Recommendations for choosing single-case data analytical techniques

Rumen Manolov[1] and Mariola Moeyaert[2]

[1] Department of Behavioural Sciences Methods, University of Barcelona, Spain

[2] Department of Educational Psychology and Methodology – State University of New York, NY

Running head: SCED analysis guidelines

Contact author

Correspondence concerning this article should be addressed to Rumen Manolov, Departament de Metodologia de les Ciències del Comportament, Facultat de Psicologia, Universitat de Barcelona, Passeig de la Vall d'Hebron, 171, 08035-Barcelona, Spain. Phone number: +34934031137. Fax: +34934021359. E-mail: rrumenov13@ub.edu

Abstract

The current paper responds to the need to provide guidance to applied single-case researchers regarding the possibilities of data analysis. The amount of available single-case data analytical techniques has been growing during recent years and a general overview, comparing the possibilities of these techniques, is missing. Such an overview is provided here referring to techniques that yield results in terms of a raw or standardized difference, procedures related to regression analysis, as well as nonoverlap and percentage change indices. The comparison is provided in terms of the type of quantification provided, the data features taken into account, the conditions in which the techniques are appropriate, the possibilities for meta-analysis, and the evidence available on their performance. Moreover, we provide a set of recommendations for choosing appropriate analysis techniques, pointing at specific situations (aims, types of data, researchers' resources) and the data analytical techniques that are most appropriate in these situations. The recommendations are contextualized using a variety of published single-case datasets in order to illustrate a range of realistic situations that researchers have faced and may face in their investigations.

Keywords: single-case designs, data analysis, recommendations

During the last decade there has been a great proliferation of data analytical techniques for single-case experimental designs (SCEDs) and an intensified discussion on the topic. A bibliographic search performed on September 8, 2015 via the PsycINFO database for years 2005-2014 using "(single-case OR single-subject) AND (analysis)" as keywords to be found in the abstract suggested the following number of papers 3 in 2005 and 2006, 7 in 2007, 6 in 2008, 7 in 2009, 5 in 2010, 10 in 2011, 13 in 2012, 15 in 2013 and 35 in 2014. The amount of works (including papers, PhD dissertations, and book chapters) that propose, test or discuss SCED data analysis illustrates the current relevance of the topic. Despite this increased attention to SCED analysis, a common requirement made by SCED article reviewers and journal editors has been to provide concrete recommendations regarding connecting specific conditions (e.g., design and data characteristics and purpose of the study) with appropriate SCED analytical techniques. In contrast to data analysis, guidelines for conducting SCEDs are already available in the form of rubrics and standards for assessing the methodological quality of SCED studies (see Maggin, Briesch, Chafouleas, Ferguson, & Clark, 2014, and Smith, 2012 for reviews). A similar broad overview regarding SCED data is lacking and this is why we provide it here.

The current SCED data analysis situation is well illustrated by Waddell, Nassar, and Gustafson's (2011) statement that "the problem of how to statistically analyze the data […] is perhaps the most confusing, daunting, and disjointed element of this experimental method" (p. 161). These authors also state that the amount of analytical techniques and formulae makes the issue even more confusing. With the current paper, we offer an overview and tentative recommendations based on the idea that there is no single analytical technique optimal for all situations (aims, data features, researchers' resources), but that one data analytic technique might be more appropriate in certain conditions compared to another.

In Search for Criteria and Recommendations

Solid and updated state of the art summaries of SCED analysis can be expected to be provided in Special Issues in journals with peer-review process. However, there are a few problems with the assumption that Special Issues might provide sound recommendations. First, the choice of focus of the Special Issue may not be based on the appropriateness of the techniques, but rather on: (a) a desire to provide the full spectrum of possibilities; (b) the guest editors knowing some of the techniques or some of the authors better than others;(c) the need to cover different topics as compared to previous special issues. Second, there may not be an explicit effort to point at the most appropriate analytical technique(s), as each research team presents the techniques it has been working on and the guest editors might not be willing to act as judges, due to (a) lack of knowledge; or (b) lack of journal space for a formal public discussion with the authors of the different papers[1]. Third, it is possible to find different foci and recommendations in the different special issues. Accordingly, an informal review of all the SCED data analysis special issues that we know of shows that some of the special issues pay more attention to techniques related to regression analysis (Journal of School Psychology in 2014; Shadish, 2014), whereas others focus on randomization tests (Journal of Contextual Behavioral Science in 2014; Vilardaga, 2014) or on nonoverlap indices (Journal of Behavioral Education in 2012, Burns, 2012; Remedial and Special Education in 2013; Maggin & Chafouleas, 2013). Another group of special issues covers a variety of techniques (Evidence-Based Communication Assessment and Intervention in 2008, Shadish, Rinsdskopf, & Hedges, 2008; Neuropsychological Rehabilitation in 2014, Evans, Gast, Perdices, & Manolov, 2014). Finally, two papers dealing with data analysis from special issues on SCED methodology ought to be mentioned. One of them (Vannest & Ninci, 2015) is focused

[1] It would not be ethical to invite a research team to submit a paper, review it, accept it for publication, and then publicly criticize the technique proposed/described without providing them the opportunity to respond.

on nonoverlap indices, whereas the other one (Gage & Lewis, 2013) reviewed several techniques before stressing the lack of agreement among researchers, stating that “a preference for standard mean difference, non-overlap, or regression-based approaches is also without empirical support” (p. 55). Thus, the lack of clear consensus (Kratochwill et al., 2010; Smith, 2012; Tate et al., 2013) and indications suggest that the current paper is necessary, as we consider that more discussion is needed apart from more research (Gage & Lewis, 2013).

Wolery, Busick, Reichow, and Barton (2010) were the first to suggest a set for criteria for SCED analytical procedures: (1) focus on the replication logic of SCED, (2) use all the data of the study; (3) estimate the magnitude of the effects across replications, (4) take into account all the characteristics of the data: level, trend, and variability; (5) show high agreement with careful visual analysis; (6) not violate the assumptions about the nature of the data, such as serial dependency; (7) have some method of allowing analyses of moderator variables. On the other hand, Manolov, Gast, Perdices, and Evans (2014) suggest that: (1) the technique chosen should reflect the aim of the analysis: statistical significance vs. effect size in a common metric vs. unstandardized effect size; (2) the output of the analysis should be easy to interpret: includes whether the quantification provided is meaningful and whether there are any interpretative benchmarks available; (3) the analysis should be easy to compute: includes hand calculation and software availability and user-friendliness; (4) the technique must take into account design requirements and data assumptions: includes randomization, absence of trend, absence of serial dependence; (5) the technique should be supported by evidence of appropriate performance with typical SCED data: includes both simulation studies and field tests with real data.

Additionally, it is possible to trace criteria closely associated to specific procedures. For instance, Kratochwill et al. (2013) suggest that an effect size should be comparable to the ones obtained in group design studies (in reference to the *d*-statistics by Hedges, Pustejvosky, & Shadish, 2012, 2013). Another criterion is that the analytical technique should not rely on the rarely possible random sampling to ensure the validity of inferential results (Dugard, 2014), with randomization tests being a procedure that meets this criterion.

We also consider relevant the following additional criteria: (1) in relation to the general recommendations for reporting results in Psychology (Wilkinson & Task Force on Statistical Inference, 1999) it could be useful for the technique to offer the possibility of constructing a confidence interval around the effect size estimate; (2) regarding design structures that meet evidence standards (Kratochwill et al., 2010), it would be necessary that the procedure is easily extensible to designs beyond AB, which is related to Wolery et al.'s criteria 2 and 3; (3) considering that visual analysis is commonly the initial step and sometimes the only step in data analysis (Perdices & Tate, 2009; Smith, 2012), it is possible to rephrase criterion 5 by Wolery et al. (2010) to "the types of data features on which visual analysis focuses (see Gast & Spriggs, 2010; Kratochwill et al., 2010) are also quantified by the procedure"; (4) according to the aim of the study, it may be useful that the technique offers the possibility of obtaining both results in the metric of the outcome and in comparable metrics: related to criterion 1 by Manolov et al. (2014); (5) according to the aim of the study, it may be also useful that the technique offers quantifications that can be used in a posterior meta-analysis.

Regarding the way in which the criteria are presented here, a flowchart representation would have required a solid basis for sorting the criteria, which is lacking. Another option was to provide an overall score, but such a score is not easily justified due to two complementary issues: (a) it is not clear that either all criteria can be considered equally important so that the same weight is assigned to all of them; and (b) in case some criteria are considered more important, it is not clear which would the appropriate weights be. Additionally, the importance of the criteria may depend on the aims of the study and the characteristics of the data at hand. In the present paper yes/no scoring is used, because practically all the criteria allow for such scoring. The criteria presented in the following section are the basis for the specific recommendations made and illustrated with real published single-case data sets later in the text.

Presentation of the Criteria

For each of the criteria listed in Tables 1 to 6, it is initially specified whether the criterion has to be met by a data analysis technique for it to be considered appropriate and useful or, in contrast, the importance of the criterion depends on the study aim or the characteristics of the data. The reader should not interpret the order of these criteria as indicating their relative importance. The criteria are rather grouped according to their object: (a) Type of quantification provided by the analytical technique: overall quantification (whether the technique is focused on two-phase AB comparisons and how it can be extended to more complex design structures including within-subject replication, such as ABAB, or across-subjects replication such as a multiple baseline design), possibility of obtaining a *p*-value and a confidence interval, metric in raw and/or comparable terms, overlap, change in level and/or in slope. (b) The data features taken into account: variability within a series and across series, baseline trend. (c) The conditions in which the techniques are appropriate: type of outcome to which the analytical technique is applicable (binary, ordinal, interval, ratio scale, including the need for continuous vs. discrete metric), type of functional form of the data to which the analytical technique is applicable (e.g., lack of trend or only linear trend can be handled), whether autocorrelation is dealt with explicitly and whether the application of the technique requires additional assumptions; (d) Meta-analytical features: how a meta-analysis can be performed (classical refers to using inverse variance as a weight and the possibility to use fixed effect and random effects models; whereas averaging refers to obtaining an unweighted average or an average weighted, for instance, by the number of measurements in the series [*n*-based weight], Kratochwill et al., 2010; Shadish et al., 2008) and the possibility to carry out moderator analysis; (e) The use of the analytical technique: what aspects relevant for visual analysis are quantified by the procedure, the availability of commercial and free software[2], the possibility of obtaining the numerical values by hand calculation, the ease of interpretation, and the assumptions and requirements for the correct functioning of the technique. An additional aspect that could have been considered is whether a technique is actually being used currently by applied researchers and to what extent. This information could serve as an indication of whether the use of the technique would require that applied researchers be specifically trained, but there is not sufficient evidence of the amount of use of all techniques[3], except for the frequently used visual analysis (Kratochwill, Levin, Horner, & Swoboda, 2014; Parker & Brossart, 2003) and the Percentage of nonoverlapping data (PND; Schlosser et al., 2008). (f) Evidence on performance – including the main conclusions and references that include tests of the analytical techniques, based either on simulated data with known characteristics (for most of the techniques) or real published data (mainly for the nonoverlap indices). For an in-depth review and discussion of the performance (i.e., Type I error rates, statistical power, bias, mean square error) of the techniques we refer the reader to the original work of the authors.

[2] See the papers by Bulté and Onghena (2013) and L-.T. Chen, Peng, and M. E. Chen (2015) for descriptions of software for SCED, as well as the tutorial on free software available for single-case data analysis available upon request from the first author and also at https://www.researchgate.net/publication/289098041_Single-case_data_analysis_Software_resources_for_applied_researchers

[3] It is not easy to assess the amount of use of different statistical techniques as many primary studies do not specify that (see Smith's, 2012, review) or the distinction between statistical techniques may even be missing from reviews (e.g., Perdices & Tate, 2009). It is even harder to quantify the use of statistical techniques in unpublished studies from professionals' everyday practice).

Applying the Criteria to Several Single-Case Analytical Techniques

Visual analysis. The classical defendants of visual analysis in SCED research (Parsonson & Baer, 1986, 1992) set the base for current recommendations for visual analysis (as summarized in Kratochwill et al., 2010) explaining a series of data features that need to be taken into account, such the inclusion of a sufficient number of data points, the evaluation of the baseline for stability or presence of trend, the assessment of change in trend within and/or across phases, the inspection of whether any change in level is immediate or delayed, considering variability within phases and across phases (overlap), and the assessment of the overall data pattern (e.g., whether the data from an ABAB design correspond to what is expected: similarly low desirable behavior in the A phases and similar improvements in the B phases) . The current leading texts on visual analysis focus on the same data features – for instance, Kratochwill et al. (2010) suggest inspecting six data aspects: level, trend, variability, immediacy of the effect, overlap, and consistency of data patterns across similar phases, also starting the analysis with the evaluation of the baseline data. The steps in performing visual analysis entail evaluating within-phase patterns in terms of variability, level, and trend, comparing adjacent phases and, finally, assessing whether there are enough demonstrations of the intervention effect.

Another way of carrying out visual analysis involves relying on visual aids such as mean, trend, and/or range lines that are supposed to help the analyst identify the main data features more objectively. Lane and Gast (2014) suggest not only representing those lines on the plot but also quantifying the mean or median level and trend within phases, as well as the PND (Scruggs, Mastropieri, & Casto, 1987) across phases. Their proposal also takes variability into account, by constructing a stability envelope around the mean or the trend line. Such envelopes are well-aligned with the remark made by Franklin, Gorman, Beasley, and Allison (1997) that outliers may affect the estimation of mean levels and trend could be confounded with variability. Structured criteria related to visual analysis were also suggested by Hagopian et al. (1997) and conservative dual criterion by Fisher, Kelley, and Lomas (2003), who propose drawing standard deviation lines above and below the mean level. In that sense, these lines can be used to detect changes beyond what is expected considering the baseline level and variability. These examples of “structured visual analysis” include formal decision rules about behavioral change and we distinguish this approach from the “traditional visual analysis” as described in Kratochwill et al. (2010), although there have been efforts to make the application of these latter standards more systematic (Maggin, Briesch, & Chafouleas, 2013).

Finally, visual analysis can be used for response-guided experimentation, that is, for evaluating continuously whether the behavior of interest has reached a predefined criterion (Franklin et al., 1997). A proposal intended to deal with the risks of detecting intervention effects too frequently, entails using data analysts who are blind to which participant is selected for treatment at each designated intervention time (Ferron & Jones, 2006), which is why this kind of visual analysis is called “masked” in Table 1. Ferron and Levin (2014) provide examples for the application of masked visual analysis to several SCED designs.

