Psicológica (2016), 37, 209-234.

A discussion of alternatives for establishing empirical benchmarks for interpreting single-case effect sizes

Rumen Manolov[1 2], Matthew Jamieson[3 4], Jonathan J. Evans[3], & Vicenta Sierra[2]

[1] *Dept. of Behavioural Sciences Methods, University of Barcelona, Spain*

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

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

[4] *Glasgow Interactive Systems Group, School of Computing Science, University of Glasgow, Scotland, UK*

In this paper we reflect on the numerous calls for the development of benchmarks for interpreting effect size indices, reviewing several possibilities. Such benchmarks are aimed to provide criteria so that analysts can judge whether the size of the effect observed is rather "small", "medium" or "large". The context of this discussion is single-case experimental designs, for which a great variety of procedures have been proposed, with their different nature (e.g., being based on amount of overlap vs. a standardized mean difference) posing challenges to interpretation. For each of the alternatives discussed we point at their strengths and limitations. We also comment how such empirical benchmarks can be obtained, usually by methodologists, and illustrate how these benchmarks can be used by applied researchers willing to have more evidence on the magnitude of effect observed and not only whether an effect is present or not. One of the alternatives discussed is a proposal we make in the current paper. Although it has certain limitations, as all alternatives do, we consider that it is worth discussing it and the whole set of alternatives in order to advance in interpreting effect sizes, now that computing and reporting their numerical values is (or is expected to be) common practice.

---

[1] This work was partially supported by the *Agència de Gestió d'Ajust Universitaris i de Recerca de la Generalitat de Catalunya* grant 2014SGR71. 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

In the era of information technology it is relatively easy for researchers willing to publish their studies to know what journal editors require (e.g., report effect size measures in empirical studies comparing conditions or studying the relation between variables) and to get it by means of the software tools available. Effect sizes offer an objective way of summarizing the results of a study and make possible further meta-analyses, but we here focus on another issue related to their use: how can the numerical values obtained be interpreted? In the following, we present the general context in which effect sizes were endorsed, afterwards focussing specifically on the particularities of single-case experimental designs. The main part of the article discusses different sources of benchmarks that can help interpreting the effect sizes, pointing at the strengths and limitations of these approaches, as well as explaining how they can be followed.

General Analytical Context

The criticism directed toward the excessive use and the misuse of null hypothesis testing (Cohen, 1990, 1994; Lambdin, 2012; Nickerson, 2000) has been complemented by the search of alternative ways of presenting the results. Among these alternatives can be mentioned the possibility to test not only statistically significant differences, but also statistical equivalence, as well as allowing for inconclusive results (Tryon, 2001). Another proposal deals with defining a difference that is clinically significant and statistically reliable and not just different from zero (Jacobson & Truax, 1991). Additionally, overcoming one of the limitations of null hypothesis testing, there have been proposals for obtaining the probability of replicating an effect (Killeen, 2005; Lecoutre, Lecoutre, & Poitevineau, 2010). In this context, a recommendation made to researchers working in psychology has been to compute and report effect size indices (e.g., standardized mean difference, correlation, risk ratio), as well as confidence intervals around them to inform about the precision of the estimate of the effect (Wilkinson, & APA Task Force on Statistical Inference, 1999). Despite the strengths of effect size measures such as their independence from sample size and the focus on the magnitude of effect, rather than on a probability of observing the effect obtained under the null hypothesis, these measures cannot be considered a panacea as they also entail difficulties in translating the effects in a correct way and into an understandable and useful language (Cortina & Landis, 2011).

The Context of Single-Case Experimental Designs

Effect size indices are currently also a frequent element of single-case experimental designs (SCED) data analysis, alongside visual inspection, due to the influence of the movement for evidence-based practice (Jenson, Clark, Kircher, & Kristjansson, 2007) and due to the fact that they help assessing the relative strength of specific treatments (Beeson & Robey, 2006). In SCED, it is not as straightforward as in group designs to choose one universally accepted effect size measure, given that there is no consensus currently on that matter (Kratochwill et al., 2010; Smith, 2012). Moreover, it is known that the use of different effect size measures can lead to different conclusions (McGrath & Meyer, 2006; Parker et al., 2005). Thus, the interpretative benchmarks of what a large vs. small effect is should be created for each specific indicator, until there is greater agreement which techniques are more appropriate and when. The characteristics of SCED analytical techniques are relevant for discussing the applicability of the alternatives for establishing benchmarks in the SCED context.

