Single-case experimental designs:

Reflections on conduct and analysis

Rumen Manolov[1,2], David L. Gast[3], Michael Perdices[4,5] and Jonathan J. Evans[6]

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

[2]ESADE Business School, Ramon Llull University, Spain

[3]Communication Sciences and Special Education, University of Georgia, USA

[4]Department of Neurology, Royal North Shore Hospital, Australia

[5]Discipline of Psychiatry, Sydney Medical School, University of Sydney, Australia

[6]Institute of Health and Wellbeing, University of Glasgow, Scotland, UK

Running head: SCED conduct and analysis

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.

Authors' note

This work was partially supported by the *Agència de Gestió d'Ajust Universitaris i de Recerca de la Generalitat de Catalunya* grant *2009SGR1492*.

Abstract

In this editorial discussion we reflect on the issues addressed by, and arising from, the papers in this Special Issue on Single Case Experimental Design (SCED) study methodology. We identify areas of consensus and disagreement regarding the conduct and analysis of SCED studies. Despite the long history of application of SCEDs in studies of interventions in clinical and educational settings, the field is still developing. There is an emerging consensus on methodological quality criteria for many aspects of SCEDs, but disagreement on what are the most appropriate methods of SCED data analysis. Our aim is to stimulate this ongoing debate and highlight issues requiring further attention from applied researchers and methodologists. In addition we offer tentative criteria to support decision making in relation to selection of analytical techniques in SCED studies. Finally, we stress that large-scale interdisciplinary collaborations, such as the current Special Issue, are necessary if SCEDs are going to play a significant role in the development of the evidence base for clinical practice.

Introduction

Single-case experimental design (SCED) studies have been used for several decades to examine the effectiveness of interventions in clinical and educational settings. Despite this long history SCED methodology continues to develop, stimulated in recent years by the evidenced-based practice movement (Tate et al., 2013). If SCED studies are to make a significant contribution to the development of clinical and educational practice guidelines, there is a need for consensus on what constitutes high quality methodology. For some aspects of SCED methodology such a consensus exists, whilst for others there remains ongoing debate, with the hottest topic of consideration being over how SCED data should be analysed. This Special Issue focused on two complementary themes - how single-case experimental design studies should be conducted and how the results should be evaluated. Our aim was to illustrate SCED research methodology to applied researchers working in clinical and educational fields, highlighting the strengths and limitations of the methodology, and to stimulate further the debate between methodologists as to how best to conduct studies and analyse data. Whilst this is a special issue of the journal Neuropsychological Rehabilitation, a field in which SCED studies are commonly used, the issues that are addressed here are relevant to scientist practitioners in all clinical and educational contexts in which SCEDs are used.

It is not our intention to summarise all of the information presented in the different contributions. Rather, we will identify areas of consensus, discuss key issues, and pose questions for future research and debate, both in terms of how SCED studies should be planned and carried out and how the data gathered should be analysed to enable conclusions to be drawn regarding the presence and magnitude of change over the time course of an intervention. Relatively more space is dedicated to analysis than to conduct, given that this topic is more controversial, something that is reflected in the number of papers focusing on methods of analysis in this Special Issue.

Conducting a SCED Study

Conducting a SCED study: Aspects on which there is a reasonable degree of consensus

We begin with a brief description of the steps that should be followed when conducting a SCED study, though for more details we recommend consulting the major textbooks on the topic (e.g., Barlow, Nock, & Hersen, 2009; Gast, 2010b; Kazdin, 2011; Kennedy, 2005).

First, the aim of the study should be specifically established with respect to the client(s), the problematic issues or behaviours, and the type of intervention(s) whose effect is to be studied empirically. At this point the researcher decides whether a single-case experimental design is the optimal choice for gathering evidence on intervention effectiveness.

Second, the most appropriate design should be identified (see Kratochwill et al., 2010; see also Smith, 2012). A key question in relation to deciding on the best design is the question of whether the target behaviour is considered to be reversible or not. For return-to-baseline, or withdrawal designs (e.g., ABAB), it is necessary to establish that it is feasible to 'withdraw' an intervention, and that the target behaviour is expected to reverse (or return to near baseline levels) upon withdrawal. For example, in a study of a psychological therapy, whilst one can stop therapy sessions, it is not possible (or desirable) to withdraw the learning that has taken place in therapy. Therefore ABA/ABAB designs are not suitable for this type of intervention. This challenge may lead one to rely on a simpler AB design or non-concurrent multiple baseline design, but as Tate et al. (2013) note, the AB design does not have sufficient experimental control (i.e., it does not control for simple change over time). The same problems of experimental control are applicable with the non-concurrent multiple baseline design (Gast & Ledford, 2010). AB designs may therefore be useful in testing the feasibility of an intervention, but cannot provide definitive evidence of treatment effectiveness.

To obtain stronger evidence for the functional relationship between an intervention and a particular outcome, it is necessary to replicate the manipulation, that is, to have available several programmed changes in the conditions (from non-intervention to intervention and vice versa). If the behaviour cannot be reversed, alternative designs such as a multiple-baseline design can be used across several behaviours, settings, or participants (Gast & Ledford, 2010), representing either a within-paricipant replication of the treatment effect (in the former two cases) or a between-participants replication of the treatment effect (in the latter case). The general rule of changing only one variable at a time is recommended, in order to be able to pinpoint the cause of any behavioural changes contingent with the changes in the conditions (Barlow et al., 2009).

However in applied research the independent variable may be an intervention package, in which case it is important to identify the procedural differences between adjacent phases At this point, it can be decided whether randomisation can be included to improve internal validity (Kratochwill & Levin, 2010; Wampold & Furlong, 1981).

Third, the specifics of the intervention are determined, according to the needs of the client, and taking into account evidence for the functional relationship between the intervention techniques and the primary outcome target. Ideally this should be done in negotiation with the client, discussing the form and timing of the intervention.

Fourth, when data collection begins, so data analysis then also commences, and visual analysis is used to perform an ongoing assessment of the data. In SCEDs, evaluation of the baseline data is crucial. Baseline data provides knowledge on the initial situation and behaviour level and is necessary before introducing the intervention (Kratochwill et al., 2010). Tate et al. (2013) recommend at least five measurement points per phase. A stable baseline increases confidence that any subsequent changes are due to the intervention (Kazdin, 2001; Smith, 2012). It has also been suggested that unstable baseline data are exactly what makes statistical analysis necessary (Kazdin, 1978). Visual analysis is also useful for monitoring the progress of the client and making timely adjustments in experimental procedures, therefore saving inappropriate allocation of resources (Fahmie & Hanley, 2008; Johnston & Pennypacker, 2009).

Fifth, when all data have been gathered the researcher can perform a visual analysis of the whole data pattern and to describe patterns of change within and between phases (Gast & Spriggs, 2010; Kratochwill et al., 2010). Furthermore, a method of quantification of the differences in target outcome measures between phases can be selected. This should provide a measure of effect size and enable statistical decision making regarding confidence in change in key outcome measures. The choice of a quantitative technique will depend on the aims of the study and also on the characteristics of the data. Some data features such as the presence of baseline trend and the amount of data variability can be assessed through visual inspection (Parker, Cryer, & Byrns, 2006), but for others such as autocorrelation, the question of their presence and magnitude is less straightforward (Huitema & McKean, 1991; Solanas, Manolov, & Sierra, 2010). Later, we provide some tentative criteria to help researchers select the appropriate type of analysis for their study.