Table 1. Main characteristics of several single-case data analytical techniques: Visual analysis

Feature	**Desirable?**	**Traditional visual analysis**	**Structured visual analysis (visual aids and quantifications)**	**Masked (response-guided) visual analysis**
Overall quantification across replications	Yes	No quantification		Required combining probabilities
Metric in raw terms	If aimed	No	Yes	No
p value	If aimed	No	No	Yes
Confidence interval	Yes	No	No	No
Metric comparable across studies	If aimed	No	Yes (if PND is used)	Yes
Overlap	If aimed	Yes	Yes (if PND is used)	Yes
Change in level	According to data	Yes	Yes, quantitatively	Yes
Change in slope	According to data	Yes	Yes, quantitatively	Yes
Variability within a series	Yes	Yes	Stability envelopes Variability lines	Yes
Variability across series	If aimed	Yes	No	Yes
Baseline trend	According to data	Yes	Yes (Lane & Gast) No (Fisher et al.; Hagopian et al.)	Yes
Applicable to outcome?	According to data	All types	All types	All types
Type of functional form required?	According to data	Stable baseline (Kazdin, 1978)	No	Stable baseline
Deals with serial dependence	According to data	No	No	No
Additional assumptions	The fewer, the better	Agreement between judges	Normality for Hagopian et al.'s (1997) and Fisher et al.'s (2003) proposals	Analyst blind to intervention Random assignment of interventions to participants / behaviors/ settings in each measurement occasion
Option for meta-analysis	Yes	No	Averaging PND values	Combining probabilities
Moderator variable	Yes	No	Separate analysis	Separate analysis
Complements visual analysis	Yes	Not applicable	Not applicable	Not applicable
Software implementation	Yes	Yes	Yes	No
Free software	Yes	Yes	Yes	Yes
Hand calculation	If no (free) software	No calculation involved	No calculation involved	No calculation involved
Ease of interpretation	Yes	Yes	Yes	Yes
Evidence on performance	Yes	Unknown exactly how analysts perform visual analysis if not instructed. Not acceptable agreement in general, worse for experienced raters (Ninci et al., 2015) except Kahng et al. (2010)	Type I error rates control and increased accuracy (Stewart et al., 2007; Young & Daly III, 2016) for Fisher et al.'s (2003) proposal; increased agreement for Hagopian et al. (1997)	Increased Type I error rates if the analyst is not blind (Allison, Franklin, & Heshka, 1992), controlled Type I error rates if blind (Ferron & Jones, 2006)

Table 1 reflects the fact that visual analysis can be used to evaluate many data features, although according to the type of visual analysis only some of them are quantified. We have assumed that in masked visual analysis the focus is put on the same data features as in traditional visual analysis. Finally, the evaluation of the performance of traditional visual analysts is scattered over multiple publications (e.g., Danov & Symons, 2008; Matyas & Greenwood, 1990; Ottenbacher, 1993; Rojahn & Schulze, 1985) with the additional difficulty that it is not always clear what criteria visual analysts use, unless they are instructed to follow a specific protocol (Wolfe & Slocum, 2015) or decision rule (Fisher et al., 2003).

Nonoverlap indices. Several nonoverlap indices, expressing the result in percentages or proportions, have been proposed (see Parker, Vannest, & Davis, 2011, for a review) and they all quantify the proportion of measurements in the intervention phase improve the baseline measurements. We here focus on the Nonoverlap of all pairs (NAP; Parker & Vannest, 2009), Improvement rate difference (IRD; Parker, Vannest, & Brown, 2009), Tau-U (Parker, Vannest, Davis, & Sauber, 2011), as these are the most recent and complete techniques, as well as on the PND, as it is the most frequently used technique (Schlosser, Lee, & Wendt, 2008; Scruggs & Mastropieri, 2013) and on the Percentage of nonoverlapping corrected data (PNCD, Manolov & Solanas, 2009), given that is allows controlling for linear trend. The Percentage of all nonoverlapping data (Parker, Hagan-Burke, & Vannest, 2007) was not included, as its authors (Parker & Vannest, 2009) indicate that part of the computational procedure may be confusing and suggest NAP as an improvement. The percentage of data points exceeding the median (Ma, 2006) was also not included, as there is evidence of that it does not agree with visual analysis and that IRD performs better (Parker & Hagan-Burke, 2007).

Table 2. Main characteristics of several single-case data analytical techniques: Nonoverlap indices

Feature	Desirable?	PND	NAP	Tau	IRD	PNCD
Overall quantification across replications	Yes	Requires averaging ABs	Requires averaging ABs	Requires averaging ABs	Requires averaging ABs	Requires averaging ABs
p value	If aimed	No	Yes	Yes	No	No
Confidence interval	Yes	No	Yes	No	Yes	No
Metric in raw terms	If aimed	No	No	No	No	No
Metric comparable across studies	If aimed	Yes	Yes	Yes	Yes	Yes
Overlap	If aimed	Yes	Yes	Yes	Yes	Yes
Change in level	According to data	No	No	No	No	No
Change in slope	According to data	No	No	Yes	No	No
Variability within a series	Yes	No	Yes	Yes	Yes	No
Variability across series	If aimed	No	No	No	No	No
Baseline trend	According to data	No	No	Yes	No	Yes
Applicable to outcome?	According to data	Ordinal	Ordinal	Ordinal	Ordinal	Interval
Type of functional form required?	According to data	No trend	No trend	Monotonic trend, if controlled	No trend	No or linear trend
Deals with serial dependence	According to data	No	No	No	No	No
Additional assumptions	The fewer, the better	No baseline outliers	No trend	None	None	None
Option for meta-analysis	Yes	Averaging	Averaging	Averaging	Averaging	Averaging
Moderator variables	Yes	Separate analyses	Separate analyses	Separate analyses	Separate analyses	Separate analysis
Complements visual analysis	Yes	Overlap	Overlap	Overlap; trend	Not directly	Overlap, trend
Software	Yes	Yes	Yes	Yes	Yes	Yes
Free software	Yes	Yes	Yes	Yes	Yes	Yes
Hand calculation	Yes, if no free software	Yes	In some cases	In some cases	In some cases	In some cases
Ease of interpretation	Yes	Yes	Yes	Yes	Yes	Yes
Evidence on performance	Yes	Mixed (Scruggs & Mastropieri, 2013); Allison & Gorman (1994)	Related to R^2; no floor effects (Parker & Vannest, 2009)	No floor or ceiling effect, no effect of autocorrelation (Parker et al., 2011)	Reliably calculated; no floor effects (Parker et al., 2009)	Controls effectively for linear trend, not affected by autocorrelation (Manolov & Solanas, 2009)

Note. PND – percentage of nonoverlapping data. NAP – nonoverlap of all pairs. IRD – improvement rate difference. PNCD – percentage of nonoverlapping corrected data

Table 2 identifies the indices controlling for baseline trend control (Tau deals with monotonic trends and PNCD only with an approximately linear trend) and also the ones that allow obtaining *p*-values (NAP, Tau) or confidence intervals (Tau), if desired by the researcher.

Descriptive indices quantifying changes in level and in slope. The percentage change index (PCI; Hershberger, Wallace, Green, & Marquis, 1999; or percentage reduction data, as referred to by Wendt, 2009) quantifies the difference between the last three baseline and intervention phase measurements, relative to the former. A similar index focusing on all the measurements is called Mean baseline reduction (MBLR) by Campbell (2004). Another procedure quantifying change in level, but in the same metric as the dependent variable rather than in terms of a percentage is Slope and level change (SLC; Solanas, Manolov, & Onghena, 2010). In SLC, baseline linear trend, before quantifying change in slope, and finally net change in level. The Mean phase difference (MPD; Manolov & Solanas, 2013) can be conceptualized as quantifying change in level and slope jointly, as it compares the projected baseline trend with the actually obtained intervention phase measurements. Both techniques are extended in Manolov and Rochat (2015) to allow for standardizing and application beyond two-phase designs.

Finally, a specific way of quantifying a change in level is the percentage of zero data (Scotti, Evans, Meyer, & Walker, 1991): the percentage of measurement occasions for which the behavior does not appear once the first problem-free measurement occasion is achieved.

Table 3. Main characteristics of several single-case data analytical techniques: Descriptive indices quantifying changes in level and in slope

Feature	Desirable?	SLC	MPD	PCI / MBLR	PZD
Overall quantification across replications	Yes	Yes, via extensions	Yes, via extensions	Requires averaging ABs	Requires averaging ABs
p value	If aimed	No	No	No	No
Confidence interval	Yes	No	No	No	No
Metric in raw terms	If aimed	Yes	Yes	No	No
Metric comparable across studies	If aimed	Yes; via extensions	Yes; via extensions	Yes	Yes
Overlap	If aimed	No	No	No	No
Change in level	According to data	Yes	Not specifically; Overall difference	Yes	No
Change in slope	According to data	Yes	Not specifically; Overall difference	No	No
Variability within a series	Yes	Yes; Standardized version	Yes; Standardized version	No	No
Variability across series	If aimed	In weights (Manolov & Rochat, 2015)	In weights (Manolov & Rochat, 2015)	No	No
Baseline trend	According to data	Yes	Yes	No	No
Applicable to outcome?	According to data	Interval scale	Interval scale	Ratio scale	Ratio scale
Type of functional form required?	According to data	No or linear trend	No or linear trend	No trend	No specific pattern
Deals with serial dependence	According to data	No	No	No	No
Additional assumptions	The fewer, the better	None	None	None	Behavior takes place before the intervention
Option for meta-analysis	Yes	Index extensions	Index extensions	Averaging Classical	Averaging
Moderator variables	Yes	Separate analyses	Separate analyses	Separate analyses	Separate analyses
Complements visual analysis	Yes	Change in slope & level	Projected vs. actual data	Change in level	Zero data is not a specific part of visual analysis
Software	Yes	Yes	Yes	Yes	Yes
Free software	Yes	Yes	Yes	Yes	Yes
Hand calculation	If no (free) software	In some cases	In some cases	In some cases	Yes
Ease of interpretation	Yes	Yes	Yes	Yes	Yes
Evidence on performance	Yes	No bias; small effect of serial dependence (Solanas et al., 2010)	No bias; small effect of serial dependence (Manolov & Solanas, 2013); good sensitivity (Solomon et al., 2015)	Not found	Not found

Note. SLC - slope and level change. MPD – mean phase difference. PCI – percentage change index. MBLR – mean baseline reduction. PZD – percentage zero data

The information presented in Table 3 is aimed to aid researchers choose a technique according to whether only change in level is the focus (PCI), change in level and in slope separately (SLC) or an overall quantification of both is desired (MPD). Another difference is that Hershberger et al. (1999) provide an expression for the variance of PCI making classical (inverse variance) weighting possible in a meta-analysis. Finally, the combination of MPD and SLC values takes variability across replications into account.

Standardized mean difference. Standardized mean difference (SMD) indices could have been included in the previous section about indices quantifying changes in level, but they offer the possibility for inferential analysis beyond description, thanks to the fact that their sampling distributions are known or can be approximated, under certain assumptions (Borenstein, 2009).

An initial application of SMDs to SCED data focused on the between-group designs *d*-statistic using pooled standard deviation in the denominator (Cohen, 1992) or only the baseline standard deviation, referred to as Glass' Δ or delta (Busk & Serlin, 1992; Glass, McGaw, & Smith, 1981). Given that the inferential use of these indices in SCED is problematic (Beretvas & Chung, 2008), an alternative (referred to as the HPS *d*-statistics here) was proposed by Hedges, Pustejovsky, and Shadish (2012, 2013) specifically for SCED. These indices were developed to take into account autocorrelation and between-subjects variability, apart from within-subject variability. The HPS *d*-statistics are comparable to Cohen's *d* as obtained from between-group design studies. In a specific domain such as neuropsychology another version of the classical SMD was proposed by Beeson and Robey (2006), comparing a maintenance phase (A_2) with the initial (A_1), but due to its more restricted application, we do not included in Table 4.

Finally, a *d*-statistic can be obtained on the basis of autoregressive integrated moving average (ARIMA) models, the application of which usually requires long data series and complex modelling (Brossart, Parker, Olson, & Mahadevan, 2006). A possibility to avoid the initial model identification (Harrop & Velicer, 1985) and a recent illustration (Harrington & Velicer, 2015) prompted us to include this analytical option here, as it could offer descriptive and inferential information about a change in level and in slope, while dealing with autocorrelation. The summary provided in the corresponding table is based on this recent application.

Table 4. Main characteristics of several single-case data analytical techniques: Standardized mean difference indices

Feature	**Desirable?**	**Glass' Δ (delta)**	**Cohen's *d***	**HPS *d*-stat.**	**ITSA**
Overall quantification across replications	Yes	Requires averaging ABs	Requires averaging ABs	Yes for AB^k and MBD	Yes
p value	If aimed	For group designs only	For group designs only	No	Yes
Confidence interval	Yes	For group designs only	For group designs only	Yes	Yes
Metric in raw terms	If aimed	Raw mean difference	Raw mean difference	Raw mean difference	No
Metric comparable across studies	If aimed	Yes	Yes	Yes	Yes
Overlap	If aimed	No	Can be converted, if normally distributed	No	No
Change in level	According to data	Yes	Yes	Yes	Yes
Change in slope	According to data	No	No	No	Yes
Variability within a series	Yes	Yes	Yes	Yes	No
Variability across series	If aimed	No	No	Yes	No
Baseline trend	According to data	No	No	No	Yes
Applicable to outcome?	According to data	Continuous	Continuous	Continuous	Continuous
Type of functional form required?	According to data	No trend	No trend	No trend	No trend for descriptive measure
Deals with serial dependence	According to data	No	No	Yes	Yes
Additional assumptions	The fewer, the better	Baseline data are not constant	Similar variance across phases	Normality	Model correctly specified
Option for meta-analysis	Yes	Averaging (Classical for group designs)	Averaging (Classical for group designs)	Classical	Averaging
Moderator variable	Yes	Separate analyses	Separate analyses	Separate analyses	Separate analyses
Complements visual analysis	Yes	Average change in level	Average change in level	Average change in level	Change in level and in slope
Software implementation	Yes	Yes	Yes	Yes	Yes
Free software	Yes	Yes	Yes	Yes	No
Hand calculation	If no (free) software	Yes, for short data series	Yes, for short data serie	No	No
Ease of interpretation	Yes	Yes	Yes	Moderate	Moderate
Evidence on performance	Yes	Somewhat affected by autocorrelation; distinguishes effect vs. no (Manolov & Solanas, 2008); consistency with PND and MBLR (Olive & Smith, 2005)	Somewhat affected by autocorrelation; distinguishes effect vs. no (Manolov & Solanas, 2008) ; consistency with PND and MBLR (Olive & Smith, 2005)	OK, can improve for variance estimator (Hedges et al., 2012)	Model converging for almost all data sets; low agreement with visual analysis (Harrington & Velicer,2015)

Note. HPS refers to the initials of the surnames of authors of this version of *d*-statistic: Hedges, Pustejovsky, and Shadish (2012, 2013). ITSA – interrupted time series analysis referring to autoregressive moving integrated moving average (ARIMA) models (1,0,0), (3,0,0) or (5,0,0) .

The information presented in Table 4 is aimed to aid researchers choose a technique according to whether inferential information (*p*-values, confidence intervals) and explicit accounting for autocorrelation is desired with the “cost” of using a more complex technique (the HPS *d*-statistic or ITSA – interrupted time series analysis using specific ARIMA models) or simpler quantifications are preferred.

Procedures based on regression analysis. Several proposals have been made for using regression analysis in the context of SCED, starting several decades ago (Gorsuch, 1983). Given that the proposals of Gorsuch (1983) were found to yield excessively low effect size values (Brossart et al., 2006; Manolov, Arnau, Solanas, & Bono, 2010), whereas the proposal by Allison and Gorman (1993) provides too large R^2 values even in absence of effect (Brossart et al., 2006; Manolov & Solanas, 2008), neither of these options is included in Table 5. The focus is rather on a piecewise regression (Center, Casey, & Skiba, 1985-1986), making possible obtaining quantifications of immediate change in level and change in slope, while controlling for baseline trend and on a more recent proposal based on generalized least squares estimation (Swaminathan, Rogers, &Horner, 2014), which allows controlling for autocorrelation and obtaining a quantification expressed as an average mean difference.

Multilevel models (including hierarchical linear models), as an extension of the piecewise regression, constitute a technique based on the nesting of the data (measurements within individuals and individuals within studies), making it possible to quantify different types of effects (e.g., immediate change in level, change in slope, amount of variation within individuals, between individuals and between studies), while also taking autocorrelation into account. A review of the possible uses of multilevel models for analysis and meta-analysis of SCED data is available in Moeyaert, Ferron, Beretvas, and Van Den Noortgate (2014).