Different Agents Involved

On the one hand, in order to make clear to whom this article is directed, we need to distinguish between *methodologists* and *applied researchers*. We use the term methodologists here to refer to those individuals who make, test, and discuss analytical proposals such as the ones leading to obtaining the effect size measures. We consider that methodologists are more likely to be the professionals interested in establishing interpretative benchmarks. In contrast, applied researchers are likely to be more interested in using the interpretative benchmarks for assessing the effects obtained in their studies (in which empirical data are gathered, for instance, comparing conditions). When discussing the alternatives, we comment on the implications for these two types of agents.

On the other hand, in the context of how the benchmarks are established, we need to distinguish between *primary authors* and *proponents*. Primary authors are the individual(s) who carry out a study, write a report, and make a statement regarding the magnitude of effect observed on the basis of whatever criteria they have used. Thus, primary authors are usually applied researchers. Proponents are the individuals that develop, adapt, extend, or suggest an analytical technique and also propose a set of benchmarks for interpreting the numerical values obtained via this technique. Thus, proponents are usually methodologists.

Alternatives for Establishing Benchmarks

Before presenting the alternatives, it is necessary to state explicitly that there have been claims that categorizing the numerical values is unnecessary and can be potentially misleading (see Kelley and Preacher, 2012, for a review). An inappropriate categorization is more likely to take place when benchmarks are used in a mechanicistic, inflexible, and universal way (Thompson, 2001), regardless of the characteristics of the data or the fact that in a certain disciplines the effects are usually larger than in others (Kelley & Preacher, 2012; Vacha-Haase & Thompson, 2004), or when not paying attention to the wider context in which the research study is set (Kazdin, 1999). If such inappropriate categorization is used, there would actually be "no wisdom" (Glass, McGaw, & Smith, 1981, p. 104) in it and the effect sizes would be more useful if they remain purely numerical, as an objective summary and as ingredients for meta-analyses comparing several interventions.

However, we still consider that benchmarks and verbal interpretations of the numerical values can be useful, as well-informed interpretations provide benefits such as reduced cognitive demands (Henson, 2006; Rosch, 1978), common language and completion of journal requirements (Kelley & Preacher, 2012). Accordingly, it is noted that "reporting and interpreting effect sizes in the context of previously reported effects is essential to good research" (Wilkinson & APA Task Force on Statistical Inference, 1999, p.599), which is the reason why we review several alternatives for aiding the interpretation of effect size indices.

In our discussion of how the alternatives can be used, we will illustrate their application with the result of two studies included in a recent meta-analysis (Jamieson, Cullen, McGee-Lennon, Brewster, & Evans, 2014). Labelle and Mihailidis (2006) perform a study with residents from a long-term-care unit having been diagnosed with dementia and presenting moderate-to-severe cognitive impairment. Audiovisual prompts are used to help the individuals in a handwashing task, measuring the number of steps completed without a caregiver (increase expected) and number of caregiver interventions (decrease expected) as dependent variables. In the Jamieson et al. (2014) meta-analysis, the Nonoverlap of all pairs (NAP; Parker & Vannest, 2009) was used as effect size measure, obtaining the value of 0.91 for the Labelle and Mihailidis (2006) data. We stick with NAP in our illustrations, as it is a commonly used and respected measure in the SCED context, which has also been shown to perform well in certain conditions (Manolov, Solanas, Sierra, & Evans, 2011). In the second study, the results of which we include in our illustrations, Chang, Chou, Wang, and Chen

(2011) use a technological device based on Microsoft Kinect for supervising the completion of the steps in a food preparation task. The participants are individuals presenting different conditions such substance abuse, dementia, and paranoid schizophrenia. The dependent variable is the success rate in the task, i.e., the number of steps completed correctly. Jamieson et al. (2014) obtained NAP = 1.00 for Chang et al.'s (2011) data.