Sixth, when writing up a SCED study reporting guidelines should be followed (Tate et al., this issue; Tate, Togher, Perdices, McDonald, & Rosenkoetter, 2012). Accurate reporting is not just a formality, but is critical to allowing study findings to be critically appraised and studies to be replicated. Good reporting is also vital in terms of the impact of research. The increasing use of systematic reviews for developing clinical guidelines means that poorly described studies fail to be reflected in clinical guidelines, however well the studies may have actually been carried out.

Once the study has been completed, the need for replication should be considered. Well-conducted SCEDs have strong internal validity (Howick et al., 2011), but external validity is necessarily related to gathering repeated evidence on different participants, in different contexts, and by different researchers (Gast, 2010a; Sidman, 1960). For instance, Kratochwill et al. (2013) suggest the 5-3-20 rule for establishing the evidence basis of an intervention, requiring at least five SCED studies conducted by at least three different research teams with a minimum of 20 cases in total. Tate et al. (this issue) emphasise that the within-design replications carried out to establish cause-effect relations (e.g., in a multiple baseline or an ABAB design) does not constitute replication of the study, which can be designed as a direct replication (done by the same researcher with different participants or in different settings in order to establish that the effect observed is reliable) or systematic replications (varying certain factors to establish the generality of the findings).

Conducting a SCED study: Issues for further research and discussion

In this section we would like to highlight issues relating to the conduct of SCED studies that require further reflection and discussion. The development of methodological standards (e.g. Ledford & Gast, this issue; Kratochwill et al., 2010; Kratochwill et al., 2013) and measures of methodological robustness (Tate et al., 2013) has raised the bar on what constitutes methodological rigour in a SCED study. But, is there a risk here if it is not possible for some, or even most, real-world studies to meet all the criteria? Will high expectations of study quality put off clinicians, educators, or researchers from conducting SCED studies? Will journal editors and reviewers err on the side of rejection of papers not meeting the highest standards? Some methodological quality criteria are very difficult to achieve in many situations where SCEDs

would be run (e.g., blinding of therapist and participant to study phase, an item in Tate et al.'s RoBiN-T scale). However, we would argue that this shouldn't put off applied researchers. The same issues affect group studies, in that double-blind placebo-controlled randomised controlled trials are not feasible for many psychological interventions. We would argue that there is a place for studies that do not reach the highest level of methodological rigour, particularly in documenting evidence for interventions that may be useful (a form of feasibility or pilot study), but that the presence of methodological standards and quality scales should serve to highlight those aspects of a study that should be considered in the planning stages. The study planning process often has to balance requirements for scientific rigour against clinical pressures, particularly the need to start an intervention urgently. A common question when attempting to establish a stable baseline as a reference against which to measure change, is whether it is feasible or ethical to wait (and for how long) before offering the client a potentially useful intervention? But quality standards may prompt the researcher to consider whether, for example, it is possible to include randomisation in order to enhance conclusion validity? Is it feasible to have more than one person assessing the information so that outcome measure reliability can be determined?

There are also issues regarding the intervention. How does a practitioner find the balance between offering a tailored intervention focussed on the individual client while also maintaining procedural fidelity (Ledford & Gast, this issue) to ensure replicability and comparability?

Another challenging issue is the question of when a SCED study is just good clinical practice (e.g., a clinician making sure that an individualised intervention is actually working) and when is it research? If it is research there is the inevitable large quantity of paperwork associated with an ethics approval process that typically takes several weeks in most countries with well-developed ethics review processes. This is clearly something of a grey area. SCED studies are focused on interventions aimed at helping an individual so there will always be the potential for the participant to benefit from the intervention. So one might argue that running a SCED study should not be defined as research. But if the intervention is novel and if there is an alternative, well established standard practice intervention, then using the novel intervention would constitute research. Furthermore, if the intention from the outset is to publish the data, then again perhaps this defines the study as research. But what seems most important is that these issues to

do not put off clinicians from using SCED studies opportunistically and publishing data that will contribute to the evidence base. It would be useful for a consensus statement on ethics and SCEDs to be developed.

Finally, it is clear that methodological standards and quality scales enable the consumers of research reports to identify methodological short-comings and hence facilitate the critical appraisal of the strengths and limitations of study findings. Notwithstanding this, it would be important not to discard the evidence gathered only because a study does not meet all requirements – it is a question of the degree of confidence that researchers and practitioners can have in the evidence rather than labelling it as useful or not.

Analysing SCED Data

We turn now to the issue of analysis of SCED data, which has been a focus of several papers in this Special Issue. Whilst visual analysis continues to be an important component of the process of interpreting SCED results (see Lane and Gast, this issue), the arguments for the use of statistical analysis techniques for SCED studies are, we believe, overwhelming. At the very least all SCED reports should present raw data to facilitate meta-analysis. Randomised N of 1 trials have been recognised as presenting the highest level of evidence for the effectiveness of an intervention (Howick et al., 2011), and therefore have excellent internal validity. However external validity depends upon replication and provision of raw data allows effect sizes to be calculated to be incorporated into meta-analyses, which are critical to building a solid evidence base to support the development of practice standards and clinical guidelines (Busse, Kratochwill, & Elliott, 1995; Parker & Hagan-Burke, 2007). However, whilst we might accept that statistical analysis is important, what is less clear is just what form of statistical analysis is most appropriate for SCED data.

In this Special Issue our intention was to present as many theoretically and empirically supported alternative methods of analysis of SCED data as possible. However, inevitably not every potential approach is addressed. All of the techniques discussed have focused on the individual assessment of the behaviours of participants in which each participant serves as his/her own control. In that sense, a set of techniques potentially useful in Neuropsychology, but

not covered here, are the ones focused on a case-controls design (i.e., the comparison of an individual, measured in a single point in time, to a small, well-matched, control sample representing normative behavioural levels; Crawford & Garthwaite, 2012; Crawford & Howell, 1998). There have been several developments in relation to this analytical option, including statistical tests and effect size estimates for comparing the individual's score to the reference provided by the control group (Crawford, Garthwaite, & Porter, 2010), as well as the possibility of detecting a differential deficit in some tasks or behaviours but not in others (Crawford & Garthwaite, 2005). These techniques have potential application in intervention outcome research, but have not been used in this way to our knowledge, and hence the following discussion is related to the set of techniques focussing on traditional SCED studies.

Our aim in this paper is to not to advocate for one method of analysis over another, but rather to propose some tentative criteria to help researchers identify the most appropriate technique for each specific experimental situation. These criteria are related to Ruscio's (2008) ideas that effect size indices should be simultaneously easily understood, show appropriate performance in a variety of conditions, and require fewer assumptions about the data features. Although the criteria themselves are proposed for discussion and are in no way definitive, we consider them to be useful for making a structured comparison between the different approaches to SCED analysis. Selecting the right analytical procedure may be crucial, given the influence of the choice of an effect size measure on the interpretation of the results (McGrath & Meyer, 2006). As with the previous section on the conduct of SCED studies we again discuss aspects of analysis on which there is some consensus, and areas where further investigation and discussion are needed.

Analysing SCED Data: Aspects on which there is a reasonable degree of consensus

Those aspects of analysis that are commonly accepted are reflected in recent reporting guidelines (Tate et al., this issue; Tate et al., 2012), design and analysis standards (Kratowchwill et al., 2010) and a methodological quality scale (Tate et al., 2013) although there are some differences between the standards developed in different fields (Smith, 2012).