An index developed by Pustejovsky, Hedges, and Shadish (2014; hereinafter, the PHS *d*-statistic) has the same underlying idea as the HPS *d*-statistic, but the former uses restricted maximum likelihood estimation and offers the possibility to obtain a standardized mean difference from a variety of multilevel models fitted (e.g., controlling for baseline trend and taking into account change in slope).

Table 5. Main characteristics of several single-case data analytical techniques: Procedures based on regression analysis (including hierarchical linear models).

Feature	**Desirable?**	**Multilevel models**	**PHS *d*-statistic**	**Piecewise regression**	**Generalized least squares regression**
Overall quantification across replications	Yes	Yes via modeling	Yes via modeling	Requires averaging ABs	Requires averaging ABs
p value	If aimed	Yes; for the effects modeled	For comparisons between models (aspects modelled)	Can be obtained for regression coefficients	Yes
Confidence interval	Yes	Yes	Yes	Yes	Yes
Metric in raw terms	If aimed	Yes	No	Yes	Yes
Metric comparable across studies	If aimed	Yes, after standardizing	Yes	Yes, after standardizing	Yes, after standardizing
Overlap	If aimed	No	No	No	No
Change in level	According to data	According to model	According to model	Yes	Yes, if flat slopes in the phases compared
Change in slope	According to data	According to model	According to model	Yes	Yes
Variability within a series	Yes	Yes	No	Yes; standardized version	Yes; standardized version
Variability across series	If aimed	Yes	Yes	No	No
Baseline trend	According to data	Yes	Yes	Yes	Yes
Applicable to outcome?	According to data	Continuous, counts and binary (Shadish et al., 2013)	Continuous	Continuous	Continuous, counts and binary (Swaminathan et al., 2014)
Type of functional form required?	According to data	Can handle linear, quadratic trend, etc.	Can handle linear, quadratic trend, etc.	Linear trend	Can handle linear, quadratic trend, etc.
Deals with serial dependence	According to data	Yes	Yes	No	Yes
Additional assumptions	The fewer, the better	Normality, if outcome is continuous	Normality	Normality	Normality, if outcome is continuous
Option for meta-analysis	Yes	Part of the model	Classical	Averaging	Averaging
Moderator variable	Yes	Including	Including covariates	Separate analyses	Separate analyses

		covariates			
Complements visual analysis	Yes	Aspects according to model	Aspects according to model	Immediate change; change in slope	Difference in fitted slope lines
Software implementation	Yes	Yes	Yes	Yes	Yes
Free software	Yes	Yes	Yes	Yes	Yes
Hand calculation	If no (free) software	No	No	No	No
Ease of interpretation	Yes	Moderate	Moderate	Yes	Moderate
Evidence on performance	Yes	OK estimation of effects, improvable for variance (Moeyaert et al., 2014).	OK for bias, precision can be improved (Pustejovsky et al., 2014).	Statistical power; free of autocorrelation (Parker & Brossart, 2003)	Slight effect of nonnormality and heteroscedasticity (Manolov & Solanas, 2013); less than optimal sensitivity (Solomon et al., 2015)

Note. PHS: initials of the surnames of authors of this version of *d*-statistic: Pustejovsky, Hedges, and Shadish (2014)

The information presented in Table 5 is aimed to aid researchers choose a technique according to whether the nested structure of the data is taken into account with the possibility to obtain directly an overall quantification across replications (multilevel models and PHS *d*-statistic) or separate quantifications for each two-phase comparison are performed (piecewise and GLS regression). For some proposals (multilevel, GLS) the possibility to handle data that are continuous has been explicitly discussed. Finally, note that multiple papers provide evidence regarding the performance of multilevel models (e.g., Ferron, Farmer, & Owens, 2010; Moeyaert, Ugille, Beretvas, Ferron, & Van Den Noortgate, 2013; Ugille, Moeyaert, Beretvas, Ferron, & Van Den Noortgate, 2012, 2014), apart from the ones included in Table 5.

Procedures whose main output is a *p*-value. Randomization tests offer a way, via data re-arranging, of obtaining statistical significance of the results based on random assignment in the design (e.g., choosing at random the start of the intervention condition). This procedure can be applied to a variety of designs and situations (Levin, Ferron, & Kratochwill, 2012) and it is also possible to use an effect size index as a test statistic (Heyvaert & Onghena, 2014a). Several studies provide evidence on randomization tests (e.g., Ferron & Onghena, 1996; Ferron &

Sentovich, 2002; Levin, Lall, & Kratochwill, 2011; Sierra, Solanas, & Quera, 2005), beyond the summary papers included in Table 6.

Simulation modeling analysis (Borckardt et al., 2008) offers a way, via bootstrap and/or Monte Carlo simulation, of obtaining the statistical significance of the difference between conditions quantified via the point biserial correlation coefficient (i.e., Pearson's correlation applied to the case in which one of the variables is dichotomous, as when 0 marks the baseline ad 1 the intervention phase).

Table 6. Main characteristics of several single-case data analytical techniques: Procedures whose main output is a *p* value

Feature	**Desirable?**	**Simulation modeling analysis**	**Randomization test**
Overall quantification across replications	Yes	Requires combining AB-probabilities	Yes (see)
p value	If aimed	No	According to test statistic chosen
Confidence interval	Yes	Yes	Yes
Metric in raw terms	If aimed	No	According to test statistic chosen
Metric comparable across studies	If aimed	Yes: correlation	According to test statistic chosen
Overlap	If aimed	No	According to test statistic
Change in level	According to data	Yes	According to test statistic
Change in slope	According to data	No	According to test statistic
Variability within a series	Yes	No	No
Variability across series	If aimed	No	No
Baseline trend	According to data	Yes	According to test statistic
Applicable to outcome?	According to data	Continuous	According to test statistic
Type of functional form required?	According to data	No or linear trend	According to test statistic
Deals with serial dependence	According to data	Yes	No
Additional assumptions	The fewer, the better	Normality if Monte Carlo is used instead of bootstrap	Randomization
Option for meta-analysis	Yes	Combining probabilities	Combining probabilities
Moderator variables	Yes	Separate analyses	Separate analyses
Complements visual analysis	Yes	Chance likelihood of the change in level	According to test statistic chosen
Software	Yes	Yes	Yes
Free software	Yes	Yes	Yes
Hand calculation	If no (free) software	No	In some specific cases
Ease of interpretation	Yes	Moderate	Moderate
Evidence on performance	Yes	Positive (Borckardt & Nash, 2014), but possibly insufficient	Positive regarding Type I error rates, improvable power (Heyvaert & Onghena, 2014b)

The information presented in Table 6 highlights the difference between the procedures for obtaining *p* values. Specifically, the randomization test can be applied to several types of design and it is also possible to choose a test statistic according to the effect expected, but it requires that randomization is present in the design of the study. In comparison, a strength of simulation modeling analysis is that it deals explicitly with autocorrelation. Other procedures such as ITSE and ITSACORR can also provide *p*-values and they have been recommended elsewhere (Robey, Schultz, Crawford, & Sinner, 1999), but we do not include them here, due to the evidence available on their inadequate performance (Huitema, 2004; Huitema, McKean, & Laraway, 2007) and lack of relation to the results of other techniques (Brossart et al., 2006).

Presentation of the Recommendations

In the following we have included a series of criteria in a tabular format. For each of the criteria, we have included an initial column entitled “Desirable?” which specifies whether the criterion has to be met necessarily by a technique for it to be considered appropriate and useful (answered by “Yes”) or the importance of the criterion depends on the specific aim of the study or the effect expected (“If aimed”), or the characteristics of the data (“According to data”).

Given that it is hard to establish an order of importance of the criteria, or to assign meaningful numerical weights, and given that the relevance of some of the criteria is subjected to the aims of the study or the characteristics of the data at hand, we cannot point unequivocally at the most appropriate SCED analytical technique. Nevertheless, depending on the desirable features, one technique is more appropriate than another. In the current section, a list of recommendations is offered pointing at specific aims (e.g., descriptive vs. inferential) and data patterns (e.g., amount of variability, presence of trend) and the most appropriate techniques in these conditions. This list should not be considered comprehensive, as it is possible that a specific combination of data characteristics and study aims is not contemplated here. We consider that an analytical technique is chosen on the basis of three main pillars: (a) the aims of the study and the type of quantification that is desired *a priori*; (b) the characteristics of the data as assessed by visual inspection, as well as the assumptions one is willing to make about the data; and (c) the knowledge and computational resources one needs to perform the analysis. In what follows we will discuss these three pillars more in detail, making a case for the need to base the choice of an analytical technique on one of them: the features of the data actually obtained.

According to Researchers’ Aim

This first pillar for making a choice amongst data analytical techniques states that a technique can be chosen according to what the researcher considers a meaningful comparison between conditions. Analytical aims can include the following: assess a variety of data features in order to have a detailed understanding of the data, obtain a global quantification of the results that summarizes all the data, obtain specific quantifications for each relevant comparison, evaluate the statistical significance of the results, and perform a meta-analysis (see Online Table 7). There can be different ways of accomplishing these aims and choosing one technique or another is ultimately related to the features of the data. Moreover, it is possible that the characteristics of the data do not allow accomplishing the analytical aim. For instance, if obtaining statistical significance is the goal, it may not be possible to obtain it via a randomization test in absence of randomization, via a multilevel analysis if the algorithm does not converge to a solution, or via Tau, if autocorrelation is considered problematic for the *p* values associated with it. Thus, it may not be possible to choose only on the basis of initial aims without taking data features into account.

Table 7. Recommendations for choosing analytical techniques according to aim.

Situation	Example	Analysis	Justification
If you want to take into account a variety of data aspects (stability; trend; floor and ceiling effects; outliers; type of intervention effect) and/or assess the whole data pattern	Figure 1 shows variable baseline data with potential trend and ceiling effect in the intervention phase	Visual analysis Visual aids for systematic assessment (Kratochwill et al., 2010; Lane & Gast, 2014)	Establish effectiveness Formal decision rules
If you want to have an overall quantification of the effect that summarizes within-subject or across subjects replications	Figure 2 shows a replication of an ABAB design across participants, without a clear overall data pattern	HPS *d*-statistic: may require detrending; at least three replications across cases PHS *d*-statistic (standardized) or the underlying multilevel model (also raw): requires greater statistical knowledge Randomization test: requires randomization in the design of the study	Impossible to handle visually Models different data aspects (e.g., trends) Statistical significance for the whole study
If you want the result to be expressed in a statistically sound metric that is comparable across studies	Burns et al. (2012) and Jamieson et al. (2014) meta-analyze group design and single-case studies, using NAP obtained from *phi* and convertible to Cohen's *d*	HPS *d*-statistic: may require detrending; at least three replications across cases PHS *d*-statistic: models different data aspects; requires greater statistical knowledge *d*-statistic from OLS regression: does not control for autocorrelation *d*-statistic from generalized least squares regression: requires data transformation	Makes possible combining SCED and group-design studies together, if desired
If you want to have specific quantifications for each comparison between pairs of phases + decide yourself how to combine these quantifications (e.g., unweighted or weighted mean, median) + you want the result to be expressed in a metric meaningful for you	Strain et al. (1998) argue for using only the initial AB; Moss & Nicholas (2006) construct their own practically meaningful measure	Mean baseline reduction Mean phase difference: if you want to compare projected baseline level and trend with actual treatment phase measurements; can be standardized Slope and level change: can be standardized Piecewise regression: can be standardized	If no trend Comparable across studies If the changes in level and in slope are in the same direction Average changes in slope & level Change in slope and immediate change in level
If you want to obtain statistical significance + you can choose at	Winkens et al. (2014)	Randomization test: choose intervention start point at random for an AB design	Possibility to use a meaningful effect size as a test statistic

random when to change conditions	Sil et al. (2013)	Randomization test: restricted randomization for alternating treatments design	
If you want to obtain statistical significance + you cannot choose at random when to change conditions	Tunnard & Wilson (2014) used Tau quantifying the difference between pairs of conditions and for its *p* value	Multilevel models: modeling different data aspects	*p* values for effects and variances
		Tau: controlling for trend	*p* values for nonoverlap
		Simulation modeling analysis: taking autocorrelation into account	*p* value for point-biserial correlation
If you want to carry out a statistically sound meta-analysis (incl. weighted average, confidence intervals, heterogeneity tests)	Graves, Roberts, Rapoff, & Boyer (2010) use Cohen's *d* and its standard error for confidence intervals and its inverse variance for weighting	Multilevel models: consider nested structure of the data (measurements within participants within studies)	Quantify effects and variances
		HPS *d*-statistic: for designs as multiple-baseline or $(AB)^k$	Inverse variance weight
		PHS *d*-statistic: after multilevel analysis	Inverse variance weight
		Randomization tests: by combining *p* values (e.g., Edgington, 1972).	Rosenthal (1978)
If you want to carry out a meta-analysis, after deciding which comparisons to choose (e.g., Vannest, Harrison, Temple-Harvey, Ramsey, & Parker, 2011, choose specific planned comparisons) and how to combine the quantifications available for two-phase comparisons	Heinicke and Carr (2014): first baseline and last treatment phase; Jamieson et al. (2014): pool all data; Parker et al. (2011): use initial AB only	Simulation modeling analysis	Combining probabilities
		Nonoverlap indices	Mean, median, *n*-based weight
		Mean baseline reduction	Mean, median, *n*-based weight
		Percentage change index: see Hershberger et al. (1999) for its variance formula	Mean, median, *n*-based weight or inverse variance weight
		Slope and level change and Mean phase difference (Manolov & Rochat, 2015)	Two possible weighting strategies

Note. All figures numbers refer to the Online Repository.

According to the Characteristics of the Data

This second pillar responds to the fact that the analytical techniques are more appropriate in certain conditions, taking into account how the quantifications are obtained (e.g., whether baseline data are expected to be stable or whether a linear trend is expected to be clearly identifiable). Another reason for this approach is the common use of visual analysis as an initial step in data analysis (Kratochwill et al., 2010; see Davis et al., 2013 for an example), influencing the choice of an analytical technique and helping validate its results (Parker, Cryer, & Byrns, 2006). We include this pillar in the second place here only because the exact characteristics of the data are known after the researcher has already defined his/her analytical aims (included in pillar 1) and collected the data. The recommendations according to this pillar are presented in Online Table 8, which explains how Online Figures 3 to 11 illustrate the situations calling for different analytical options.

Table 8. Recommendations for choosing analytical techniques according to data features.

