Tau-U. However, additional analyses are needed given so few studies in this review met SCR design standards (n =6 studies). Also, designs with functional relations (n = 12) were in the studies that met design standards with or without reservations.

4.3. Future research

Results of this study highlighted the need for transparent and consistent visual analysis standards and procedures. Standardized or transparent processes for visual analyses might minimize the disagreements regarding functional relations. Single case researchers expect data to be presented in a minimally transformed manner to allow for independent judgments about data patterns and the presence of functional relations. As substantiated by the current evaluation, visual analysis continues to be the gold standard for the evaluation of single case data (Johnston & Pennypacker, 2009). Despite the importance of visual analysis, multiple concerns about the interrater reliability of visual analysis have existed for decades (e.g., Brossart, Parker, Olson, & Mahadevan, 2006; DeProspero &

Cohen, 1979; Lieberman et al., 2010). Inconsistent results across different visual analysts can impact both the credibility of individual studies and the associated research methodology (Kratochwill & Levin, 2014). The validity of a study's results and the broad acceptability of single case research as a valid and rigorous experimental methodology are dependent on the reliability and soundness of the analysis procedures. Poor reliability in visual analysis has several negative implications for SCR and has precluded the inclusion of SCR in meta-analyses and systematic reviews of evidence-based practice.

Reliable and valid procedures for visually determining and quantifying the magnitude of effect might be particularly useful for SCR. Overlap measures (i.e., Tau-u and NAP) fail to accurately quantify magnitude of change (they measure the amount of overlap across adjacent conditions), which limits their utility as an effect size. Visual analysts' magnitude evaluation generally involve ratings—as was the case in the current study—which can be difficult to quantify or synthesize given the range of metrics and designs used in SCR. Effect sizes such as the BC-SMD allow researchers to quantify the magnitude of change and aggregate across studies to make judgements about a body of research and inform evidence-based practices. The results of the current study, although limited by the quality/rigor and types of designs used, point to the need for additional meta-analyses of SCR.

Consistent procedures and terminology might advance the science of SCR or at a minimum increase the transparency and reliability of visual analyses. In fact, researchers have argued for the creation and use of formal guidelines to operationalize the visual analysis process (Furlong & Wampold, 1982; Kazdin, 1982). For example, Maggin, Briesch, and Chafouleas (2013) developed a protocol for visual analysis based on the WWC SCR evidence criteria (Kratochwill et al., 2013). The protocol guides visual analysts through a series of questions related to data patterns (Maggin et al., 2013). Although not validated, their protocol has been reliably used in systematic reviews of SCR (Fettig and Barton, 2014; Qi et al., 2015); however, their protocol does not assist the researcher in *synthesizing* data across conditions to make decisions about functional relations. Wolfe et al. (2016) created and evaluated visual analysis protocols that emphasize the data characteristics outlined in the WWC SCR design standards and guides researchers in systematically making decisions about data patterns within and across conditions and tiers to make judgments about functional relations. However, their protocols only address SCR studies using A-B-A-B or multiple baseline designs, which limits broad application.

Our findings indicate NAP or Tau-U and visual analyses do not consistently match, although these interpretations should be taken with caution given study author(s) and independent visual analyses also disagreed. The low reliability between visual analysis and non-overlap methods is problematic, and continued work in this area is critical. In the paper commissioned by IES, Shadish et al. (2015) recommended use of between case effect sizes for synthesizing SCR. They present three different between case effect sizes including the BC-SMD used in the current analyses (i.e., between-case *d*-statistics, Hedges et al., 2012; multi-level modeling R-package, Pustejovsky et al., 2014; Bayesian multi-level approach, Swaminathan, Rogers, & Horner, 2014). They also call for continued development of methods for calculating between case effect sizes, synthesizing effect sizes, conducting moderator analyses, and analyzing publication bias. Although interpretations of our BC-SMD are limited, we hope our findings highlight that the potential for progress is strong for meta-analyses of SCR.

4.4. Limitations

Our analyses should be viewed in light of some limitations. First, we used a specific set of literature (parent implemented FA-based interventions) to conduct our analyses. A different literature might produce differential outcomes and additional analysis should be conducted to corroborate our findings. Further, most of the studies within this literature did not meet contemporary design standards (IES, 2017; see Tables 2 & 4), which limits their utility in meta-analyses. For example, the WWC only includes studies that meet all design standards in their syntheses (IES, 2017). Second, we identified specific terms recommended in SCR quality indicators (Horner et al., 2005), design standards (IES, 2017; Kratochwill et al., 2013), and texts, which are primarily relevant within special education. This might have inadvertently excluded other terms authors used to describe data patterns when using visual analysis. Third, authors of the reviewed literature might have been limited by manuscript length guidelines, which could limit the full discussion of the analysis processes. Fourth, the authors' independent visual analyses did not account for magnitude or strength of the functional relation. This characterization is a critical consideration in education research, and conducting visual analyses without this characterization might have been inauthentic. Fifth, the NAP calculations did not account for directionality or trend of the data. This limits its interpretations of studies in which decreasing trend is the desired outcome. Lastly, the BC-SMD calculations were limited to five studies, none of which met contemporary design standards, which precluded additional analyses.