The most basic requirement is to report the raw data obtained at each measurement point for each participant, setting, or target behaviour. This condition is commonly met using a graphical representation, which is consistent with the historical origins of SCED analysis (Michael, 1974; Skinner, 1938), and has two main benefits: a) readers of the report can reach their own conclusions about intervention effectiveness (Barlow et al., 2009; Johnston & Pennypacker, 2009); b) the raw data can be retrieved and used in further analyses using a different numerical indicator or they may become part of quantitative integrations. In addition, presenting raw data in tabular form would also facilitate subsequent analysis and meta-analysis, removing the need for data to be extracted from graphs, something that is not always easy (see also Shadish & Sullivan, 2011). With increasing opportunities to deposit supplementary data accompanying papers on publishers' websites, this is perfectly possible. Alternatively it might be possible to develop a SCED data depository, an idea we explore later in this paper.

A second aspect accepted by many researchers (e.g., Davis et al., 2013; Fisch, 2001; Franklin, Gorman, Beasley, & Allison, 1996; Smith, 2012) is the need to complement the visual analysis with a quantitative summary of the results, whose interpretation is augmented by the visual inspection itself (Parker et al., 2006). The justification for this numerical summary is based on the need for objective and replicable outcomes (Robey, Schultz, Crawford, & Sinner, 1999), which can then be used for meta-analytical purposes (Busse et al., 1995). As this Special Issue has demonstrated, what is less clear is which of the myriad of techniques available should be used to provide the quantitative indicator. At the present time it is the case that no single procedure is appropriate for all aims and for all types of data, regardless of how these are gathered.

In the following section we outline some tentative criteria that researchers could consider when choosing how to analyse their SCED data.

Analysing SCED Data: Tentative Criteria for Choosing an Appropriate Method of Analysis.

The method should reflect the aim of the analysis. The first task when selecting a method of analysis is to identify the aim of the analysis. As Smith (2012) observed, the analytical

approach should be chosen according to the research question(s) asked or hypothesis tested. A likely primary aim is to obtain evidence that an intervention is having an effect, i.e., whether the change in conditions can be considered to have a functional relationship with changes in behaviour over time. Demonstrating experimental control requires that the design offers sufficient opportunities to explore whether the change in condition is associated with a change in the behaviour of interest (Horner et al., 2005; Kratochwill et al., 2010). If the aim is to explore whether a study with an appropriate design has shown a clear intervention effect, visual analysis may be sufficient, if carried out in a systematic way as suggested by Kratochwill et al. (2010). A second aim might be to assess the statistical significance of the results, in order to make inferences in studies in which random sampling has taken place, or as an indicator of the likelihood that changes in data are the result of the intervention as compared to the likelihood of there being no relationship between phase changes and outcome measures (i.e., if the null hypothesis were true). In this case, techniques such as randomisation tests (Heyvaert & Onghena, this issue) and Simulation modelling analysis (Borckardt & Nash, this issue) can be used, and *p* values can also be obtained for the standardized effect arising from Swaminathan, Rogers, Horner, Sugai, & Smolkowski's (this issue) regression model, and Tau-U indicator (Brossart, Vannest, Davis, & Patience, this issue). A third possible aim of analysis might be to compute an effect size measure in terms of a common metric; thus, in contrast to the previously mentioned aim, here the focus would be put on the strength of association (as a descriptive measure) and not on statistical significance for inferential purposes. One of the most common effect size metrics is R-squared (i.e., variability in the behaviour explained by the different conditions), which can be thought of as a general quantification of the intervention effect and obtained from a regression analysis. Alternatively one might focus on a more specific aspect of the data such as the amount of data overlap between phases (e.g., the *Nonoverlap of all pairs* [Parker & Vannest, 2009] or the Tau-U indices), the change in level (e.g., a standardized mean difference such as the *d* statistic; Shadish et al., this issue; the unstandardized *Mean phase difference;* Manolov & Solanas, 2013; see also Swaminathan et al., this issue) or the change in slope (e.g., the *Slope and level change* procedure [Solanas, Manolov, & Onghena, 2010], also offering unstandardized estimates). Of these, the R-squared and the *d* statistic are classic parametric effect size indices, though the nonoverlap measures are commonly applied to SCED studies. A fourth potential aim of analysis is that there might be the need to incorporate moderator variables in the analysis, in order to

quantify the effect of, say, participant's characteristics; multilevel analysis (Baek et al., this issue) can be used for that purpose . Thus, the choice of a technique would be based on the relative importance given by the researcher to one (or more) of these four possible aims of analysis. It is quite possible of course that all four aims would apply.

The output of the analysis should be easy to interpret. Being easy to comprehend is a positive feature of any analytical technique (Maggin & Chafouleas, 2013). But beyond this general prerequisite, three specific issues are considered.

First, it is relevant whether the result yielded is standardized or not. An unstandardized measure may be useful when the target behaviour is measured in meaningful terms (Cumming, 2012), such as the number of times medication is forgotten in a memory rehabilitation study, the number of cigarettes smoked when studying nicotine dependence or number of binge/purge episodes when studying eating disorders. Davis and colleagues (2013) have suggested that meaningful measurement units help assess the practical significance of the results, beyond their statistical significance. Thus unstandardized measures are called for when the absolute terms help assess clinical relevance of the change over time. By contrast, when the aim is to make the results of different studies comparable, standardized (e.g, *d* statistic) and bounded (e.g., R-squared, or a percentage of nonoverlap index) indices are more useful. These latter measures also make post-hoc combinations of results from different studies easier. For standardized measures an additional benefit is the option of converting the result into other common effect size measures (Borenstein, 2009). Moreover, in some cases (e.g., the *d* statistic for SCED) knowing the sampling distribution enables assessing the statistical significance of the effect size measure, apart from focusing on its magnitude. Nevertheless, sampling distributions are not usually known, especially if autocorrelation has to be taken into account.

Second, it is important to know whether interpretation of the numerical value can be aided by conventional benchmarks. Whilst such conventions can be helpful, they are more useful if it is possible to provide a measure of degree of freedom when labelling the magnitude of the effect observed as being "small" or "large". However, this issue is still controversial as it might be said that such categorization can be potentially misleading and it is important that specific benchmarks are not treated as absolute cut-offs in the way that $p \leq 0.05$ is sometimes treated in inferential statistics (Cohen, 1994).

Third, it has been suggested that given the frequent use of visual analysis, it might be helpful for analytical techniques to be related to the graphical representation of the data (Parker et al., 2006; Wolery, Busick, Reichow, & Barton, 2010). Techniques that meet this criteria are those that quantify one (or preferably several) of the data features relevant to visual analysis: level, slope, variability, overlap, etc. (Kratochwill et al., 2010).

The analysis should be easy to compute. This criterion considers the ease or difficulty of the calculation procedure. Scruggs and Mastropieri (2013) suggest that easy-to-obtain techniques lead to fewer errors in the computational process. Kratochwill and colleagues (2013) suggested that technical difficulties are a potential reason for the limited use of statistical analysis in SCED studies. Schlosser, Lee, and Wendt (2008) consider that the "theoretical strengths and weaknesses" (p.184) of the techniques need to be considered together with issues related to their application in everyday psychological practice.