Situation	**Example**	**Analysis**	**Justification**
If the data pattern is clear + the measurements obtained allow for meaningful interpretations	Figures 3: achieving optimal performance; Figure 4: clear separation between conditions; Figure 5: achieving predefined criterion	Visual analysis for your within-study interpretation of results	Obvious differences
		Quantification according to the data pattern and design structure (change in level or in slope)	Communication Future meta-analysis
If you want to carry out a visual analysis using visual aids + the data show a clear pattern	Figure 3: focusing on within-phase level	Represent graphically a measure of central tendency (mean, median)	Stable data
		Standardized mean difference or Mean baseline reduction index or Percentage change data	Helps interpretation
	Figure 6: quantify the rate of improvement (clear intervention data trend)	Represent trend graphically (e.g., using the split middle method; Miller, 1985, without the binomial test; Crosbie, 1987, or using ordinary least squares regression)	Baseline trend
		Slope and level change (compares mean levels as well) or Piecewise regression (quantifies immediate effect as well)	Quantify change in slope
If you want to carry out a visual analysis using visual aids + the data show considerable variability + you are willing to explore whether the data in a phase fall generally in the range of values expected in case no behavioral change has taken place	Figure 7: variable data with no trend	Standard deviation bands (Pfadt & Wheeler, 1995), as a visual aid or a statistical tool via the conservative dual criterion (Fisher, Kelley, & Lomas, 2003)	Stable baseline with no clear trend
	Figure 8, upper panel: variable but improving trend	Trend stability envelope (Gast & Spriggs, 2010; Manolov, Sierra, Solanas, & Botella, 2014).	Baseline trend and variability to take into account
	Figure 2: variable data; change in variability as type of effect	Range lines as a very general (but sensitive to outliers) approach, especially, if an overlap measure is to be used	Level and trend do not represent well the data
If you want to carry out visual analysis using visual aids + the baseline trend in the data is not clear	Figure 1: possible baseline trend with a lot of variability	Compare the fit of the different methods to the data: split-middle, ordinary least squares (e.g., piecewise regression), differencing (e.g., Mean phase difference), trisplit method (Parker, Vannest, & Davis, 2014), running medians (Tukey, 1977)	Gain a better understanding of the trend in the data
If there is considerable	Figure 9: several	HPS *d*-statistic (Hedges, Pustejovsky, &	No clear trends

data variability in the measurements within a study + you want to take it into account when comparing the results across studies + you want to obtain an overall quantification	replications with unequal effect Figure 10: several replications per case; difficult task for visual analysis only	Shadish, 2012, 2013) PHS *d*-statistic (Pustejovsky, Hedges, & Shadish, 2014) on the basis of multilevel modeling	and no need for initial detrending Need to model intercepts, trends, and treatment effects as random factors
If the data variability in the measurements within a study is considerable + you want to know how much variability is there between cases	Figure 10: nested data (measurements within baselines within cases); varying effects	Multilevel models	Quantify the variability of data patterns between cases and the unexplained variance
If the within-phase data variability in the measurement in one or several two-phase comparisons is considerable	Figure 11: variable data with no clear trend and certain overlap to be quantified	Nonoverlap of all pairs	Estimates of within-phase trend and level are not expected to be informative
If baseline trend is sufficiently clear + seems reasonable to project it into the subsequent phase	Figure 8, lower panel: clear improvement, whose projection stays within reasonable limits for the data gathered	Generalized least squares regression: if you want to control for autocorrelation and compare two sets of predicted values; possibility to standardize the result. Mean phase difference: if you want to compare actual with predicted intervention phase values; possibility to standardize the result.	Baseline phase is not too short; the intervention phase is not too long, and the projection will not lead to impossible values
If you are concerned about autocorrelation + want to have the result expressed in a comparable metric + make quantitative integrations possible	Carey & Matyas, (2005) focus on serial dependence and variability; Solomon, Klein, & Politylo (2012) assess autocorrelation as well	Multilevel models and PHS *d*-statistic: requires data to be previously standardized (Van den Noortgate & Onghena, 2008) HPS *d*-statistic: may require initial detrending Generalized least squares regression: may require transforming the data iteratively Simulation modeling analysis using point biserial correlation: does not handle trend	Can handle trend and include moderators Does not require pre-standardizing Can handle trend Separate quantification for each AB

Note. All figures numbers refer to the Online Repository.

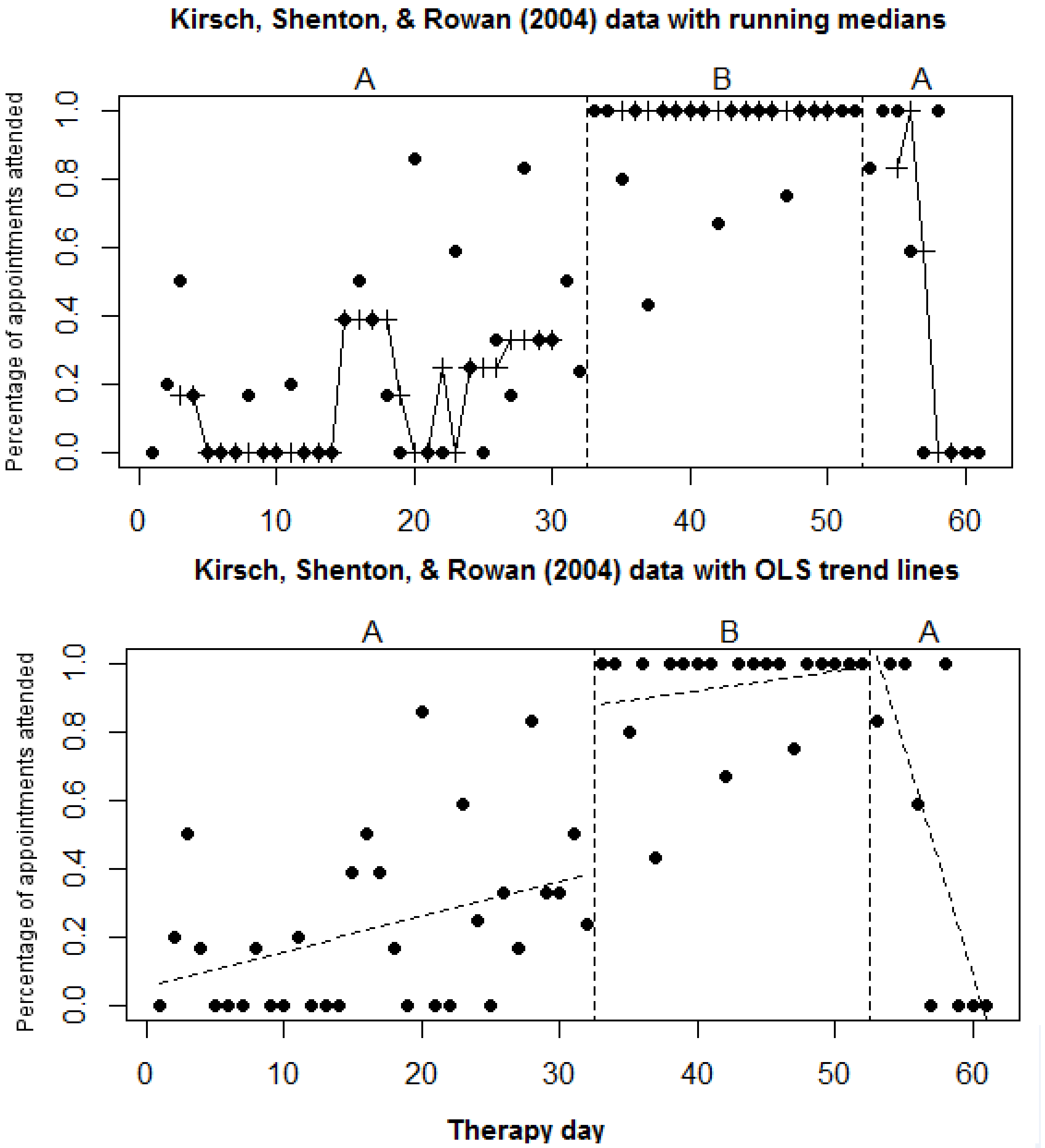


Figure 1. Data collected by Kirsch, Shenton, & Rowan (2004) on the effect of a paging system for prospective activity impairments in a participant with traumatic brain injury. OLS: ordinary least squares.

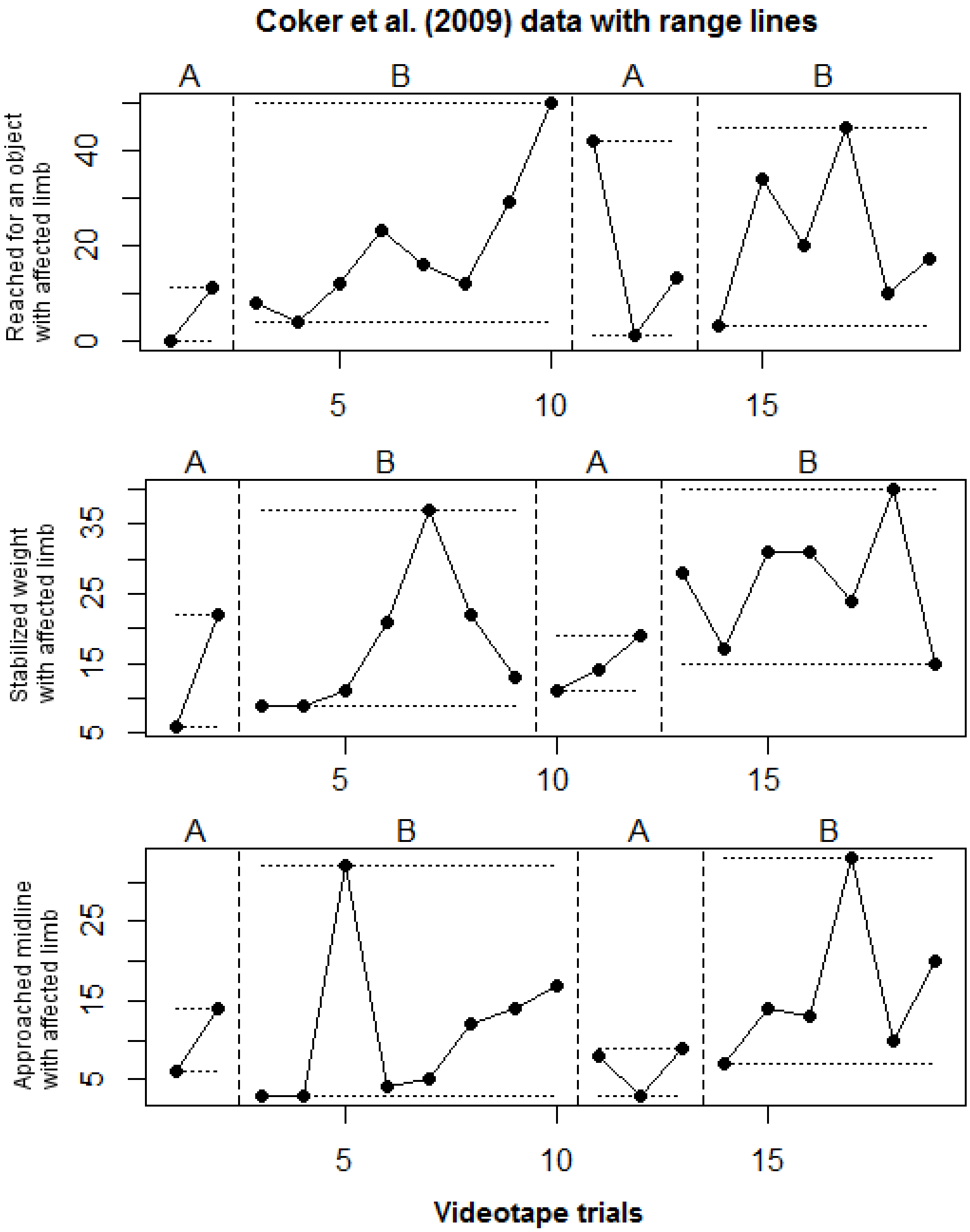


Figure 2. Data gathered by Coker et al. (2009) on the effects of constraint-induced movement therapy for a child less than one year of age with a diagnosis of hemiplegic cerebral palsy.

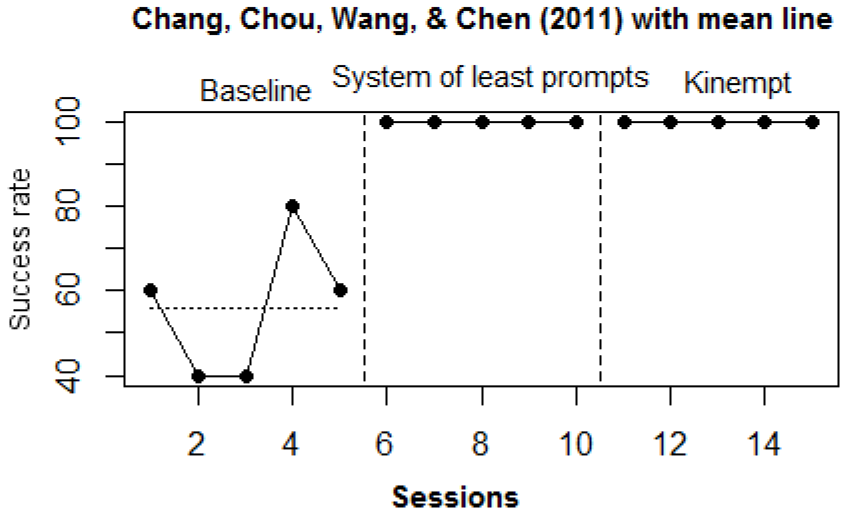
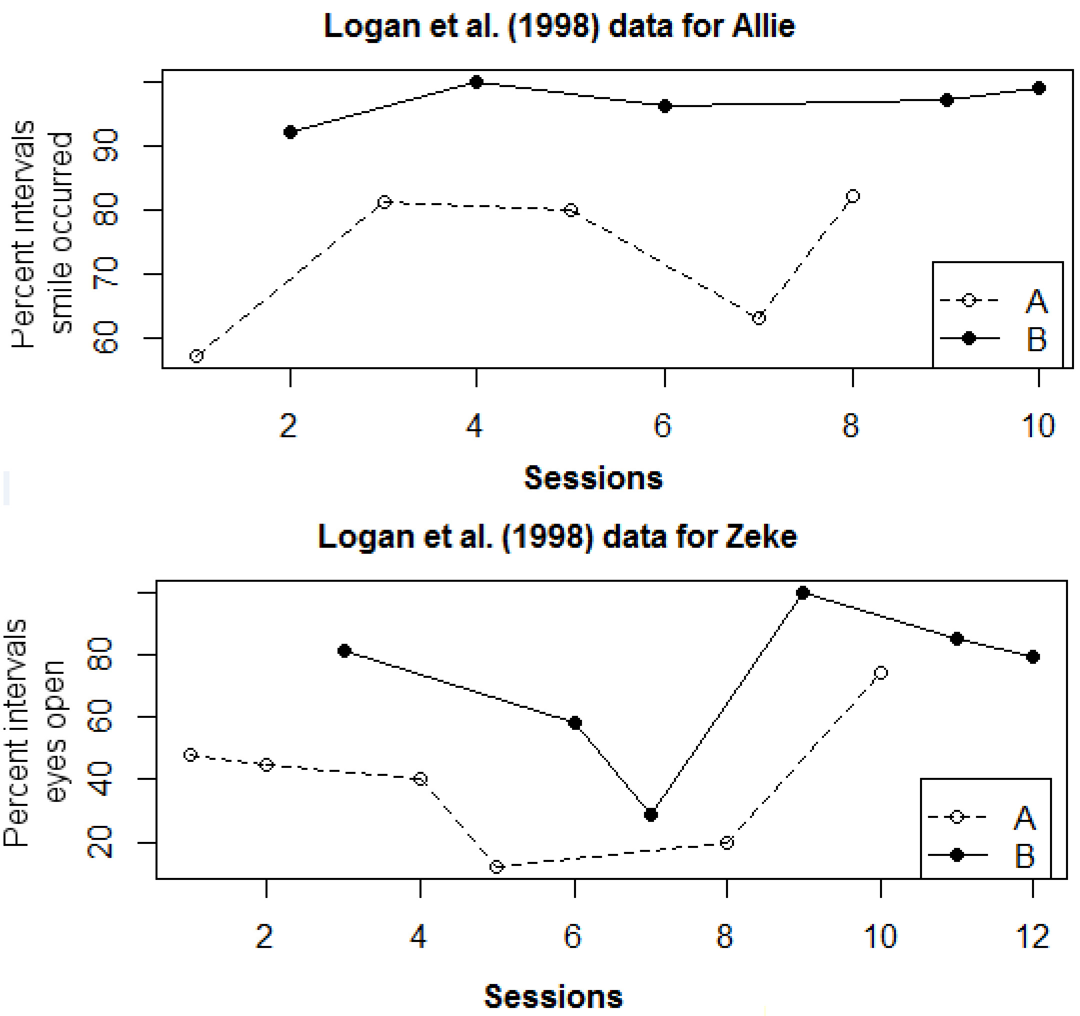


Figure 3. Data gathered by Chang et al. (2011) on a vocational task prompting system called Kinempt for individuals with cognitive impairments.