Alternative 1: Cohen's Benchmarks

Features. Cohen's (1988; 1992) rules-of-thumb have arguably been the most frequently used interpretative guidelines, although they were also suggested in order to promote the use of power analysis before carrying out a study. This widespread use takes place despite the fact that Cohen himself suggested using empirical evidence for interpreting effect size values, when available, instead of using his criteria universally. Interestingly, Sun, Pan, and Wang (2010) stress the need for alternative ways of interpretation, but they (a) use Cohen's benchmarks when comparing conclusions based on effect sizes and *p* values conclusions and (b) provide an example of good use of effect sizes based on Cohen's benchmarks.

Methodologists: establishing benchmarks. Cohen's benchmarks are already available and the need not be set anew. However, if these benchmarks are to be used in the SCED context, methodologists may need to justify their appropriateness considering data characteristics such as likely nonnormality and serial dependence. For instance, methodologists need to explain whether the benchmarks for point-biserial R^2 (.01, .06, and .14) are also applicable, taking the square root, to the correlation used in Simulation modelling analysis (Borckardt et al., 2008) or whether the benchmarks for R^2 arising from multiple regression analysis (.02, .13, and .25, attributed to Cohen by Kotrlik, Williams, & Jabor, 2011) are applicable to Allison and Gorman's (1993) model. Actually, it could be argued that Cohen's benchmarks for standardized mean difference (0.2, 0.5, and 0.8) are applicable to the *d*-statistics developed by Hedges, Pustejovsky, and Shadish (2012; 2013), provided that these authors demonstrate that the measures are comparable to the ones obtained in group designs. Nevertheless, it has been suggested that Cohen's benchmarks are not appropriate for SCED due to the common presence of larger effects (Parker et al., 2005) so this direct application can also be questioned.

Applied researchers: using benchmarks. The steps are as follows: 1) compute an effect size for the data gathered; 2) consult Cohen's (1988) book for the corresponding type of effect size measure: in this case, in order to illustrate how Cohen's benchmarks can be used we could use the fact that

a d=2 corresponds to 81.1% of nonoverlap and so, 91% and 100% nonoverlaps, from the Labelle and Mihailidis (2006) and Chang et al. (2011) studies, are associated with d>2. However, it could be claimed that the nonoverlap entailed in the interpretation of d is not the same as the one quantified by NAP. In order to address such criticism, it should be noted that Cohen's d can also be expressed as a probability of superiority of one set of measures over another (Citrome, 2014; Lakens, 2013). In this case, the way in which NAP is computed is actually identical to the nonparametric version of the probability of superiority (Grissom & Kim, 1994). According to this conversion, a probability of superiority of 91% corresponds to d=1.9, whereas 100% would correspond to d=3.8 (Fritz, Morris, & Richler, 2012); 3) locate the value obtained in the range of values suggested: given that NAP=0.91 and NAP=1.00 would correspond to d=1.8 or 3.8, both of which are greater than 0.8, the effects could be labelled as "large".

Strengths. Cohen's benchmarks are well-known and widely used and, despite their limitations, their use is likely to be less questioned that the use of other relatively unknown sets of benchmarks.

Limitations. There have been repeated calls to find suitable alternatives (Institute of Education Sciences, 2015), emphasising the need to consider previous findings and knowledge of the area studied in order to determine what constitutes the different levels of effectiveness (Kotrlik et al., 2011). Cohen's benchmarks are not specific to single-case designs in which the characteristics of the data (e.g., repeated measurements of one or few participants) and of the analytical procedures need to be taken into account. In relation to this lack of specificity, some mathematically possible conversions (e.g., between R^2 and d) may not be justified (Shadish, Rindskopf, & Hedges, 2008). Moreover, standardized mean differences such as Cohen's d and Glass' delta (Glass et al., 1981) computed for SCEDs only reflect within-subject variability, whereas in the between groups context in which they were proposed between subjects variability is also having an impact (Beretvas & Chung, 2008; Hedges et al., 2012). Thus the values obtained are not directly comparable. Finally, no Cohen criteria are available for interpreting any of the several nonoverlap measures that have been developed later, although he did offer an interpretation of standardized mean difference in terms of overlap between *normally* distributed data.