In line with these comments, it has to be mentioned that with some methods of analysis, it is possible to obtain the result by hand calculation relatively quickly, but with other methods specialist software is necessary for carrying out the more tedious computations. The choice of technique might therefore be related to whether the necessary software is available to the researcher, though it should be noted that several analysis programs are available for free. Finally, even when software is available (for techniques requiring intensive computations, data transformation, or estimation of parameters), it is important to take into account the number of steps necessary to obtain the results, given that it is desirable that applied researchers can understand the output provided by the method of analysis.

We do not advocate choosing a technique only on the basis of ease of calculation, given that this does not guarantee meaningful or correct results. Ease of use of any technique is in part related to level of training. In that sense, both experienced and new researchers should try to obtain as much training as possible, or collaborate with more expert peers.

The method of analysis must take into account design requirements and data assumptions. Some methods of analysis have certain requirements and hence place demands on study design. For example, application of randomisation tests depends upon having a minimal number of possible random assignments (and therefore data points) to make it theoretically

possible to obtain a sufficiently small *p* value (in a simple AB design this is 20). Some other methods also require long data series (e.g., ARIMA), or baseline stability (e.g., the current version of *d* statistic by Hedges and colleagues). With some other methods valid interpretation of numerical output requires assumptions that the data or the residuals have certain features (e.g., normality, independence). Therefore, there may be an iterative study planning process, considering both design and method of analysis, with each influencing each other. This highlights the importance of considering what analysis technique will be used as part of the study planning process and not as an afterthought, albeit there has to be some flexibility to change the method according to characteristics of the data (as is the case in use of parametric and non-parametric methods in group studies).

Select a method based on evidence of performance with typical SCED data. Choice of an analytical technique should also depend on whether the technique has been demonstrated to function properly. Schlosser and Sigafoos' (2008) referred to this in terms of the need for "empirically guided selection of metrics" (p. 118). The criteria by which a method of analysis is judged may be classical, based on statistical properties such as Type I error rates (i.e., the probability of getting a "positive" result when there is actually no intervention effect) and power (i.e., the probability of getting a "positive" result when there *is* an intervention effect). Similarly, there should be evidence of the accuracy of effect size estimates (Kratochwill et al., 2013). Performance can be obtained from simulation studies in which an intervention effect is programmed to be present (with a specified magnitude) or not. Thus simulations studies can determine how well the procedures discriminate between data with and without effect. Also relevant is precision of effect estimates, i.e. which procedures and statistics show lower standard error and, thus, offer the researcher a higher degree of confidence that the numerical values yielded are an accurate representation of the data gathered. Performance of the procedures has to be assessed taking into account the common features of SCED data, such as the presence of autocorrelation, baseline trend, or outliers. Do these data features distort the numerical values provided by the technique and to what extent? Sometimes a procedure may not include an a priori assumption regarding the characteristics of the data, but its application may be restricted to certain conditions, such as only working with independent data, stable data, or data presenting linear trend.

In addition to simulation studies, evidence of effective performance with real psychological data sets is also valuable. For instance, the information on typical values (including the range and several key percentiles) can be used to assess whether values yielded by a quantitative procedure allow discrimination between data sets with different magnitudes of effect. This will identify undesirable features such tendency to produce floor and ceiling effects, representing respectively lack of discrimination due to accumulation of values at the lower or upper bound of the numerical indicator. Moreover, when the analysis techniques are applied to real data, it is possible to validate the results they produce (Scruggs & Mastropieri, 2013), comparing outputs from the analysis with the professional's opinion of treatment effects, with the evolution of the behaviour of interest at a follow-up stage, or in comparison with other studies on the topic.

Final remarks on the tentative criteria. Four further issues need to be highlighted in relation to the criteria presented above. First, whilst the criteria have been presented and discussed, no weights have been given to them. The optimal situation would be for one or more analysis methods to meet all of the criteria and indeed to have other positive features not included here. But in the (common) situation in which only some of the criteria are met, questions perhaps arise as to which of the criteria should be prioritised. For instance, is it more important for a technique to be easily calculated and understood (i.e., interpreted correctly) or to function properly for a variety of situations? Ease of calculation and interpretation are of little value if the analysis outcome does not adequately represent the intervention effect. On the other hand, if a procedure estimates well the intervention effect, but its values are prone to miscalculation or misinterpretation due to the complexity of the technique, is this procedure actually useful? Finally, is an intuitive technique that only performs appropriately in a very restricted set of conditions to be recommended? Which characteristics of analytic techniques are therefore most important? From a methodological perspective, it is essential for an analytical technique to have desirable statistical properties, to be as unbiased and efficient as possible, to be sensitive to effects that are genuinely present and not distorted by extraneous variables. From an applied perspective, it is essential to be able to apply the technique without additional costs associated with training. Our position is that if, or when, an optimal analytical technique is identified on the basis of evidence (from simulation and real-data studies), its degree of complexity should not be an issue if there is an expert methodologist/statistician on the research

team, ensuring correct application and interpretation of the analysis, something that is common in clinical trials.

Second, what is clear from this Special Issue is that there is currently no single universally optimal SCED analysis technique. Different analytical techniques are optimal in different circumstances and so it may be necessary to choose a procedure not only as a function of the researcher's aim, but also considering the features of the data set to be analysed (e.g., number of measurements available, baseline trend, autocorrelation, variability). Therefore the techniques need not be seen as competitors for the ultimate prize of best SCED analysis; rather they can be used together in many cases. The joint use of visual and quantitative analyses is particularly useful, given that the former can serve as an initial exploration of the data, as a means of choosing the latter and might also help interpreting the output of the quantitative analysis.

Third, we hope that the collection of papers in this special issue will introduce applied researchers to a number of recently developed analysis techniques. To some extent the most useful techniques will be defined by those that are actually used in practice. Which techniques are used will in part depend on provision of training and support for their use (such as in the form of computer analysis packages) and we return to this issue later.

Fourth, the issues considered in the preceding section, and a lot of those raised in the next section, have been explored and discussed primarily (if not exclusively) in the context of withdrawal/reversal and multiple-baseline designs, given that they represent the vast majority of designs used (e.g., 62% of the designs according to Shadish and Sullivan, 2011; and 86% according to Smith's, 2012, review). In contrast, other design structures may raise their own specific issues as discussed in the following section.

Analysing SCED Data: Issues for Further Research or Discussion

In this section we will point to issues that remain to be addressed by researchers if optimal methods of analysis for SCED data are to be determined. We group them into three subsections, according to whether they relate to the interpretation of results, to the mechanical application of data analysis procedures, or to the integration of results of several studies.

Analytical and interpretative challenges. In the last few years, in the context of the evidence-based practice movement (e.g., APA Presidential Task Force on Evidence-Based Practice, 2006; Jenson, Clark, Kircher, & Kristjansson, 2007) and changes in journal policies (Wilkinson & The Task Force on Statistical Inference, 1999), more attention is paid to effect size indices. Thus, the long standing call to overcome the limitations of *p* values (Cohen, 1994) seems to have been heard. However, it is important not to repeat some of the same mistakes that have occurred in interpreting *p* values when interpreting effect sizes. For example whilst categorising results into "small", "medium", and "large" effects might provide some benchmarking of outcomes, reification of these categories and the precise cut-offs will lead to the same problems that have beset *p* value interpretation (Cortina & Landis, 2011). In the SCED context there is an additional complication, given that Cohen's (1988) interpretative guidelines are considered to be unsuitable (Matyas & Greenwood, 1990; Parker et al., 2005). In fact, one of the priorities in the field, as noted by the US Institute of Education Sciences (2013) is to revise these benchmarks and substitute them with more suitable ones. Research is necessary though to identify such benchmarks. A promising approach to developing SCED benchmarks involves the combined use of visual and statistical analyses (Parker & Vannest, 2009; Petersen-Brown, Karich, & Symons, 2012) to obtain evidence on what values of the numerical indicators can reasonably be labelled "small" or "large", using visual analysis results as the reference standard. A related challenge is how to compare intervention effects quantified with different primary indicators, for instance, a standardized mean difference arising from generalised least squares regression (Maggin et al., 2011) and the Nonoverlap of all pairs (Parker & Vannest, 2009). This issue is easily solved when raw data are available, but when not additional numerical evidence might be used (Manolov & Solanas, 2012), although further discussion is needed.