Logan et al. (1998) data for Allie

Logan et al. (1998) data for Zeke

Figure 4. Data gathered by Logan et al. (1998) on the impact of peers (A: peers with disabilities; B: typical peers) on the perceived happiness of students with profound multiple disabilities multiple.

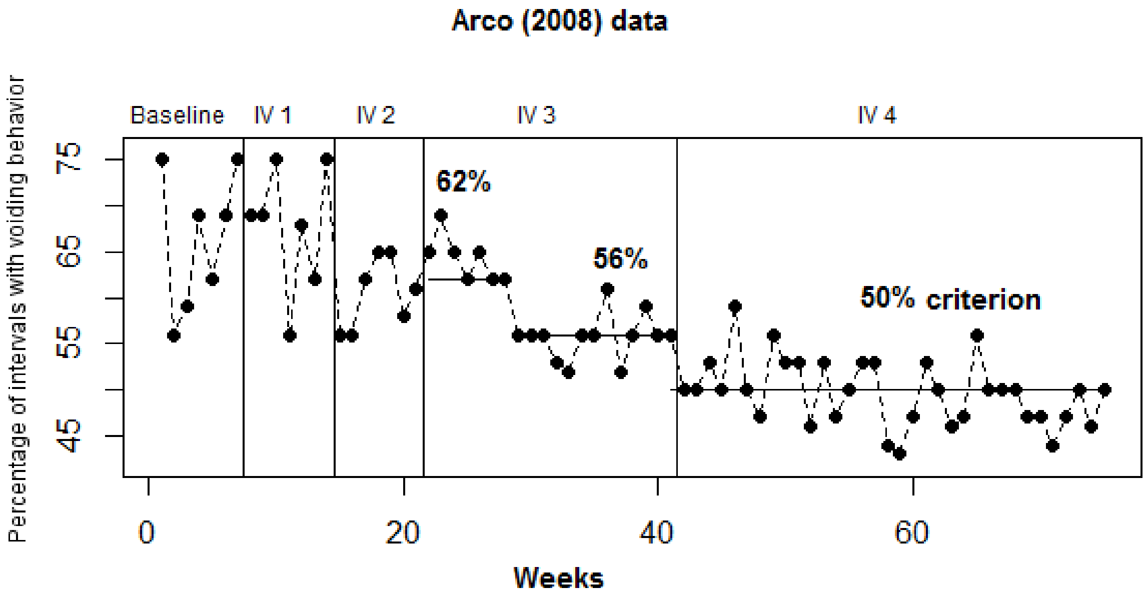


Figure 5. Data gathered by Arco (2008) on the effect of neurobehavioral intervention with a participant who had sustained frontal-temporal lobe brain trauma. The three criteria were negotiated with the participant

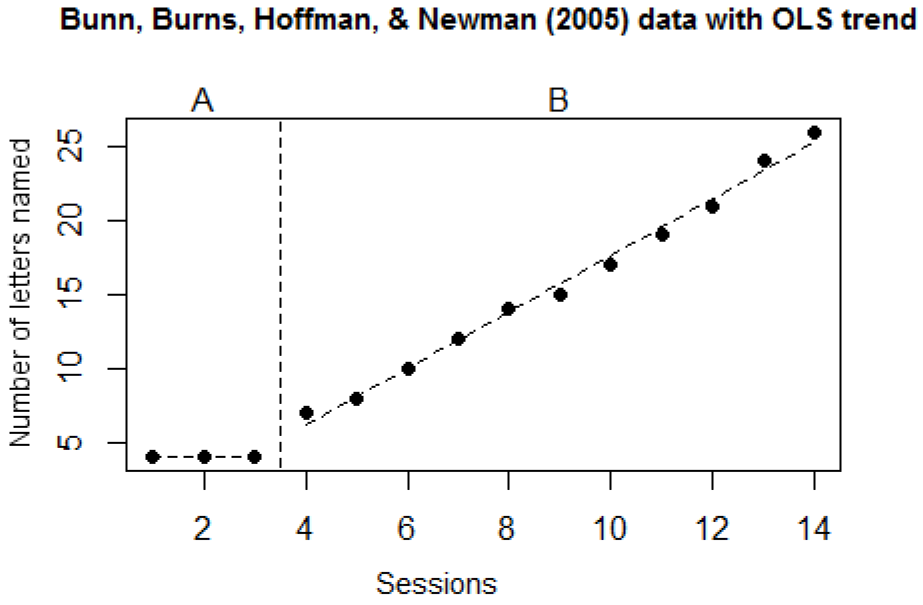


Figure 6. Data gathered by Bunn et al. (2005) on the effect of incremental rehearsal for teaching a 4-year old girl letter identification. OLS: ordinary least squares.

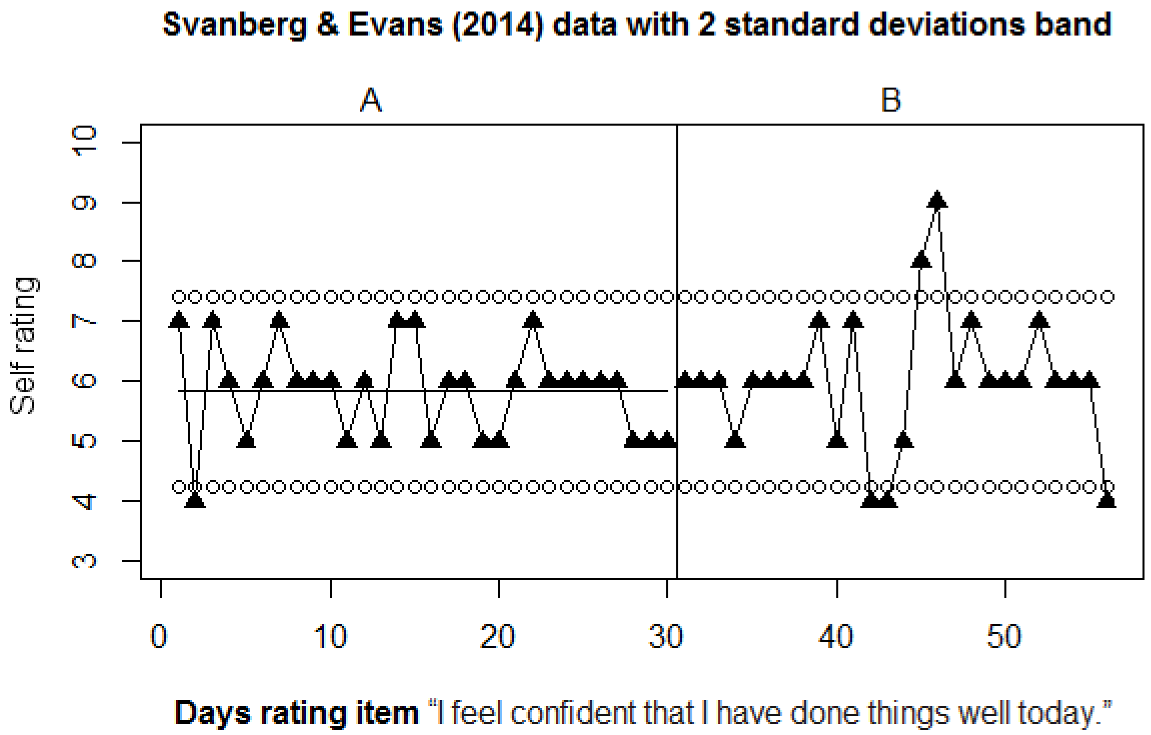


Figure 7. Data gathered by Svanberg and Evans (2014) on the effect of SenseCam on the mood of a person with Korsakoff syndrome.

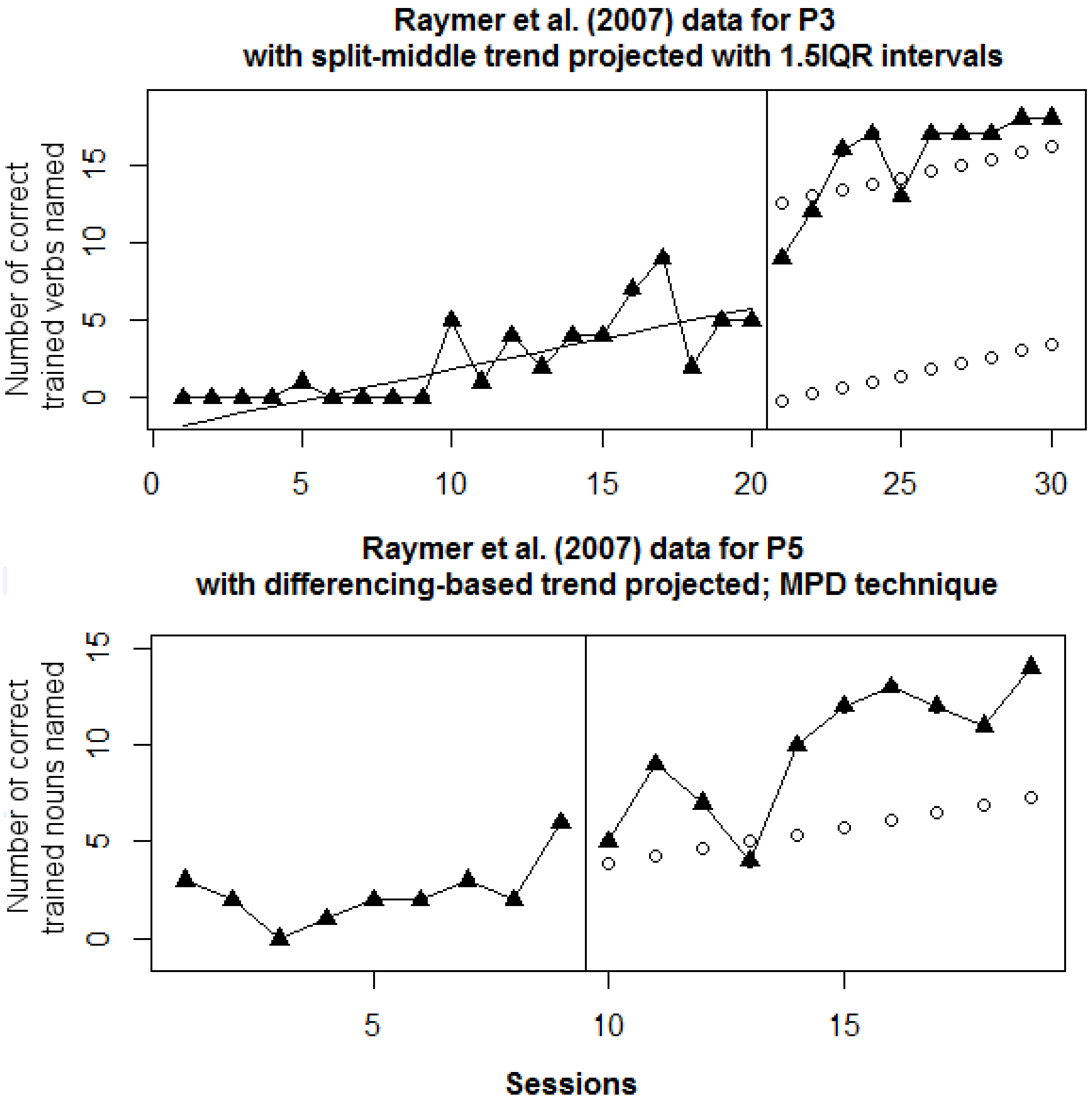


Figure 8. Data gathered by Raymer et al. (2007) on the effect of semantic-phonologic treatment for noun and verb retrieval impairments in participants (P3 and P5) with naming disorders induced by left hemisphere strokes. IQR: interquartile range. MPD: Mean phase difference.

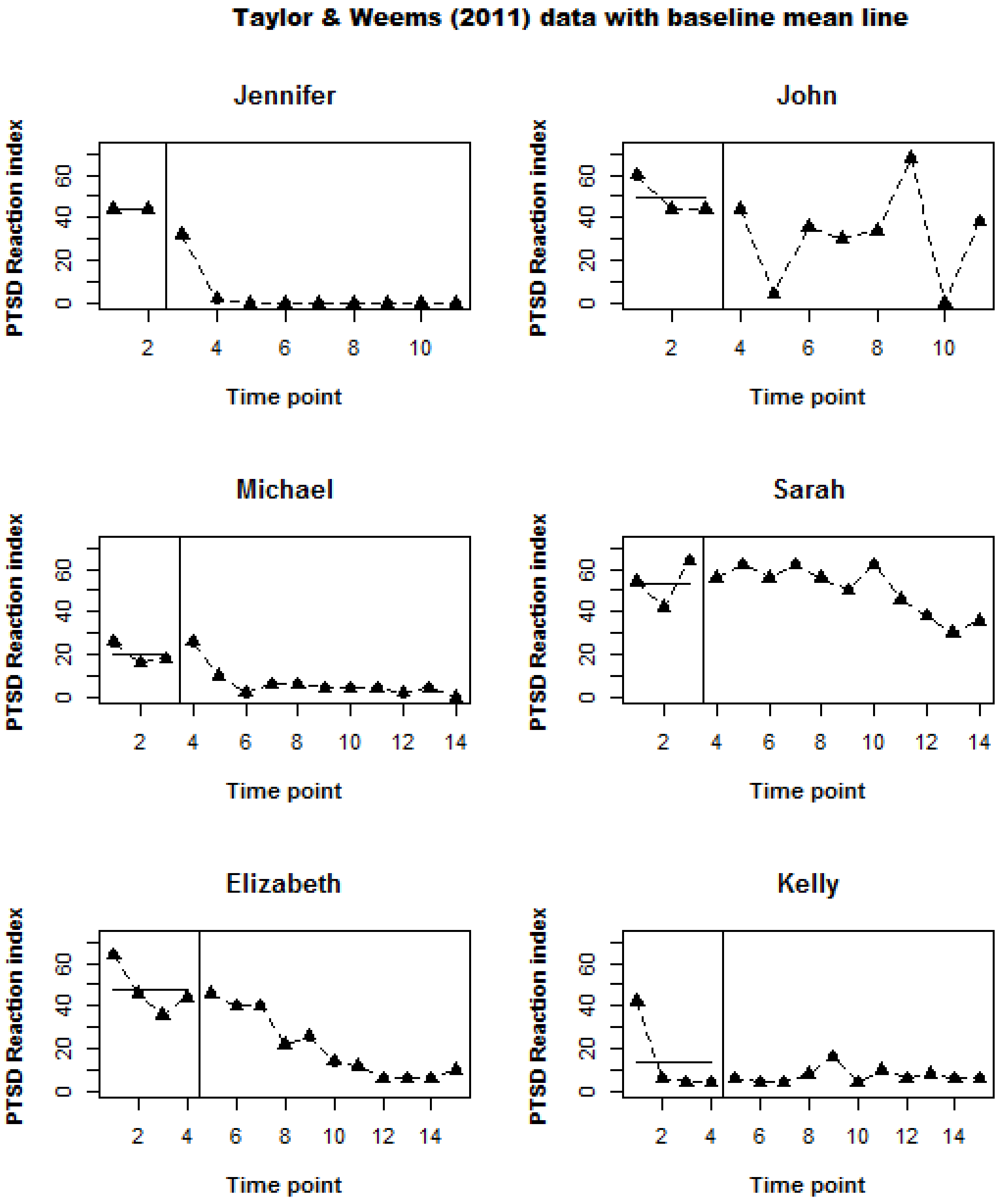


Figure 9. Data gathered by Taylor and Weems (2011) on the effect of cognitive-behavior therapy for disaster-exposed youth. PTSD: posttraumatic stress disorder, measured by the self-report questionnaire by Frederick, Pynoos, & Nadar (1992).