4.5. Conclusions

The overarching goal of our analyses was to advance the science of SCR by identifying a need for refining the processes by which visual analyses are conducted and reported. However, visual analysis is currently the only way to identify functional relations, which is required before determining the magnitude of the effect. We recommend that guidelines for visual analyses should be ecologically valid and consider the range of professionals and researchers who conduct SCR. Simultaneously, we need to build consensus regarding the use of effect size measures to increase the likelihood that SCR will be used and accepted by broader audiences across the social sciences. Given the substantial progress that has been made in recent years as evidenced by Shadish et al. (2015) the potential for additional progress is irrefutable.

References[1]

Barton, E. E., Pustejovsky, J. E., Maggin, D. M., & Reichow, B. (2017). Technology-aided instruction and intervention for students with ASD: A meta-analysis using novel methods of estimating effect sizes for single-case research. *Remedial and Special Education, 38*(6), 371–386.

Barton, E. E., Ledford, J. R., Lane, J. D., Decker, J., Germansky, S. E., Hemmeter, M. L., & Kaiser, A. (2016). The iterative use of single case research designs to advance the science of EI/ECSE. *Topics in Early Childhood Special Education, 36*(1), 4–14.

Fettig, A., & Barton, E. E. (2014). Parent implementation of function[HYPHEN]based intervention to reduce children's challenging behavior: A literature review. *Topics in Early Childhood Special Education, 34*(1), 49–61.

Bowman-Perrott, L., Davis, H., Vannest, K., Williams, L., Greenwood, C., & Parker, R. (2013). Academic benefits of peer tutoring: A meta-analytic review of single-case research. *School Psychology Review, 42*, 39.

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

Brossart, D. F., Vannest, K. J., Davis, J. L., & Patience, M. A. (2014). Incorporating nonoverlap indices with visual analysis for quantifying intervention effectiveness in single-case experimental designs. *Neuropsychological Rehabilitation, 24*, 464–491.

Chen, M., Hyppa-Martin, J. K., Reichle, J. E., & Symons, F. J. (2016). Comparing single case design overlap-based effect metrics from studies examining speech generating device interventions. *American Journal on Intellectual and Developmental Disabilities, 121*, 169–193.

Conroy, M. A., Dunlap, G., Clarke, S., & Alter, P. J. (2005). A descriptive analysis of positive behavioral intervention research with young children with challenging behavior. *Topics in Early Childhood Special Education, 25*, 157–166.

Dart, E. H., Collins, T. A., Klingbeil, D. A., & McKinley, L. E. (2014). Peer management interventions: A meta-analytic review of single-case research. *School Psychology Review, 43*, 367–384.

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analysis of intrasubject data. *Journal of Applied Behavior Analysis, 12*, 573–579.

*Duda, M. A., Clarke, S., Fox, L., & Dunlap, G. (2008). Implementation of positive behavior support with a sibling set in a home environment. *Journal of Early Intervention, 30*, 213–236.

*Dunlap, G., & Fox, L. (1999). A demonstration of behavioral support for young children with autism. *Journal of Positive Behavior Interventions, 1*, 77–87.

*Dunlap, G., Ester, T., Langhans, S., & Fox, L. (2006). Functional communication training with toddlers in home environments. *Journal of Early Intervention, 28*, 81–96.

Dunlap, G., & Fox, L. (2011). Function-based interventions for children with challenging behavior. *Journal of Early Intervention, 33*, 333–343.

*Fettig, A., & Ostrosky, M. M. (2011). Collaborating with parents in reducing children's challenging behaviors: Linking functional assessment to intervention. *Child Development Research, 2011*, 1–10.

*Fettig, A., Schultz, T. R., & Sreckovic, M. A. (2015). Effects of coaching on the implementation of functional assessment–based parent intervention in reducing challenging behaviors. *Journal of Positive Behavior Interventions, 17*, 170–180.

*Frea, W. D., & Hepburn, S. L. (1999). Teaching parents of children with autism to perform functional assessments to plan interventions for extremely disruptive behaviors. *Journal of Positive Behavior Interventions, 1*, 112–122.