Regarding the specific issues raised for specific design structures, we should focus on alternating treatment designs (ATD) where the experimental effect in is demonstrated if levels of the dependent variable for each intervention do not overlap. For changing criterion designs, experimental effect is demonstrated if criteria that trigger the next phase change are met. But definitions and "metric" of these requirements are, at best, vague. Specifically, the lack of overlap in levels of the dependent variable in an ATD study gives only qualitative information about the relative effect of the treatments, whereas quantitative criteria are missing. For instance, how much difference should there be between treatments before it is considered significant (i.e.,

that one treatment is really better than the other)? How would treatment effects be calculated in this situation if (as is often the case in ATDs) there is no baseline? Is the difference between two treatments best determined in terms of effect sizes reflecting changes in mean, level, slope?

As far as the breadth of applicability of the analytical techniques is concerned, it has to be stressed that most methodological research and most illustrative papers have focussed on immediate level or slope changes. Therefore, more attention should be paid to situations in which the effect is either temporary, gradual, or delayed (Lieberman, Yoder, Reichow, & Wolery, 2010). Is the demonstration of the functional relationship between condition and behaviour compromised when the effect is gradual or delayed? Is it justified to compute a summary indicator considering only part of the data? How do techniques perform in such situations?

A final question regarding application is whether applying any statistical analysis (aimed to provide an objective summary) is preferred to using only visual analysis, even if there is (still) no conclusive evidence on the appropriateness of the former. On this issue, Tate et al.'s (2013) methodological quality scale awards the highest score on the "data analysis" item if any of the following three apply: a) a systematic visual analysis is used, b) visual analysis is aided by quasi statistical techniques, or c) statistical methods are used and a rationale for their suitability is provided.

In terms of interpretation of summary indicators, there is another more specific question that needs to be answered in the SCED context. The question arises from the fact that many designs entail replications, either between or within participants. These designs, including multiple baseline and ABAB designs, are used commonly (Shadish & Sullivan, 2011; Smith, 2012) and meet the requirement of including several attempts to demonstrate a functional relationship between intervention and outcome (Kratochwill et al., 2010; Tate et al., 2013). With such designs the interpretative challenges comes from the fact that to date the performance of analysis methods have typically been evaluated using AB designs. By contrast, the current need is to obtain a single indicator for the whole design, even when it includes more than two phases. One question arising in this context is whether the whole-design indicator should be a combination of the effect size indices for each two-phase comparison. This seems like the logical option and Schlosser at al. (2008) review several ways in which different researchers have obtained

quantifications for ABAB, multiple baseline, and alternating treatments designs, among others. However, a number of issues remain to be determined:

a) Should the two phases being compared be adjacent (Gast & Spriggs, 2010) or is it also justified to compare an initial baseline phase with the final intervention phase (Campbell, 2003; Faith, Allison, & Gorman, 1996)?

b) Should all two-phase comparisons be considered or are there some comparisons with specific characteristics and issues (as discussed by Parker and Vannest, 2012 and Schlosser et al., 2008)?

c) If all possible comparisons are considered, would it be important to consider a "statistical correction' that is analogous to the Bonferroni correction used in between-group comparisons?

d) Should the two-phase comparisons be weighted and which are the appropriate weights? Among the possibilities for weighting, the following have to be suggested: the standard error of the summary indicator, phase length, and data variability. It is debateable which the optimal solution is when the sampling distribution, and thus the standard error, of the indicator is unknown or cannot be approximated reasonably well asymptotically.

e) When is it justified to combine? Is it necessary to demonstrate first a functional relationship (i.e., that the data pattern observed matches the expected one according to the design structured), considering the importance given to this demonstration (Kratochwill et al., 2010; Parker & Vannest, 2012)? Or would such a conditional combination lead to computing and reporting effect sizes only when there is a clear effect, leading potentially to publication bias (i.e., misrepresenting the actual degree of effectiveness of an intervention)? We believe that there has not yet been sufficient discussion of this topic.

f) Is combining two-phase comparisons the same as combining effect sizes from different studies? Is it necessary with the former for the comparisons (i.e., the data sets) to be independent, as it has been suggested when combining probabilities (Strube, 1985) or effect sizes meta-analytically (Cooper, 2010)? Beretvas and Chung (2008) state that the two-phase comparisons in an ABAB or a multiple baseline design should not be considered as independent, but it is still debatable how the outcomes from multiple phase comparisons in a single study should be treated

before including them in a meta-analysis. Using the average or selecting only one of the outcomes are the main options that Beretvas and Chung's (2008) review concluded were appropriate. These are also options when performing meta-analysis of group-design studies with more than one outcome (Borenstein, Hedges, Higgins, & Rothstein, 2009; Lipsey & Wilson, 2001).

Availability of data analytical techniques. If analysis methods are going to be taken up by applied researchers, they need to be easily accessed. Several of the procedures discussed in the current Special Issue can actually be computed by hand. However, even with relatively simple procedures, it is preferable that calculations are done by computer, reducing likelihood of human error. Moreover, in most analyses the calculations are tedious and/or time-consuming. The list of tools provided below comprises options such as using the open source software R (R Core Team, 2013), stand-alone executable files that do not require any particular software to be installed in the operating system, or web-based applications. Ideally all of the SCED analysis procedures, or at least the ones whose use is deemed advisable, should be implemented in one (or more) packages. The development of such a major package would require, in addition to substantial funding, the collaboration of several researchers, each of whom is expert in one or more techniques, but making access to such tools free of charge will be important in preventing exclusion of applied researchers on economic grounds.

As well as specialist programming, the development of computer based analysis packages also requires documentation to explain the theoretical basis of the procedure, how the data should be entered and how to interpret the outputs. In addition, training workshops may be required, in the form of pre-congress workshops, online presentations, or video tutorials.

Integration of single-case studies. In this Special Issue three alternatives for combining results from SCED studies have been discussed: combining probabilities (Solmi & Onghena), multilevel models (Baek et al.), and the *d* statistic (Shadish et al.). A key issue that remains to be resolved is the question of what weights should be used in cases when the optimal weight (i.e., standard error; Hedges & Olkin, 1985; Whitlock, 2005) cannot be used? Is standard error actually the optimal choice for weighting SCED studies or has such an assumption been accepted without discussion? Some initial evidence is available on that topic (Manolov, Guilera, & Sierra, 2014), but more research is necessary.