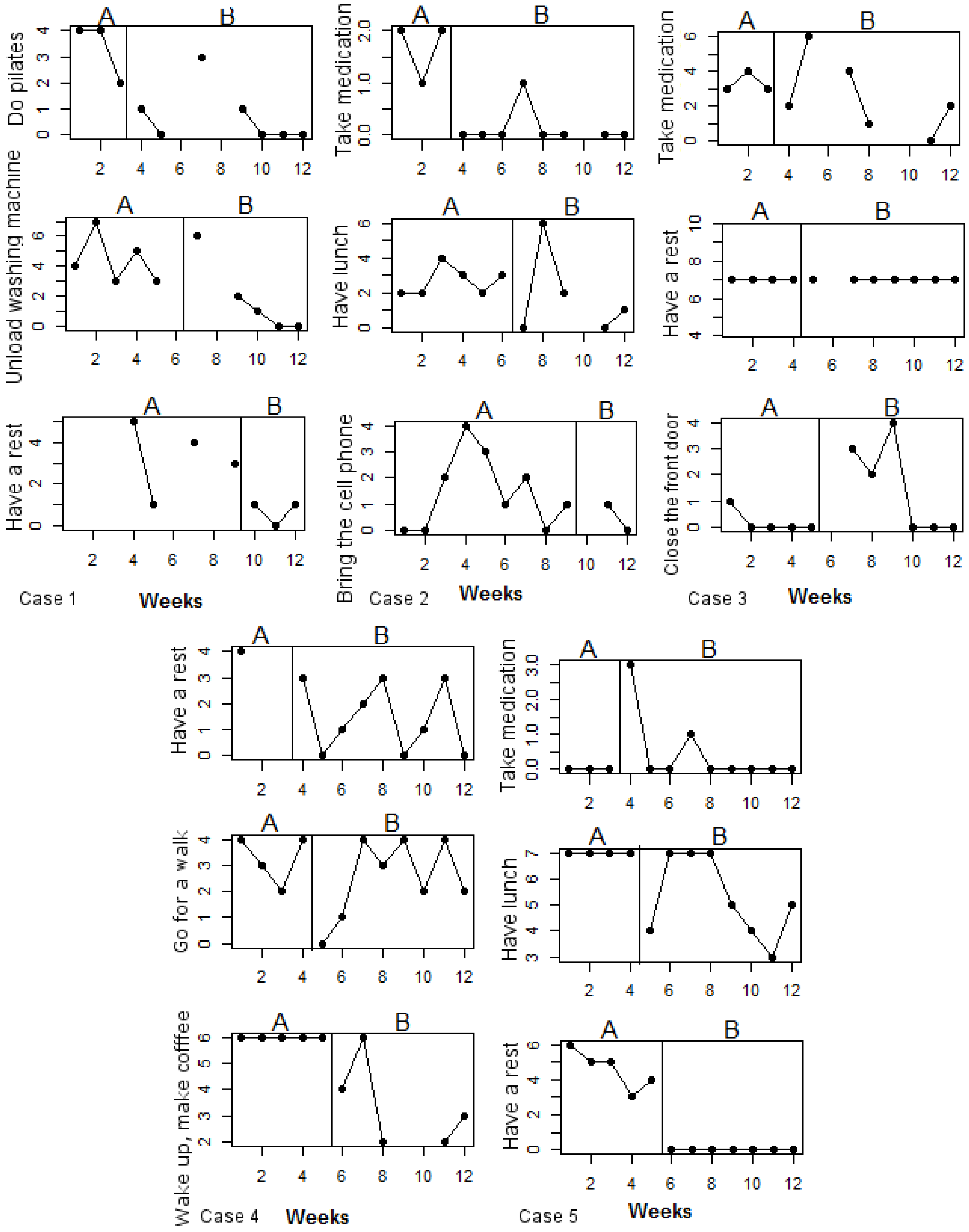


Figure 10. Data gathered by Boman et al. (2010): activities missed without and with reminders.

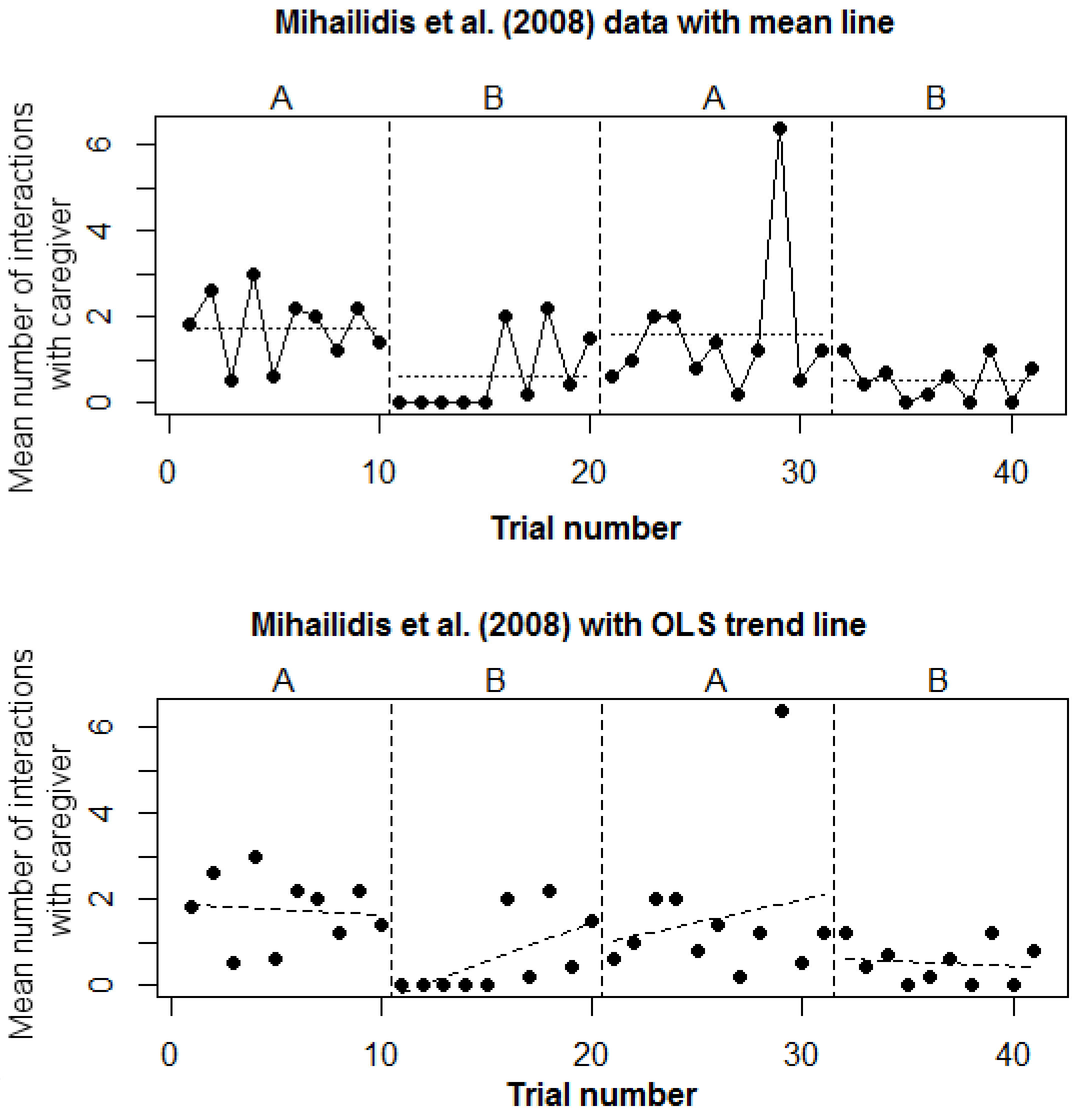


Figure 11. Data gathered by Mihailidis et al. (2008) on the effect of COACH prompting system (B phases) to assist older adults with dementia through handwashing. The points represent averages for six adults. OLS: ordinary least squares.

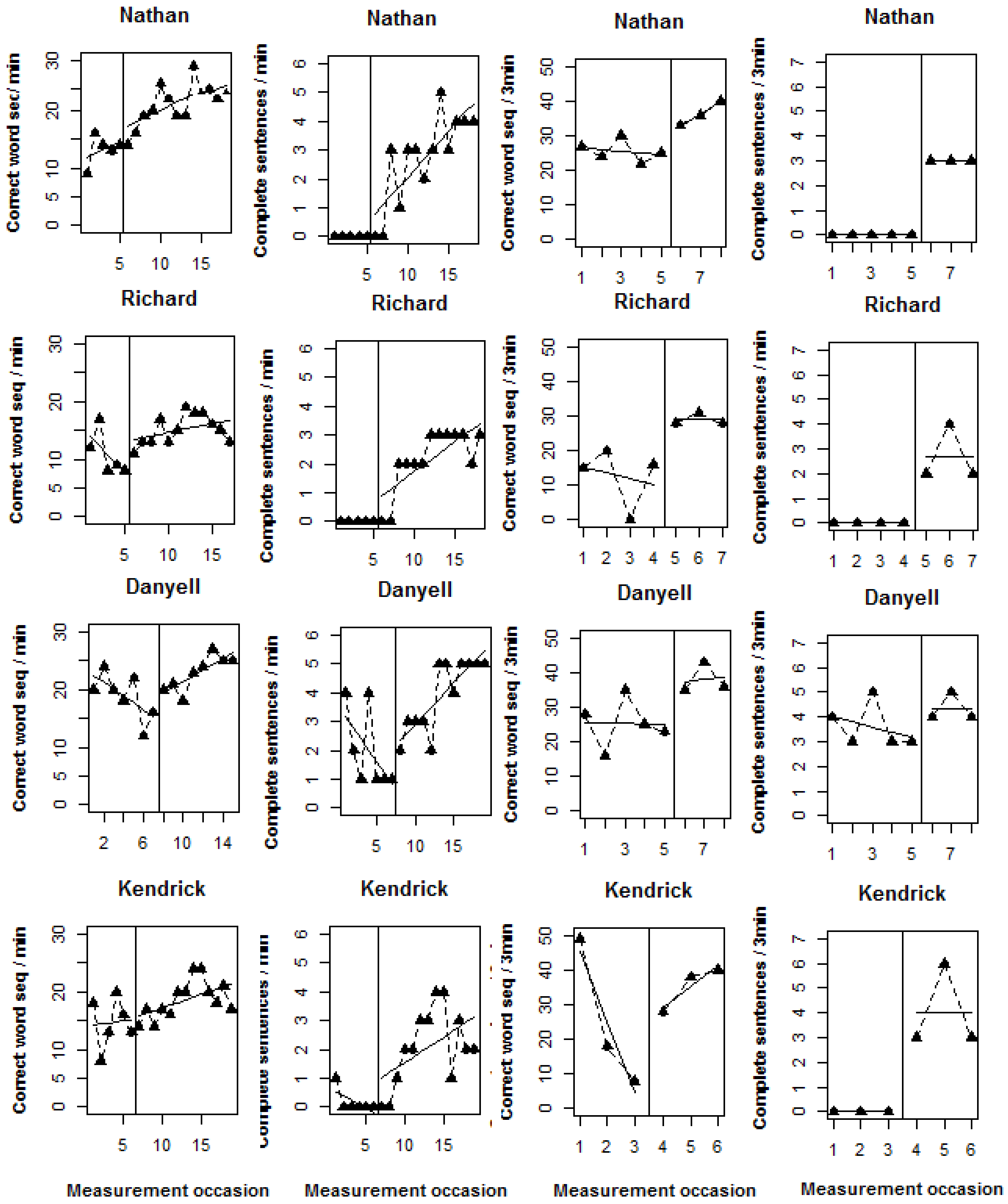


Figure 12. Data gathered by Datchuk (2015) on the effect of a multicomponent intervention on four outcomes for four adolescents with writing difficulties. The continuous lines represent within-phase ordinary least squares regression of the score on the measurement occasion.

According to the Researchers' Resources

The third pillar is expressed in terms of how researchers can be expected to analyze the data according to the knowledge and access to software they have. However, we consider that it is more appropriate to collaborate with researchers who do have such resources and base the choice of an analytical technique on the research aim or the data characteristics instead. Therefore, this pillar (the recommendations for which are included in Online Table 9) should be considered as the last option and an emergency-only solution to be avoided, if possible.

Table 9. Recommendations for choosing analytical techniques according to researchers' resources.

Situation	Analysis	Justification
If you have the knowledge about how to model different aspects of the data + the data include replication across participants	Multilevel models	Model autocorrelation Estimate variance of the intervention effect between participants and/or studies Separate estimates of the intervention effect per participant and/or per study, apart from average effects
If you have the knowledge how to model different aspects of the data + the data include replication across participants + you want to obtain an overall quantification	PHS *d*-statistic	Make use of multilevel modeling of the relevant data aspects
If you or your collaborators have the knowledge to understand statistical formulations + the data include replication across participants + you want to obtain an overall quantification	HPS *d*-statistic	Meta-analyze group-design and SCED studies together using inverse index variance as a weight May require detrending
If you want to be autonomous + not need to spend too much time learning + you are used to performing visual analysis	Nonoverlap of all pairs	Possibility to obtain *p* values If the data show no trend
If you want to be autonomous + not need to spend too much time learning + you are used to performing visual analysis + the data show trend or you want to quantify intervention phase trend	Tau	Possibility to obtain *p* values
If you want to be autonomous + not need to spend too much time learning + you are used to performing visual analysis + the data show no trend + you want to quantify the difference between conditions even if there is complete nonoverlap	Mean baseline reduction Percentage change index	Comparability across studies, given that the index is expressed as a percentage If there is substantive reason to focus on the last three measurements per

		phase
If you want to be autonomous + not need to spend too much time learning + you are used to performing visual analysis + you want a separate quantifications of change in level and in slope, expressed in the same metric as the dependent variable	Slope and level change	Controls for baseline trend Possibility to obtain a standardized measure
If you want to be autonomous + not need to spend too much time learning + you are used to performing visual analysis + you want an overall quantification between predicted and actual intervention phase measurements, expressed in the same metric as the dependent variable	Mean phase difference	Possibility to control for baseline trend Possibility to obtain a standardized measure
If you are used to performing regression analysis + you are acquainted with regression analysis + you want a quantification in the same metric as the dependent variable	Piecewise regression	Controls for baseline trend Possibility to obtain a standardized measure
If you are used to performing regression analysis + are comfortable with interpreting transformed data + you want a quantification in the same metric as the dependent variable + you want to control for autocorrelation	Generalized least squares regression	Controls for baseline trend Possibility to obtain a standardized measure

An illustration

In the current section we will illustrate how an applied researcher can use the provided recommendations. This illustration is based on a dataset identified via a PsycINFO search performed on September 18, 2015 using the keyword “single-case design” and looking for the most recent publication. The article identified was a study by Datchuk (2015) concerning multicomponent intervention on the writing behavior of adolescents with writing difficulties published online on September 10, 2015. The data are gathered according to a multiple-baseline design across 4 participants measured on 4 outcome variables each. Datchuk (2015) bases his analysis on comparing phases in terms of their mean levels. Thus, we can assume that his aims are to quantify the differences in average levels and also to take progressive improvements into account, as well as an overall quantification that can support his discussion about whether the intervention is effective across the participants.

First, according to Table 7 from the Online Repository, visual analysis is necessary for an initial assessment of whether the data pattern corresponds to what is expected from an effective intervention (i.e., a change in the target behavior only when the intervention is introduced) and of the baseline (Datchuk, 2015, correctly mentions that baselines are stable or worsening; see Figure 12 from the Online Repository). Such analysis also suggests that there are improving trends in some of the intervention phases (also mentioned by Datchuk, 2015), that there are no outliers and that data variability is not excessive.