Furlong, M. J., & Wampold, B. E. (1982). Intervention effects and relative variation as dimensions in experts' use of visual inference. *Journal of Applied Behavior Analysis, 15*, 415–421.

*Galensky, T. L., Miltenberger, R. G., Stricker, J. M., & Garlinghouse, M. A. (2001). Functional assessment and treatment of mealtime behavior problems. *Journal of Positive Behavior Interventions, 3*, 211–224.

Gast, D. L., & Ledford, J. R. (2014). *Single case research methodology: Applications in special education and behavioral sciences.* Baltimore, MD: Paul H. Brookes.

Gast, D. L., & Spriggs, A. (2014). Visual analysis of graphic data. In D. Gast, & J. R. Ledford (Eds.). *Single case research methodology: Applications in special education and behavioral sciences*New York: Routledge pp. 234–375.

*Harding, J. W., Wacker, D. P., Berg, W. K., Lee, J. F., & Dolezal, D. (2009). Conducting functional communication training in home settings: A case study and recommendations for practitioners. *Behavior Analysis in Practice, 2*, 21–33.

Harrington, M., & Velicer, W. F. (2015). Comparing visual and statistical analysis in single-case studies using published studies. *Multivariate Behavioral Research, 50*, 162–183.

Hedges, L. V., Pustejovsky, J., & Shadish, W. R. (2012). A standardized mean difference effect size for single-case designs. *Research Synthesis Methods, 3*, 224–239.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S. L., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practices in special education. *Exceptional Children, 71*, 165–179.

Ingram, K., Lewis-Palmer, T., & Sugai, G. (2005). Function-based intervention planning: Comparing the effectiveness of FBA function-based and non-function-based intervention plans. *Journal of Positive Behavior Interventions, 7*, 224–236.

Institute for Education Sciences (2017). *What Works Clearinghouse^{TM} procedures and standards handbook, Version 4.0.* Retrieved from:Washington, DC: Author. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf.

Johnston, J. M., & Pennypacker, H. S. (2009). *Strategies and tactics of behavioral research* (3rd ed.). Hillsdale, NJ: Erlbaum.

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

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

Kratochwill, T. R., & Levin, J. R. (Eds.). (2014). *Single-case intervention research: Methodological and statistical advances*. Washington, D.C: APA.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., et al. (2013). Single-case intervention research design standards. *Remedial and Special Education, 34*, 26–38.

*Koegel, L. K., Stiebel, D., & Koegel, R. L. (1998). Reducing aggression in children with autism toward infant or toddler siblings. *Journal of the Association for Persons with Severe Handicaps, 23*, 111–118.

Kurtz, P. F., Boelter, E. W., Jarmolowicz, D. P., Chin, M. D., & Hagopian, L. P. (2011). An analysis of functional communication training as an empirically supported treatment for problem behavior displayed by individuals with intellectual disabilities. *Research in Developmental Disabilities, 32*, 2935–2942.

Lane, J. D., & Gast, D. L. (2014). Visual analysis in single case experimental design studies: Brief review and guidelines. *Neuropsychological Rehabilitation, 24*, 445–463.

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

*Lucyshyn, J. M., Albin, R. W., Horner, R. H., Mann, J. C., Mann, J. A., & Wadsworth, G. (2007). Family implementation of positive behavior support for a child with autism: Longitudinal, single-case, experimental, and descriptive replication and extension. *Journal of Positive Behavior Interventions, 9*, 131–150.

Maggin, D. M., Briesch, A. M., & Chafouleas, S. M. (2013). An application of the what works clearinghouse standards for evaluating single subject research: Synthesis of the self-management literature base. *Remedial and Special Education, 34*, 44–58.

*Marcus, B. A., Swanson, V., & Vollmer, T. R. (2001). Effects of parent training on parent and child behavior using procedures based on functional analysis. *Behavioral Interventions, 16*, 87–104.

*Moes, D. R., & Frea, W. D. (2002). Contextualized behavioral support in early intervention for children with autism and their families. *Journal of Autism and Developmental Disorders, 32*, 519–533.

Ninci, J., Vannest, K. J., Willson, V., & Zhang, N. (2015). Interrater agreement between visual analysts of single-case data: A meta-analysis. *Behavior Modification, 39*, 510–541.

Odom, S. L., Barton, E. E., Reichow, B., Swaminathan, H., & Pustejovsky, J. E. (2018). Between-case standardized effect size analysis of single case designs: Examination of the two methods. *Research in Developmental Disabilities, 79*, 88–96.