Alternative 2: Across-Studies Comparisons

In order to reduce the arbitrariness in the interpretation of within-study effect size indices, across-studies comparisons have been deemed

relevant (Durlak, 2009; Sun et al., 2010; Vacha-Haase & Thompson, 2004). Such comparisons are meaningful when the studies inform about "effects involving the same or similar variables" (Valentine & Cooper, 2003, p.5) and when similar outcomes are measured in similar designs (Durlak, 2009). Moreover, it is important to take into account the domain to which the set of studies serving as a reference belong. In that sense, following the across studies approach for 131 school psychology SCED studies, Solomon, Howard, and Stein (2015) provided interpretative benchmarks for several SCED analytical techniques on the basis of quartiles[1], obtaining Tau-U quartiles/benchmarks that are 0.2 lower than the quartiles reported by Parker et al. (2011) from their sample of SCED studies (including school psychology, special education, and behavioral psychology). This difference is apparently substantial, considering that (absolute) Tau-U ranges from 0 to 1. Moreover, the discrepancy observed illustrates the need to establish benchmarks for each specific discipline and the need to go beyond convenience samples (as both samples were in these reviews).

Methodologists: establishing benchmarks. An approach for across-studies comparisons consists in: 1) sorting the distribution of effect sizes in ascending order and 2) dividing this distribution into portions with the same size. Regarding the latter point, Hemphill (2003) provides three portions (lower, middle, and upper third) for correlation coefficients, Parker and Vannest (2009) and Parker, Vannest, Davis, and Sauber (2011) provide four portions according to the quartiles, as well as additional indicators of position such as percentiles 10 and 90, for NAP and Tau-U, respectively. Another option is to compute the frequencies of several ordered and equally-spaced intervals of effect size values (.00–.09, .10–.19, …, .90–1.00), as Haase, Wachter, and Solomon (1982) did for η^2. This option informs that an effect in a given interval is higher than a certain percentage of the effects observed.

Applied researchers: using benchmarks. The steps are as follows: 1) compute an effect size for the data gathered: NAP = 0.91 for the Labelle and Mihailidis (2006) data and NAP=1.00 for Chang et al.'s (2011) data; 2) consult the set of criteria established: according to Parker and Vannest's (2009) field test the values that correspond to percentiles 10, 25, 50, 75, and 90 are respectively 0.50, 0.69, 0.84, 0.98, and 1.00; 3) locate the value

[1] Note that Solomon et al. (2015) removed from the dataset used for the benchmarks data with autocorrelation above |0.4| and with high levels of nonnormality. Moreover, the set of studies was not the same for establishing the benchmarks of the different procedures as some techniques assumed lack of autocorrelation and others assumed normality. Such characteristics of the set of studies need to be kept in mind when performing comparisons.

obtained in the range of values suggested: 0.91 is greater than percentile 50 but smaller than percentile 75, whereas 1.00 is equal to percentile 90; 4) label the effect: NAP=0.91 could be labelled as "small" effect and NAP=1.00 as a "large" effect according to the across-studies comparison.

Strengths. Identifying key positions in the ordered empirical distribution of effects (e.g., percentiles) makes possible assessing whether the effect obtained in a particular study is among the smaller, intermediate, or larger ones obtained previously. This type of assessment is similar to the one performed for personality traits, such as the Big Five (Costa & McCrae, 1992): one is judged to be more or less extroverted, agreeable, etc. according to where the outcomes obtained by others.

Limitations. This alternative allows assessing the magnitude of an effect size value in relative, but not absolute terms (Haase et al., 1982; see also corollary 2 of the effect size definition by Kelley and Preacher, 2012). In that sense, an effect observed may be clinically relevant, but smaller than the ones observed in other studies due to a variety of reasons, which would make it look numerically less salient, leading to the label of "small effect". Conversely, the fact that an effect is one of the largest observed for an effect size measure does not guarantee that it has practical significance.