Second, still according to Table 7, an overall quantification for the four across-subject replications for each outcome can be obtained using the HPS *d*-statistic for multiple-baseline designs, given that there are no baseline trends to control for. Note that software for the HPS *d*-statistic (*scdhlm* package for R: https://github.com/jepusto/scdhlm) also offers the overall quantification in raw metric: in this case, words or sentences per 1 or 3 minutes, according to the outcome.

Third, in order to obtain a quantification for each pair of AB phases that is also expressed in a raw metric and that also allows quantifying the changes in slope identified visually, piecewise regression and SLC procedure can be used. Piecewise regression quantifies also the immediate change in level, whereas SLC quantifies average difference in level once any linear trends have been removed. In order to choose between the two, it should be noted that the visual inspection of the data suggests that an immediate change is not evident for all baselines and outcomes. Moreover, Datchuk (2015) compared mean levels in his analysis, which potentially indicates that his aim was not to focus on an immediate change. Thus, we would choose SLC and we can implement it using the R code available in the appendix of Solana et al. (2010). However, one should be cautious when interpreting the estimates given that the measurement occasions are not equally spaced in time (as shown on the figures presented in Datchuk, 2015). Therefore, the slope change estimate would quantify the amount of increase in the target behavior per measurement occasion, but not per natural days. This distinction is not relevant for the HPS *d*-statistics, which only compares average levels.

Fourth, statistical significance and meta-analytical integration do not seem to be among Datchuk’s (2015) aims. Therefore, the corresponding Table 7 recommendations are not relevant.

Fifth, if we look at the suggestions made in Table 8 on the basis of data characteristics, we see that a relatively clear data pattern can be made even clearer using visual aids. Given that intervention phase trend observed in some cases, it seems meaningful to represent trend lines on the graphs, keeping in mind that relatively stable baselines would lead to flat trend lines. Ordinary least squares regression can be used to fit the trend lines (as shown on Figure 12 from the Online Repository), given that the dependent variable is a rate and there are no outliers.

Sixth, following the subsequent recommendations from Table 8, the recommendations about data with considerable variability are not relevant. The only relevant analysis left is the Nonoverlap of all pairs, as the data show no baseline trend and it is not directly visually clear how much overlap is there in the data. Nevertheless, this index may not be as informative in this case (and was not among Datchuk's aims) as it does not offer a quantification in terms of the dependent variable (number of words or sentences per 1 minute and per 3 minutes).

Discussion

The multiple criteria and recommendations presented here highlight the fact that it is unlikely that a single analytical technique would become the only standard for SCED data. If we were to compare analytical techniques on the amount of methodological and statistical work done on them, multilevel models are likely the focus of most recent discussions and tests, whereas randomization tests have generated many publications in the past. On the basis of their enormous flexibility in terms of what data features and effects is modelled and estimated, multilevel models could be the standard. However, such an apparently logical choice can only be made if it is assumed that all applied researchers are able (or can learn) to specify the desired and appropriate model correctly, that they also know how to interpret correctly the results, while also assuming that the parametric assumptions are met and the amount of data (especially, number of cases) available is sufficient for obtaining unbiased and precise estimates. Thus, multilevel models do not seem to be the universal solution.

Complementarily, if we were to compare analytical techniques on the amount of applications, visual analysis (Parker & Brossart, 2003) and nonoverlap indices (especially the Percentage of nonoverlapping data) are expected to excel (Scruggs & Mastropieri, 2013; Vissenaeken, 2015). On the basis of the ease of computation and interpretation, nonoverlap indices (especially, NAP and Tau-U) could be the standard analytical option. However, such indices do not quantify (interval or ratio scale) differences beyond the complete (ordinal) nonoverlap (Solomon et al., 2015) and they assume lack of autocorrelation especially for obtaining statistical significance on the basis of statistical models that were not developed for dependent data. Moreover, a single statistically sound method for performing meta-analysis of has not been established for nonoverlap indices (e.g., Schlosser et al., 2008). This is another illustration of the difficulties involved in identifying/developing *the* SCED analytical technique. It also illustrates the reason why we here decided to point in which conditions which techniques are likely to be most useful.

Apart from proposing and testing techniques individually, a global perspective like the one offered here is necessary, putting the strengths and limitations of a procedure in the context of the remaining techniques available. Nevertheless, the global perspective offered here is only an initial step, not a final definitive judgment on the usefulness of SCED analytical techniques. A public discussion with both methodologists and applied researchers is necessary and the input of the applied researchers is crucial if the analytical techniques are to be used by them.

Part of the discussion on SCED data analysis techniques could focus on the distinction between an approach for choosing an analytical technique based on the specific aim and/or data features of a given study and a more general approach towards data analysis in the SCED context. A general approach should necessarily take into account additional information, apart from the one displayed in a graph and analyzed quantitatively. Specifically, the convergence of professional criteria, normative data (if available), clients' and relatives' perceptions on the change that has taken place is important when assessing intervention effectiveness (Kazdin, 1999). In this general analytical approach, the assessment of the intervention has to pay special attention to the operative definition of the dependent variable, procedural and treatment fidelity

(Ledford & Gast, 2014), the demonstration of experimental control, and the maintenance and generalization of the effect (see Maggin et al., 2014).

Focusing specifically on quantification, deciding how to analyze the data on the basis of the expected intervention effect (as done when choosing a test statistic in randomization tests; Heyvaert & Onghena, 2014a) would provide a conservative solution ensuring that only a theoretically-grounded effect would be quantified and treating the remaining data features as nuisance. In contrast, choosing the analytical technique on the basis of the data at hand would ensure that a quantification represents the main features of data, according to the researcher, and this is likely to make easier the detection of any kind of difference between the conditions compared, regardless of whether such a difference was expected or not.

Finally, another question is whether it is preferable to use only one or several data analytical techniques. Given that several different data features may be of interest (e.g., level, slope, overlap, variability) it seems justified to use several techniques to account for all these data features, just as visual analysts focus on several data features at a time (Parker et al., 2006). Analogously, in between-groups designs, it is also possible to apply several statistical indices (e.g., mean, standard deviation) and tests (e.g., a *t*-test, a test for the statistical significance of the correlation coefficient) to the same data. A potential problem of using several techniques would be that they might make more difficult the decision on whether the intervention was effective or not in practical terms (e.g., when there is a complete nonoverlap but a small change in level).

Limitations

Regarding the way in which the criteria and recommendations were put forward, we acknowledge that there may be other solid ways of establishing criteria for SCED data analysis and scoring the different analytical techniques. One option would have been to ask the opinion of methodologists and statisticians as participants (as in a Delphi study) regarding the set of criteria for evaluating the analytical techniques or regarding the recommendations they would make. Such an approach would have the main strength that methodologists and statisticians who develop and/or test the techniques understand them better and they can provide detailed recommendations, according to study aims, number of replications, characteristics of the data, assumptions made, etc. However, the following limitations made us discard this option. (a) It is necessary to develop an initial list of criteria and/or a set of recommendation to use as a basis for further discussion and this initial set may bias the discussion and have to great influence on the final set (Kristensen & Gärling, 1997). (b) It is possible that each researcher would defend the technique they have proposed and/or have been working on. (c) The decision regarding who is a methodologist or a statistician and who is an expert is problematic – specifically, should the number of publications or their influence (citations) be considered and what is the cut-off point. (d) It is questionable whether the views of methodologists and statisticians are shared by practitioners and applied researchers and whose criteria should prevail.

Another option would have been the same Delphi study, as described above but with applied researchers. Such an approach would have the following strengths: (a) the potential users of the techniques would be given the possibility to decide; (b) it would be possible to compare the proposals available with the techniques actually used; (c) evidence on the attractiveness of the proposals could be obtained, once applied researchers become acquainted with them; (d) such a study would present analytical techniques directly to their potential users and update their knowledge on the topic. Among the limitations of this approach we could state the following: (a) The issue with the initial list of criteria mention above is also valid here. (b) It is possible that each applied researcher would prefer to stick to the technique s/he is familiar with. (c) It is necessary to select the information on the proposals (e.g., from journal articles and book chapters) to be presented to the applied researchers, or to elaborate new information in case the existing one is not considered suitable (e.g., due to being excessively technical). (d) There is an even greater difficulty in deciding which practitioners and applied researchers would participate as experts in the study and which the criterion for expertise is. (e) The previously mentioned question about whether statisticians' or applied researchers' criteria should prevail remains.

Regarding the analytical techniques reviewed, we consider that the main and most promising options were included, but we do not claim that the list is comprehensive. Moreover, looking at the amount of recommendations made, it could appear that the degree of synthesis is not sufficient, as it has not led to simple or few recommendations. Nevertheless, we consider that the list still open, as there are probably other combinations of criteria, not made explicit here, that would lead to even more recommendations regarding the choice of an analytical technique.

Future Research

Given the way in which the current recommendations were developed, we consider that a useful follow-up study would be to provide datasets to applied researchers and methodologists, handle them the recommendations, and explore whether (a) these guidelines help them identifying the most appropriate technique depending on the assumptions they are willing to make, the purpose, the data characteristics, etc., (b) there is an agreement among applied researchers regarding the technique chosen, and (c) whether the analytic techniques are applied appropriately and the interpretation of the obtained results is correct.

References

Allison, D. B., Franklin, R. D., & Heshka, S. (1992). Reflections on visual inspection, response guided experimentation, and Type I error rate in single-case designs. *The Journal of Experimental Education, 6*, 45-51.

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research and Therapy, 31*, 621-631.

Allison, D. B., & Gorman, B. S. (1994). “Make things as simple as possible, but no simpler”. A rejoinder to Scruggs and Mastropieri. *Behaviour Research and Therapy, 32*, 885-890.

Arco, L. (2008). Neurobehavioural treatment for obsessive-compulsive disorder in an adult with traumatic brain injury. *Neuropsychological Rehabilitation, 18*, 109-124

Beeson, P. M., & Robey, R. R. (2006). Evaluating single-subject treatment research: Lessons learned from the aphasia literature. *Neuropsychological Review, 16*, 161-169.

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

Boman, I.-L., Bartfai, A., Borell,L., Tham, K., & Hemmingsson, H. (2010).Support in everyday activities with a home-based electronic memory aid for persons with memory impairments. *Disability and Rehabilitation: Assistive Technology, 5*, 339–350.

Borckardt, J., & Nash, M. (2014). Simulation modelling analysis for small sets of single-subject data collected over time. *Neuropsychological Rehabilitation, 24*, 492-506.

Borckardt, J. J., Nash, M. R., Murphy, M. D., Moore, M., Shaw, D., & O'Neil, P. (2008). Clinical practice as natural laboratory for psychotherapy research: A guide to case-based time-series analysis. *American Psychologist, 63*, 77-95.

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage Foundation.

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.

Bulté, I., & Onghena, P. (2013). The single-case data analysis package: Analysing single-case experiments with R software. *Journal of Modern Applied Statistical Methods, 12*, 450-478.

Bunn, R., Burns, M. K., Hoffman, H. H., & Newman, C. L. (2005). Using incremental rehearsal to teach letter identification with a preschool-age child. *Journal of Evidence-Based Practices for Schools, 6*, 124-134.

Burns, M. K., (2012). Meta-analysis of single-case design research: Introduction to the special issue. *Journal of Behavioral Education, 21*, 175-184.

Burns, M. K., Zaslofsky, A. F., Kanive, R., & Parker, D. C. (2012). Meta-analysis of incremental rehearsal using phi coefficients to compare single-case and group designs. *Journal of Behavioral Education, 21*, 185-202.

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research designs and analysis: New directions for psychology and education* (pp. 187-212). Hillsdale, NJ: Lawrence Erlbaum.

Carey, L .M., & Matyas, T. A. (2005): Training of somatosensory discrimination after stroke: Facilitation of stimulus generalization. *American Journal of Physical Medicine and Rehabilitation, 84*, 428-442.

Center, B. A., Skiba, R. J., & Casey, A. (1985-1986). A methodology for the quantitative synthesis of intra-subject design research. *The Journal of Special Education, 19,* 387-400.

Chang, Y. J., Chou, L. D., Wang, F. T. Y., & Chen, S. F. (2011). A kinect-based vocational task prompting system for individuals with cognitive impairments. *Personal and Ubiquitous Computing, 17,* 1-8.

Chen, L.-T., Peng, C.-Y. J., & Chen, M.-E. (2015). Computing tools for implementing standards for single-case designs. *Behavior Modification, 39*, 835-69.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Coker, P., Lebkicher, C., Harris, L., & Snape, J. (2009). The effects of constraint-induced movement therapy for a child less than one year of age. *NeuroRehabilitation, 24,* 199–208.

Crosbie, J. (1987). The inability of the binomial test to control Type I error with single-subject data. *Behavioral Assessment, 9*, 141-150.

Danov, S. E., & Symons, F. J. (2008). A survey evaluation of the reliability of visual inspection and functional analysis graphs. *Behavior Modification, 32*, 828-839.

Datchuk, S. M. (2015, September 10). Writing simple sentences and descriptive paragraphs: Effects of an intervention on adolescents with writing difficulties. *Journal of Behavioral Education*. Advance online publication. doi: 10.1007/s10864-015-9236-x

Davis, D. H., Gagné, P., Fredrick, L. D., Alberto, P. A., Waugh, R. E., & Haardörfer, R. (2013). Augmenting visual analysis in single-case research with hierarchical linear modeling. *Behavior Modification, 37*, 62-89.

Dugard, P. (2014). Randomization tests: A new gold standard? *Journal of Contextual Behavioral Science, 3,* 65–68.

Edgington, E. S. (1972a). An additive method for combining probability values from independent experiments. *Journal of Psychology, 80*, 351−363.

Evans, J. J., Gast, D. L., Perdices, M., & Manolov, R. (2014). Single case experimental designs: Introduction to a special issue of Neuropsychological Rehabilitation. *Neuropsychological Rehabilitation, 24*, 305-314.

Ferron, J. M., Farmer, J. L., & Owens, C. M. (2010). Estimating individual treatment effects from multiple-baseline data: A Monte Carlo study for multilevel-modeling approaches. *Behavior Research Methods, 42*, 930-943.

Ferron, J. M., & Jones, P. K. (2006). Tests for the visual analysis of response-guided multiple-baseline data. *The Journal of Experimental Education, 75*, 66-81.

Ferron, J. M., & Levin, J. R. (2014). Single-case permutation and randomization statistical tests: Present status, promising new developments. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research. Methodological and statistical advances* (pp. 153-183). Washington, DC: American Psychological Association.

Ferron, J. M., & Onghena, P. (1996). The power of randomization tests for single-case phase designs. *The Journal of Experimental Education, 64,* 231-239.

Ferron, J. M., & Sentovich, C. (2002). Statistical power of randomization tests used with multiple-baseline designs. *The Journal of Experimental Education*, *70*, 165-178.

Fisher, W. W., Kelley, M. E., & Lomas, J. E. (2003). Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. *Journal of Applied Behavior Analysis, 36*, 387-406.

Franklin, R. D., Gorman, D. B., Beasley, T. M. & Allison, D. B. (1997). Graphical display and visual analysis. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 119-158). Mahwah, NJ: Lawrence Erlbaum Associates.

Frederick, C. J., Pynoos, R. S., & Nadar, K. (1992). *Reaction index to psychic trauma form C (child)*. Unpublished manuscript, University of California, Los Angeles.

Gage, N. A., & Lewis, T. J. (2013). Analysis of effect for single-case design research. *Journal of Applied Sport Psychology, 25,* 46-60.