[1] References marked with an asterisk (*) indicate studies included in this review.

Ottenbacher, K. J. (1993). Interrater agreement of visual analysis in single-subject decisions: Quantitative review and analysis. *American Journal on Mental Retardation, 98*, 135–142.

Parker, R. I., & Vannest, K. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy, 40*, 357–367.

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification, 35*, 303–322.

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy, 42*, 284–299.

Parker, R. I., & Vannest, K. J. (2012). Bottom up analysis of single-case research designs. *Journal of Behavioral Education, 17*, 254–265.

Pustejovsky, J. E. (2018). *Procedural sensitivities of effect sizes for single-case designs with directly observed behavioral outcome measures*.

Pustejovsky, J. E. (2016c). *Procedural sensitivities of effect sizes for single-case designs with behavioral outcome measures.* Retrieved from:http://jepusto.github.io/working_papers/.

Pustejovsky, J. E. (2016a). *Tau-U.* Retrieved fromhttp://jepusto.github.io/Tau-U.

Pustejovsky, J. E. (2016b). *scdhlm: A web-based calculator for between-case standardized mean differences (Version 0.3.1) [Web application].* Retrieved from:https://jepusto.shinyapps.io/scdhlm.

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

Qi, C., Barton, E. E., Collier, M., Lin, Y., & Montoya, C. (2015). A systematic review of the effects of social stories for individuals with autism spectrum disorder. *Focus on Autism and Other Developmental Disabilities, 33*, 25–34. https://doi.org/10.1177/1088357615613516 Advanced online publication.

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

Rakap, S., Snyder, P., & Pasia, C. (2014). Comparison of nonoverlap methods for identifying treatment effect in single-subject experimental research. *Behavioral Disorders, 39*, 128–145.

*Robertson, R. E., Wehby, J. H., & King, S. M. (2013). Increased parent reinforcement of spontaneous requests in children with autism spectrum disorder: Effects on problem behavior. *Research in Developmental Disabilities, 34*, 1069–1082.

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

Shadish, W. R., Hedges, L. V., Horner, R. H., & Odom, S. L. (2015). *The role of between-case effect sizes in conducting, interpreting, and summarizing Single-Case Research (NCER 2015-002).* Retrieved from:Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. http://ies.ed.gov/.

Shadish, W. R., Hedges, L. V., & Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology, 52*, 123–147.

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43*, 971–980.

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

Swaminathan, H., Rogers, H. J., & Horner, R. H. (2014). An effect size measure and Bayesian analysis of single-case designs. *Journal of School Psychology, 52*, 105–122.

Swoboda, C., Kratochwill, T., Horner, R., Levin, J., & Albin, R. (2012). *Visual analysis training protocol: Applications with the alternating treatment, multiple baseline, and ABAB designs.* Retrieved fromhttp://www.singlecase.org/.

Vannest, K. J., & Ninci, J. (2015). Evaluating intervention effects in single-case research designs. *Journal of Counseling & Development, 93*, 403–411.

*Vaughn, B. J., Clarke, S., & Dunlap, G. (1997). Assessment-based intervention for severe behavior problems in a natural family context. *Journal of Applied Behavior Analysis, 30*, 713–716.

*Vaughn, B. J., Wilson, D., & Dunlap, G. (2002). Family-centered intervention to resolve problem behaviors in a fast-food restaurant. *Journal of Positive Behavior Interventions, 4*, 38–45.

Whalon, K. J., Conroy, M. A., Martinez, J. R., & Werch, B. L. (2015). School-based peer-related social competence interventions for children with autism spectrum disorder: A meta-analysis and descriptive review of single case research design studies. *Journal of Autism and Developmental Disorders, 45*, 1513–1531.

Wolfe, K., Seaman, M. A., & Drasgow, E. (2016). Interrater agreement on the visual analysis of individual tiers and functional relations in multiple baseline designs. *Behavior Modification, 40*, 852–873.

Wolfe, K., & Slocum, T. A. (2015). A comparison of two approaches to training visual analysis of AB graphs. *Journal of Applied Behavior Analysis, 48*, 472–477.

Wolery, M. (2013). A commentary: Single-case design technical document of the what works clearinghouse. *Remedial and Special Education, 34*, 39–43.

Wolery, M., Busick, M., Reichow, B. R., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single subject data. *The Journal of Special Education, 44*, 18–28.

Zimmerman, K. N., Pustejovsky, J. E., Ledford, J. R., Barton, E. E., Severini, K. E., & Lloyd, B. P. (2018). Single-case synthesis tools II: Comparing quantitative outcome measures. *Research in Developmental Disabilities, 79*, 65–76.