Although there are some statistical issues to be resolved, there is also a more practical issue of extracting data for analysis. Given that there is no consensus on what effect size to use, calculation of effect sizes for SCED data relies on use of raw data (unlike traditional group studies where reporting of means and standard deviations is often sufficient). As discussed earlier this means that provision of raw data is important so that researchers performing meta-analyses do not have to rely on "ungraph" techniques (e.g., Bulté & Onghena, 2012; Shadish et al., 2009) or on the response of primary authors to obtain the raw data (Shadish & Sullivan, 2011). A potentially useful option would be to develop a SCED data depository, organised by field of research, which would allow access to raw data to facilitate integration analyses, whilst recognising the authors of the original studies.

Analysing SCED Data: Resources for Applied Researchers

In this section we provide pointers to useful sources of information on specific methods of analysis and to software for SCED analysis.

Illustrations of procedures. For those wanting to learn how to use the various methods of analysis, illustrations of many of the techniques can be found in the literature, including applications in SCED studies. The process of visual analysis is illustrated by Lane and Gast (this issue) and Kratochwill et al. (2010). In the following list we provide reference to applied studies using the techniques included in the Special Issue. The order in this list parallels the order of the contributions in the Special Issue.

- Visual analysis including an assessment of inter-observer agreement and intervention (not procedural) fidelity: Bennett, Ramasamy, and Honsberger (2013).
- Nonoverlap indices: *Nonoverlap of all pairs* was used by O'Neill and Findlay (this issue) and prior to a meta-analysis by Jamieson, Cullen, McGee-Lennon, Brewster, and Evans (this issue), *Improvement rate difference* has been used by Ganz et al. (2012) in a meta-analysis, and Tau-U by Tunnard and Wilson (this issue), *Percentage of nonoverlapping corrected data* (Manolov & Solanas, 2009, commented by Brossart, Vannest, Davis, ^ Patience, this issue): no application published to date.
- Simulation modelling analysis: Jarrett and Ollendick (2012).

- Randomisation tests: Winkens, Ponds, Pouwels-van den Nieuwenhof, Eilander, and van Heugten (this issue).
- Quantifying specific data features: Slope and level change technique (Solanas et al., 2010): Svanberg and Evans (this issue). Mean phase difference technique (Manolov & Solanas, 2013, discussed by Swaminathan et al., this issue): no application published to date.
- The *d* statistic has been illustrated by Shadish and colleagues (this issue) and Shadish, Hedges, and Pustejovsky (2013).
- Regression analysis as described by Swaminathan and colleagues (this issue): Maggin et al. (2011).
- Bayesian analysis: apart from Rindskopf (this issue), de Vries and Morey (2013) provide an example based on a neuropsychological study by Rasquin, Van De Sande, Praamstra, and Van Heugten (2009).
- Multilevel models for integrating individual studies' results: Gage and Lewis (2012) use hierarchical linear models for meta-analytical purposes.
- Combining probabilities for integrating individual studies' results: Holden, Bearison, Rode, Rosenberg, and Fishman (1999).

Software. In this section we will offer a list of computerised SCED analysis tools (statistical packages, stand-alone programs, and web-based calculators). An initial version of a summary document has been elaborated, illustrating how most of the procedures can be applied with the resources currently available, focusing on open source rather than commercial software. It can be found as a supplementary material on the journal's website. The order of the techniques is the same as in the Special Issue.

- Visual analysis – graphing, central tendency, trend, and variability: The SCDA plug-in for R-Commander (Bulté & Onghena, 2012) offers the possibility to represent the data graphically and to add visual aids referring to average level, trend, or data variability, among other options. This plug-in was available from the R website: http://cran.r-project.org/web/packages/RcmdrPlugin.SCDA/index.html and it is also downloadable for free from the R platform itself. However, given that the plug-in is currently not maintained, the following actions need to take place: 1) install the following three

packages in R: SCMA, SCRT, and SCVA – they include the functions for meta-analysis, randomization tests, and visual analysis, respectively; 2) download the SCDA plug-in "RcmdrPlugin.SCDA_0.2.tar.gz" file from the R archive: http://cran.r-project.org/src/contrib/Archive/RcmdrPlugin.SCDA/; 3) write the following code in the R console: *install.packages(file.choose(), repos = NULL, type = "source")*; 4) select the "RcmdrPlugin.SCDA_0.2.tar.gz" file from the location where it was downloaded; 5) load RcmdrPlugin.SCDA in the R console or in R-Commander.

- Visual analysis – estimating and projecting baseline trend: R code is available on the following address https://dl.dropboxusercontent.com/s/5z9p5362bwlbj7d/ProjectTrend.R. The purpose of this code is to estimate baseline trend using the split-middle method (Miller, 1985; White, 1972) and projecting it into the treatment phase. Trend stability across conditions is estimated following the 80%-20% formula (Gast & Spriggs, 2010) and also using the interquartile range (Tukey, 1977). The code has been developed by the first author (RM).
- Visual analysis and training: Another option for visual analysis is the training protocol available at www.singlecase.org/ (content developed by Swoboda, Kratochwill, Horner, Levin, and Albin; copyright of the site: Hoselton and Horner).
- Nonoverlap indices: *Nonoverlap of all pairs* (Parker & Vannest, 2009), *Improvement rate difference* (Parker, Vannest, & Brown, 2009), and *Tau-U* (Parker, Vannest, Davis, & Sauber, 2011) can all be computed online on the website http://www.singlecaseresearch.org/ (Vannest, Parker, & Gonen, 2011). *Tau-U:* R code was developed by Kevin Tarlow and is offered by Brossart et al. (this issue) online via the URL https://dl.dropboxusercontent.com/u/2842869/Tau_U.R. *Percentage of nonoverlapping corrected data*: R code available in the article presenting the procedure (Manolov & Solanas, 2009) and also available online at https://dl.dropboxusercontent.com/s/8revawnfrnrttkz/PNCD.R.
- *Simulation modelling analysis*: available online on the website http://clinicalresearcher.org/software.htm.
- Randomisation tests: R code available in the SCRT package (Bulté & Onghena, 2008; 2009) downloadable for free from the R platform itself and available from the R website http://cran.r-project.org/web/packages/SCRT/index.html. The SCDA plug-in for R-

Commander (Bulté, 2013; Bulté & Onghena, 2012) also includes randomisation tests. Analyses via randomization tests can also be carried out using Excel as a platform, thanks to the work of Boris Gafurov and Joel Levin (http://code.google.com/p/exprt/).

- Quantifying specific data features: *Slope and level change* technique: R code available in the article presenting the procedure (Solanas et al., 2010) and also online at https://dl.dropboxusercontent.com/s/ltlyowy2ds5h3oi/SLC.R; additionally, there is an SLC plug-in for R-Commander available from the R website http://cran.r-project.org/web/packages/RcmdrPlugin.SLC/index.html and downloadable for free from the R platform itself. *Mean phase difference* technique: R code available in the article presenting the procedure (Manolov & Solanas, 2013) and also online at https://dl.dropboxusercontent.com/s/nky75oh40f1gbwh/MPD.R.
- Quantifications in terms of the *d* statistic: SPSS macros and Graphic User Interface (clickable menus) plus manuals are available via William Shadish's website: http://faculty.ucmerced.edu/wshadish/software/software-meta-analysis-single-case-design. R code for these developments is available at James Pustejovsky's page http://blogs.edb.utexas.edu/pusto/software/.
- Regression analysis as presented by Swaminathan et al. (this issue): a computer programme developed in FORTRAN 90 (Rogers & Swaminathan, 2007) has been created.
- Bayesian analysis: Rindskopf (this issue) recommends using WinBUGS program available for free at the website http://www.mrc-bsu.cam.ac.uk/bugs/. An R package called BayesSingleSub (de Vries & Morey, 2013) is available from the R website http://cran.r-project.org/web/packages/BayesSingleSub.
- Multilevel models: several alternative platforms can be used including two specifically designed programs (HLM, MLwiN), the *lme4* and *nlme* packages in R, *proc mixed* and *proc glimmix* in SAS, the *mixed* option using SPSS syntax, and the *gllamm* programme in Stata) of which only R is open-source. WinBUGS can be used also for multilevel models. See also Shadish, Kyse, and Rindskopf (2013). Additionally, a website (http://ppw.kuleuven.be/home/onderzoek/multilevel-synthesis-of-single-case-experimental-data/) is available including theoretical information, examples and code in

relation to multilevel models. This website, expected to grow in near future is also accessible from www.single-case.com.