Gast, D. L., & Spriggs, A. D. (2010). Visual analysis of graphic data. In D. L. Gast (Ed.), *Single subject research methodology in behavioral sciences* (pp. 199-233). London, UK: Routledge.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Gorsuch, R. L. (1983). Three methods for analyzing limited time-series (N of 1) data. *Behavioral Assessment, 5*, 141-154.

Graves, M. M., Roberts, M. C., Rapoff, M., & Boyer, A. (2010). The efficacy of adherence interventions for chronically ill children: A meta-analytic review. *Journal of Pediatric Psychology, 35*, 368-382.

Hagopian, L. P., Fisher, W. W., Thompson, R. H., Owen-DeSchryver, J., Iwata, B. A., & Wacker, D. P. (1997). Toward the development of structured criteria for interpretation of functional analysis data. *Journal of Applied Behavior Analysis, 30,* 313-326.

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

Harrop, J. W., & Velicer, W. F. (1985). A comparison of alternative approaches to the analysis of interrupted time-series. *Multivariate Behavioral Research, 20*, 27-44.

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

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2013). A standardized mean difference effect size for multiple baseline designs across individuals. *Research Synthesis Methods, 4*, 324-341.

Heinicke, M. R., & Carr, J. E. (2014). Applied behavior analysis in acquired brain injury rehabilitation: A meta-analysis of single-case design intervention research. *Behavioral Interventions, 29*, 77–105.

Hershberger, S. L., Wallace, D. D., Green, S. B., & Marquis, J. G. (1999). Meta-analysis of single-case data. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 107-132). London, UK: Sage.

Heyvaert, M., & Onghena, P. (2014a). Analysis of single-case data: Randomisation tests for measures of effect size. *Neuropsychological Rehabilitation*, *24*, 507-527.

Heyvaert, M., & Onghena, P. (2014b). Randomization tests for single-case experiments: State of the art, state of the science, and state of the application. *Journal of Contextual Behavioral Science, 3,* 51–64.

Heyvaert, M., Saenen, L., Maes, B., & Onghena, P. (2015). Comparing the percentage of non-overlapping data approach and the hierarchical linear modeling approach for synthesizing single-case studies in autism research. *Research in Autism Spectrum Disorders, 11*, 112-125.

Huitema, B. E. (2004). Analysis of interrupted time-series experiments using ITSE: A critique. *Understanding Statistics*, *3*, 27-46.

Huitema, B. E., McKean, J. W., & Laraway, S. (2007). Time series intervention analysis using ITSACORR: Fatal flaws. *Journal of Modern Applied Statistical Methods*, *6*, 367-379.

Jamieson, M., Cullen, B., McGee-Lennon, M., Brewster, S., & Evans, J. J. (2014). The efficacy of cognitive prosthetic technology for people with memory impairments: A systematic review and meta-analysis. *Neuropsychological Rehabilitation, 24*, 419-444.

Kahng, S. W., Chung, K.-M., Gutshall, K., Pitts, S. C., Kao, J., & Girolami, K. (2010). Consistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis*, *43*, 35-45.

Kazdin, A. E. (1978). Methodological and interpretive problems of single-case experimental designs. *Journal of Consulting and Clinical Psychology, 46*, 629-642.

Kazdin, A. E. (1999). The meanings and measurements of clinical significance. *Journal of Consulting and Clinical Psychology, 67*, 332-339.

Kirsch, N. L., Shenton, M., & Rowan, J. (2004). A generic, "in-house", alphanumeric paging system for prospective activity impairments after traumatic brain injury. *Brain Injury, 18*, 725-734.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). Single case designs technical documentation. In *What Works Clearinghouse: Procedures and standards handbook (Version 1.0).* Available at http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_scd.pdf

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

Kratochwill, T. R., Levin, J. R., Horner, R. H., & Swoboda, C. M. (2014). Visual analysis of single-case intervention research: Conceptual and methodological issues. In T. R. Kratochwill & J. R. Levin (Eds.) *Single-case intervention research: Methodological and statistical advances* (pp. 91-125). Washington D.C.: American Psychological Association.

Kristensen, H., & Gärling, T. (1997). The effects of anchor points and reference points on negotiation process and outcome. *Organizational Behavior and Human Decision Processes, 71*, 85-94.

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

Ledford, J. R, & Gast, D. L. (2014). Measuring procedural fidelity in behavioural research. *Neuropsychological Rehabilitation, 24,* 332-348.

Levin, J. R., Ferron, J. M., & Kratochwill, T. R. (2012). Nonparametric statistical tests for single-case systematic and randomized ABAB…AB and alternating treatment intervention designs: New developments, new directions. *Journal of School Psychology, 50,* 599-624.

Levin, J. R., Lall, V. F., & Kratochwill, T. R. (2011). Extensions of a versatile randomization test for assessing single-case intervention effects. *Journal of School Psychology, 49*, 55-79.

Logan, K. R., Jacobs, H. A., Gast, D. L., Murray, A. S., Diano, K., & Skala, C. (1998). The impact of typical peers on the perceived happiness of students with profound multiple disabilities. *Research and Practice for Persons with Severe Disabilities, 23*, 309-318.

Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject research: Percentage of data points exceeding the median. *Behavior Modification, 30*, 598-617.

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.

Maggin, D. M., Briesch, A. M., Chafouleas, S. M., Ferguson, T. D., & Clark, C. (2014). A comparison of rubrics for identifying empirically supported practices with single-case research. *Journal of Behavioral Education, 23*, 287–311.

Maggin, D. M., & Chafouleas, S. M. (2013). Introduction to the Special Series: Issues and advance of synthesizing single-case research. *Remedial and Special Education, 34*, 3-8.

Manolov, R., Arnau, J., Solanas, A., & Bono, R. (2010). Regression-based techniques for statistical decision making in single-case designs. *Psicothema, 22*, 1026-1032.

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

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

Manolov, R., Sierra, V., Solanas, A., & Botella, J. (2014). Assessing functional relations in single-case designs: Quantitative proposals in the context of the evidence-based movement. *Behavior Modification, 38*, 878-913.

Manolov, R., & Solanas, A. (2009). Percentage of nonoverlapping corrected data. *Behavior Research Methods, 41*, 1262-1271.

Manolov, R., & Solanas, A. (2013). A comparison of mean phase difference and generalized least squares for analyzing single-case data. *Journal of School Psychology, 51*, 201-215.

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis for single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23*, 341-351.

Mihailidis, A., Boger, J. N., Craig, T., & Hoey, J. (2008). The COACH prompting system to assist older adults with dementia through handwashing: An efficacy study. *BMC Geriatrics, 8*, 28.

Miller, M. J. (1985). Analyzing client change graphically. *Journal of Counseling and Development, 63*, 491-494.

Moeyaert, M., Ferron, J., Beretvas, S., & Van Den Noortgate, W. (2014). From a single-level analysis to a multilevel analysis of since-case experimental designs. *Journal of School Psychology, 52*, 191-211.

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S., & Van Den Noortgate, W. (2013). The three-level synthesis of standardized single-subject experimental data: A Monte Carlo simulation study. *Multivariate Behavioral Research, 48,* 719-748.

Moss, A., & Nicholas, M. (2006). Language rehabilitation in chronic aphasia and time postonset: A review of single-subject data. *Stroke, 37*, 3043-3051.

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.

Olive, M. L., & Smith, B. W. (2005). Effect size calculations and single subject designs. *Educational Psychology, 25,* 313-324.

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

Parker, R. I., & Brossart, D. F. (2003). Evaluating single-case research data: A comparison of seven statistical methods. *Behavior Therapy, 34,* 189-211.

Parker, R. I., Cryer, J., & Byrns, G. (2006). Controlling baseline trend in single-case research. *School Psychology Quarterly, 21*, 418-443.

Parker, R. I., & Hagan-Burke, S. (2007). Median-based overlap analysis for single case data: A second study. *Behavior Modification, 31*, 919-936.

Parker, R. I., Hagan-Burke, S., & Vannest, K. J. (2007). Percentage of all non-overlapping data: An alternative to PND. *Journal of Special Education, 40*, 194-204.

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

Parker, R. I., Vannest, K. J., & Brown, L. (2009). The improvement rate difference for single-case research. *Exceptional Children, 75*, 135-150.

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. (2014). A simple method to control positive baseline trend within data nonoverlap. *Journal of Special Education, 48*, 79-91.

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.

Parsonson, B. S., & Baer, D. M. (1986). The graphic analysis of data. In A. Poling & R. W. Fuqua (Eds.), *Research methods in applied behavior analysis: Issues and advances* (pp. 157-186). New York, NY: Plenum Press.

Parsonson, B. S., & Baer, D. M. (1992). The visual analysis of data, and current research into the stimuli controlling it. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research designs* *and analysis: New directions for psychology and education* (pp. 15-40). Hillsdale, NJ: Lawrence Erlbaum.

Perdices, M., & Tate, R. L. (2009). Single-subject designs as a tool for evidence-based clinical practice: Are they unrecognised and undervalued? *Neuropsychological Rehabilitation, 19*, 904-927.

Pfadt, A., & Wheeler, D. J. (1995). Using statistical process control to make data-based clinical decisions. *Journal of Applied Behavior Analysis, 28*, 349-370.

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.

Raymer, A. M., Ciampitti, M., Holliway, B., Singletary, F., Blonder, L. X., … , Gonzalez Rothi, L. J. (2007). Semantic-phonologic treatment for noun and verb retrieval impairments in aphasia, *Neuropsychological Rehabilitation, 17*, 244-270.

Robey, R. R., Schultz, M. C., Crawford, A. B., & Sinner, C. A. (1999). Single-subject clinical outcome research: designs, data, effect sizes, and analysis. *Aphasiology, 13*, 445-473.

Rojahn, J., & Schulze, H. H. (1985). The linear regression line as a judgmental aid in visual analysis of serially dependent A-B time-series data. *Journal of Psychopathology and Behavioral Assessment, 7*, 191-206.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85*, 185-193.

Schlosser, R. W., Lee, D. L., & Wendt, O. (2008). Application of the percentage of non-overlapping data (PND) in systematic reviews and meta-analyses: A systematic review of reporting characteristics. *Evidence-Based Communication Assessment and Intervention*, *2*, 163-187.

Scotti, J. R., Evans, I. M., Meyer, L. H., & Walker, P. (1991). A meta-analysis of intervention research with problem behavior: Treatment validity and standards of practice. *American Journal on Mental Retardation, 96*, 233-256.

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education, 8*, 24-33.

Scruggs, T. E., & Mastropieri, M. A. (2013). PND at 25: Past, present, and future trends in summarizing single-subject research. *Remedial and Special Education, 34*, 9-19.

Shadish, W. R. (2014). Analysis and meta-analysis of single-case designs: An introduction. *Journal of School Psychology, 52*, 109-122.

Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention, 2*, 188–196.

Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: New applications and some agenda items for future research. *Psychological Methods, 18*, 385-405.

Sierra, V., Solanas, A., & Quera, V. (2005). Randomization tests for systematic single-case designs are not always appropriate. *The Journal of Experimental Education, 73*, 140-160.

Sil, S. Dahlquist,L. M., & Burns, A. J. (2013). Videogame distraction reduces behavioral distress in a preschool-aged child undergoing repeated burn dressing changes: A single-subject design. *Journal of Pediatric Psychology, 38,* 330–341.

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

Solanas, A., Manolov, R., & Onghena, P. (2010). Estimating slope and level change in N=1 designs. *Behavior Modification, 34,* 195-218.

Solomon, B. G., Howard, T. K., & Stein, B. L. (2015). Critical assumptions and distribution features pertaining to contemporary single-case effect sizes. *Journal of Behavioral Education, 24*, 438-458.

Solomon, B. G., Klein, S. A., & Politylo, B. C. (2012). The effect of performance feedback on teachers' treatment integrity: A meta-analysis of the single-case literature. *School Psychology Review, 41*, 160–175.

Stewart, K. K., Carr, J. E., Brandt, C. W., & McHenry, M. M. (2007). An evaluation of the conservative dual-criterion method for teaching university students to visually inspect AB-design graphs. *Journal of Applied Behavior Analysis, 40*, 713-718.

Strain, P. S., Kohler, F. W., & Gresham, F. (1998). Problems in logic and interpretation with quantitative syntheses of single-case research: Mathur and colleagues (1998) as a case in point. *Behavioral Disorders, 24,* 74–85.

Svanberg, J., & Evans, J. J. (2014). Impact of SenseCam on memory, identity and mood in Korsakoff's syndrome: A single case experimental design study. *Neuropsychological Rehabilitation, 24,* 400-418.

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*, 213-230.

Tate, R. L., Perdices, M., Rosenkoetter, U., Wakima, D., Godbee, K., Togher, L., & McDonald, S. (2013). Revision of a method quality rating scale for single-case experimental designs and n-of-1 trials: The 15-item Risk of Bias in N-of-1 Trials (RoBiNT) Scale. *Neuropsychological Rehabilitation, 23,* 619-638.

Taylor, L. K., & Weems, C. F. (2011). Cognitive-behavior therapy for disaster-exposed youth with posttraumatic stress: Results from a multiple-baseline examination. *Behavior Therapy, 42*, 349–363.

Tukey, J. W. (1977*). Exploratory data analysis*. London, UK: Addison-Wesley.

Tunnard, C., & Wilson, B. (2014). Comparison of neuropsychological rehabilitation techniques for unilateral neglect: An ABACADAEAF single-case experimental design. *Neuropsychological Rehabilitation, 24,* 382-399.

Ugille, M., Moeyaert, M., Beretvas, S. N., Ferron, J., & Van Den Noortgate, W. (2012). Multilevel meta-analysis of single-subject experimental designs: A simulation study. *Behavior Research Methods, 44,* 1244-1254.

Ugille, M., Moeyaert, M., Beretvas, S. N., Ferron, J., & Van Den Noortgate, W. (2014). Bias corrections for standardized effect size estimates used with single-subject experimental designs. *Journal of Experimental Education, 82*, 358-374.

Van Den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment and Intervention, 2*, 142-151.

Vannest, K. J., Harrison, J. R., Temple-Harvey, K., Ramsey, L., & Parker, R. I. (2011). Improvement rate differences of academic interventions for students with emotional and behavioral disorders. *Remedial and Special Education, 32*, 521–534.

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

Vilardaga, R. (2014). Technical, practical and analytic innovations in single case designs for contextual behavioral scientists. *Journal of Contextual Behavioral Science, 3*, 136–137.

Vissenaeken, M. (2015). *Reviews of single-case studies: A systematic description.* Unpublished master thesis. KU Leuven, Belgium.

Waddell, D. E., Nassar, S. L., & Gustafson, S. A. (2011). Single-case design in psychophysiological research: Part II: Statistical analytic approaches. *Journal of Neurotherapy, 15*, 160-169.

Wendt, O. (2009). *Calculating effect sizes for single-subject experimental designs: An overview and comparison*. Paper presented at The Ninth Annual Campbell Collaboration Colloquium, Oslo, Norway. Retrieved June 29, 2015 from http://www.campbellcollaboration.org/artman2/uploads/1/Wendt_calculating_effect_sizes.pdf

Wilkinson, L., & The Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 694-704.

Winkens, I., Ponds, R., Pouwels-van den Nieuwenhof, C., Eilander, H., & van Heugten, C. (2014). Using single-case experimental design methodology to evaluate the effects of the ABC method for nursing staff on verbal aggressive behaviour after acquired brain injury. *Neuropsychological Rehabilitation, 24*, 349-364.

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *Journal of Special Education, 44*, 18-29.

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.

Young, N. D., & Daly III, E. J. (2016). An evaluation of prompting and reinforcement for training visual analysis skills. *Journal of Behavioral Education, 25*, 95-119.