- Combining probabilities: some options (e.g., the multiplicative approach described in Jones and Fiske, 1953, and the additive approach by Edgington, 1972) are included in the SCDA plug-in for R-Commander (Bulté, 2013; Bulté & Onghena, 2012). Additionally, R code is available on the Neuropsychological Rehabilitation website as supplemental online material to the article authored by Solmi and Onghena (this issue).

Conclusions

SCEDs have the potential to bring research and clinical practice closer together and build a strong evidence base for clinical practice. Recent developments in conduct and analysis should increase the scientific rigour of SCED studies, and whilst this rigour brings demands and complexity, there are now several resources available to support and guide applied researchers throughout the process of conducting a SCED study and analysing the data.

A number of issues remain to be resolved. Research, debate, and international collaborations are the pillars for solving the methodological issues of SCEDs. Outputs such as this Special Issue and similar ones published in other journals such as *Evidence-Based Communication Assessment and Intervention* (Volume 2, Issue 3) in 2008, *Journal of Behavioral Education* (Volume 21, Issue 3) in 2012, *Journal of Applied Sport Psychology* (Volume 25, Issue 1) in 2013, *Remedial and Special Education* (Volume 34, Issue 1) in 2013, *Journal of School Psychology* (articles available online ahead of print as of the time of preparing the present Special Issue – probably going to conform an issue in Volume 52 during 2014), and the SCRIBE project (Tate et al., this issue) show that large scale collaborations are possible in this field. These need to continue.

References

APA Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist, 61*, 271-285.

Baek, E., Moeyaert, M., Petit-Bois, M., Beretvas, S. N., Van de Noortgate, W., & Ferron, J. (2014). The use of multilevel analysis for integrating single-case experimental design results within a study and across studies. *Neuropsychological Rehabilitation, 24,* XXX-XXX.

Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior change* (3rd ed.). Boston, MA: Pearson.

Bennett, K. D., Ramasamy, R., & Honsberger, T. (2013). The effect of covert audio coaching on teaching clerical skills to adolescents with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 43*, 585-593.

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.

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

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., pp. 221-235). New York, NY: The Russell Sage Foundation.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: John Wiley & Sons.

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

Bulté, I. (2013). *Being grateful for small mercies: The analysis of single-case data*. Unpublished doctoral dissertation. KU Leuven, Belgium.

Bulté, I., & Onghena, P. (2008). An R package for single-case randomization tests. *Behavior Research Methods, 40*, 467-478.

Bulté, I., & Onghena, P. (2009). Randomization tests for multiple-baseline designs: An extension of the SCRT-R package. *Behavior Research Methods, 41*, 477-485.

Bulté, I., & Onghena, P. (2012). When the truth hits you between the eyes: A software tool for the visual analysis of single-case experimental data. *Methodology, 8*, 104-114.

Busse, R. T., Kratochwill, T. R., & Elliott, S. N. (1995). Meta-analysis for single-case consultation outcomes: Applications to research and practice. *Journal of School Psychology, 33*, 269-285.

Campbell, J. M. (2003). Efficacy of behavioral interventions for reducing problem behavior in persons with autism: A quantitative synthesis of single-subject research. *Research in Developmental Disabilities, 24*, 120-138.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

Cooper, H. (2010). *Research synthesis and meta-analysis* (4th ed., Vol. 2). London, UK: Sage.

Cortina, J. M., & Landis, R. S. (2011). The Earth is not round (p=.00). *Organizational Research Methods, 14*, 332-349.

Crawford, J. R., & Garthwaite, P. H. (2005). Testing for suspected impairments and dissociations in single-case studies in neuropsychology: Evaluation of alternatives using Monte Carlo simulations and revised tests for dissociations. *Neuropsychology, 19*, 318-331.

Crawford, J. R., & Garthwaite, P. H. (2012). Single-case research in neuropsychology: A comparison of five forms of t-test for comparing a case to controls. *Cortex, 48*, 1009-1016.

Crawford, J. R., Garthwaite, P. H., & Porter, S. (2010). Point and interval estimates of effect sizes for the case-controls design in neuropsychology: Rationale, methods, implementations, and proposed reporting standards. *Cognitive Neuropsychology, 27*, 245-260.

Crawford, J. R., & Howell, D. C. (1998). Comparing an individual's test score against norms derived from small samples. *The Clinical Neuropsychologist, 12*, 482-486.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. London, UK: Routledge.

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.

de Vries, R. M., & Morey, R. D. (2013). Bayesian hypothesis testing for single-subject designs. *Psychological Methods, 18,* 165-185.

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

Fahmie, T. A., & Hanley, G. P. (2008). Progressing toward data intimacy: A review of within-session data analysis. *Journal of Applied Behavior Analysis, 41,* 319-331.

Faith, M. S., Allison, D. B., & Gorman, B. S. (1996). Meta-analysis of single-case research. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 245–277). Mahwah, NJ: Lawrence Erlbaum.

Fisch, G. S. (2001). Evaluating data from behavioral analysis: Visual inspection or statistical models? *Behavioural Processes, 54,* 137-154.

Franklin, R. D., Gorman, B. S., Beasley, T. M., & Allison, D. B. (1996). 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.

Gage, N. A., & Lewis, T. J. (2012, May 11). Hierarchical linear modeling meta-analysis of single-subject design research. *Journal of Special Education*. Advance online publication. doi: 10.1177/0022466912443894

Ganz, J. B., Earles-Vollrath, T. L., Heath, A. K., Parker, R. I., Rispoli, M. J, & Duran, J. B. (2012). A meta-analysis of single case research studies on aided augmentative and alternative communication systems with individuals with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 42,* 60-74.

Gast, D. L. (2010a). Replication. In D. L. Gast (Ed.), *Single subject research methodology in behavioral sciences* (pp. 110-128). London, UK: Routledge.

Gast, D. L. (2010b). *Single subject research methodology in behavioral sciences*. London, UK: Routledge.

Gast, D. L., & Ledford, J. (2010). Multiple baseline and multiple probe designs. In D. L. Gast (Ed.), *Single subject research methodology in behavioral sciences* (pp. 276-328). London, UK: Routledge.

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.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York, NY: Academic Press.

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.

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

Holden, G., Bearison, D. J., Rode, D. C., Rosenberg, G., & Fishman, M. (1999). Evaluating the effects of a virtual environment (STARBRIGHT world) with hospitalized children. *Research on Social Work Practice, 9*, 365-382.

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

Howick, J., Chalmers, I., Glasziou, P., Greenhaigh, T., Heneghan, C., Liberati, A., et al. (2011). *The 2011 Oxford CEBM Evidence Table* (Introductory Document). Oxford Centre for Evidence-Based Medicine. http://www.cebm.net/index.aspx?o=5653

Huitema, B. E., & McKean, J. W. (1991). Autocorrelation estimation and inference with small samples. *Psychological Bulletin, 110,* 291-304.

Institute of Education Sciences. (2013). Request for applications: Statistical and research methodology in education. Retrieved from http://ies.ed.gov/funding/pdf/2014_84305D.pdf

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,* XXX-XXX.

Jarrett, M. A., & Ollendick, T. H. (2012, February 6). Treatment of comorbid attention-deficit/hyperactivity disorder and anxiety in children: A multiple baseline design analysis. *Journal of Consulting and Clinical Psychology*. Advance online publication. doi: 10.1037/a0027123

Jenson, W. R., Clark, E., Kircher, J. C., & Kristjansson, S. D. (2007). Statistical reform: Evidence-based practice, meta-analyses, and single subject designs. *Psychology in the Schools, 44*, 483-493.

Johnston, J. M., & Pennypacker, H. S. (2009). *Strategies and tactics of behavioral research* (3rd ed.). New York, NY: Routledge.

Jones, L. V., & Fiske, D. W. (1953). Models for testing the significance of combined results. *Psychological Bulletin, 50*, 375−382.

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. (2001). *Behavior modification in applied settings* (6th ed.). Belmont, CA: Wadsworth.

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

Kennedy, C. H. (2005). *Single-case designs for educational research*. Boston, MA: Allyn & Bacon.

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. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/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. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods, 15*, 124-144.

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

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

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

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

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.

Maggin, D. M., Swaminathan, H., Rogers, H. J., O'Keefe, B. V., Sugai, G., & Horner, R. H. (2011). A generalized least squares regression approach for computing effect sizes in single-case research Application examples. *Journal of School Psychology, 49*, 301-321.

Manolov, R., Guilera, G., & Sierra, V. (2014, February). Weighting strategies in the meta-analysis of single-case studies. *Behavior Research Methods*. Advance online publication. doi: 10.3758/s13428-013-0440-0

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

Manolov, R., & Solanas, A. (2012). Assigning and combining probabilities in single-case studies. *Psychological Methods, 17*, 495-509.

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.

McGrath, R. E., & Meyer, G. J. (2006). When effect size disagree: The case of *r* and *d*. *Psychological Methods, 11*, 386-401.

Michael, J. (1974). Statistical inference for individual organism research: Mixed blessing or curse? *Journal of Applied Behavior Analysis, 7*, 647-653.

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

O'Neill, B., & Findlay, G. (2014). Single case experimental designs in neurobehavioural rehabilitation: Preliminary findings on biofeedback in the treatment of challenging behaviour. *Neuropsychological Rehabilitation, 24*, XXX-XXX.

Parker, R. I., Brossart, D. F., Vannest, K. J., Long, J. R., Garcia De-Alba, R., Baugh, F. G., & Sullivan, J. R. (2005). Effect sizes in single case research: How large is large? *School Psychology Review, 34*, 116-132.

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). Useful effect size interpretations for single case research. *Behavior Therapy, 38*, 95-105.

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. (2012). Bottom-up analysis of single-case research designs. *Journal of Behavioral Education, 21*, 254-265.

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., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy, 42*. 284-299.

Petersen-Brown, S., Karich, A. C., & Symons, F. J. (2012). Examining estimates of effect using Non-overlap of all pairs in multiple baseline studies of academic intervention. *Journal of Behavioral Education, 21*, 203-216.

R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rasquin, S. M. C., Van De Sande, P., Praamstra, A. J., & Van Heugten, C. M. (2009). Cognitive-behavioral intervention for depression after stroke: Five single case studies on effects and feasibility. *Neuropsychological Rehabilitation, 19*, 208-222.

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.

Rindskopf, D. M. (2014). Bayesian analysis of data from single case designs. *Neuropsychological Rehabilitation, 24,* XXX-XXX.

Rogers, H. J. & Swaminathan, H. (2007). *A computer program for the statistical analysis of single subject designs*. Storrs, CT: University of Connecticut, Measurement, Evaluation, and Assessment Program.

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods, 13*, 19-30.

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.

Schlosser, R. W., & Sigafoos, J. (2008). Meta-analysis of single-subject experimental designs: Why now? *Evidence-Based Communication Assessment and Intervention, 2*, 117-119.

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., Brasil, I. C. C., Illingworth, D. A., White, K. D., Galindo, R., Nagler, E. D., & Rindskopf, D. M. (2009). Using UnGraph to extract data from image files: Verification of reliability and validity. *Behavior Research Methods, 41*, 177-183.

Shadish, W. R, Hedges, L. V, & Pustejovsky, J. E. (2013, December). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology.* Advance online publication. doi: 10.1016/j.jsp.2013.11.005

Shadish, W. R, Hedges, L. V, Pustejovsky, J. E, Boyajian, J. G., Sullivan, K. J., Andrade, A, & Barrientos, J. L. (2014). A d-statistic for single-case designs that is equivalent to the usual between-groups d-statistic. *Neuropsychological Rehabilitation, 24,* XXX-XXX.

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.

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

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. New York, NY: Basic Books.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. New York, NY: Appleton-Century.

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.

Solanas, A., Manolov, R., & Sierra, V. (2010). Lag-one autocorrelation in short series: Estimation and hypothesis testing. *Psicológica, 31*, 357-381.

Solmi, F., & Onghena, P. (2014). Combining p-values in replicated single-case experiments with multivariate outcome. *Neuropsychological Rehabilitation, 24*, XXX-XXX.

Strube, M. J. (1985). Combining and comparing significance levels form nonindependent hypothesis tests. *Psychological Bulletin, 97*, 334-341.

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*, XXX-XXX.

Swaminathan, H., Rogers, H. J., Horner, R., Sugai, G., & Smolkowski, K. (2014). Regression models for the analysis of single case designs. *Neuropsychological Rehabilitation, 24*, XXX-XXX.

Tate, R. L., Perdices, M., McDonald, S., Togher, L., & Rosenkoetter, U. (2014). The conduct and report of single-case research: Strategies to improve the quality of the neurorehabilitation literature. *Neuropsychological Rehabilitation, 24*, XXX-XXX.

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.

Tate, R. L., Togher, L., Perdices, M., McDonald, S., & Rosenkoetter, U. (2012, July). *Developing reporting guidelines for single-case experimental designs: The SCRIBE project*.

Paper presented at the 9th Conference of the Neuropsychological rehabilitation special interest group of the World federation for neurorehabilitation, Bergen, Norway.

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

Vannest, K. J., Parker, R. I., & Gonen, O. (2011). *Single Case Research: web based calculators for SCR analysis*. (Version 1.0) [Web-based application]. College Station, TX: Texas A&M University. Retrieved Friday 12th July 2013. Available from singlecaseresearch.org

Wampold, B. E., & Furlong, M. J. (1981). Randomization tests in single-subject designs: Illustrative examples. *Journal of Behavioral Assessment, 3*, 329-341.

Whitlock, M. C. (2005). Combining probability from independent tests: The weighted Z-method is superior to Fisher's approach. *Journal of Evolutionary Biology, 18*, 1368-1373.

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

White, O. R. (1972). *The split-middle: A quickie method of trend analysis*. Eugene, OR: Regional Resource Center for Handicapped Children.

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,* XXX-XXX.

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.