Alternative 3: Benchmarks Proposed by Each Proponent of an Analytical Technique

Features. When an analytical technique is developed or adapted for SCED data analysis, in certain cases, it is possible that its proponents suggest interpretative benchmarks. This is the case for the Percentage of nonoverlapping data (PND; Scruggs, Mastropieri, & Casto, 1987), the NAP (Parker & Vannest, 2009) and Improvement rate difference (IRD; Parker, Vannest, & Brown, 2009). For NAP and IRD, Alternative 4 described below was followed. However, for PND it was not explicitly stated where the benchmarks come from (e.g., "scores of 70 to 90 have been considered effectives"; Scruggs & Mastropieri 1998, p.224). We consider that the case of PND is important, given that it is arguably the most frequently used SCED analytical technique (Schlosser, Lee, & Wendt, 2008; Scruggs & Mastropieri, 2013), despite its potential flaws (Allison & Gorman, 1994). Specifically, Scruggs and Mastropieri (1998) suggest that a PND in the range 50–70 indicates small or questionable effectiveness, 70–90 an effective intervention, and values greater than 90 a very effective one.

Methodologists: establishing benchmarks. We do not envision any specific procedure for establishing benchmarks that is different from the ones discussed in Alternatives 1, 2, 4, 5, and 6. A proposal must be based

on at least one of these fundaments, although methodologists and applied researchers could consider some of these alternatives more appropriate than others.

Applied researchers: using benchmarks. In this section we describe an incorrect way of using the PND benchmarks for assessing the magnitude of an effect quantified via NAP. If Scruggs and Mastropieri's (1998) criteria were used to assess how large is NAP=0.91 and NAP=1.00, both values would have been labelled as large, given that they are greater than 90%. We do not recommend this practice, but we illustrate it here, for two reasons. First, a similar incorrect use can be seen in Ma (2010) using the criteria suggested by Scruggs & Mastropieri (1998) for PND when he interprets the values of the Percentage of data points exceeding the median (PEM; Ma, 2006). However, such a practice is not justified, as Parker and Vannest (2009) and Manolov, Solanas, and Leiva (2010) show that the values obtained for PEM are larger than the ones yielded by PND, which is expected given that the former compares intervention phase measurements with a criterion that is usually lower (the median vs. the maximum value). Second, such practice parallels the type of mistake committed when using Cohen's benchmarks for the between-groups *d* when interpreting a *d* value computed from single-case data, as in both cases the similarity is only superficial as standardization does not take place in the same way.

Strengths. If we assume that the proponents of a technique are the individuals that best know its features, it would be logical to follow their guidelines. Moreover, this would add consistency to the interpretation and avoid that researchers use their own (equally arbitrary) criteria.

Limitations. All interpretative benchmarks need some justification and consistent interpretation is not sufficient, if the labels stemming from such benchmarks are to be used by policy makers for deciding which interventions to endorse and finance.

Alternative 4: Judging the Magnitude of Effect through Visual Analysis

The first three alternatives are equally applicable to designs comparing groups, correlational designs and SCEDs. This fourth alternative is more specific to SCEDs, given the importance of visual analysis in the field (Parker, Cryer, & Byrns, 2006). This actually is the procedure followed by Parker and colleagues for NAP and IRD. This alternative has also been followed in a meta-analysis using the NAP (Petersen-Brown, Karich, & Symons, 2012).

Methodologists: establishing benchmarks. Parker and Vannest (2009) do not make explicit the steps followed ("based on expert visual judgments of these 200 datasets, we can offer very tentative NAP ranges", p.364), but Petersen-Brown et al. (2012) do explain their procedure and thus we focus on the latter. Independent observers judge the set of graphs, using structured criteria such as the ones provided by Kratochwill et al. (2010), focussing on changes in level, trend, variability and the immediacy of effects: they decide for each two-phase comparison whether there is an effect (at least two of the four criteria are met) or not. In a multiple baseline design, in case an effect is judged to be present for 75% or more of the baselines, it is labelled as "large" vs. "small" for 50% to 75% of the baselines. This procedure could be extended to any design that entails replication (e.g., ABAB, alternating treatments), as is required from all SCEDs to be considered experimental (Kratochwill et al., 2010). Afterwards, the authors used the ROC curve to identify the NAP values that correspond to large (NAP=.96) and small effect (NAP=.93) as judged by visual analysis.

Applied researchers: using benchmarks. The steps are as follows: 1) compute an effect size for the data gathered: NAP = 0.91 for the Labelle and Mihailidis (2006) data and NAP=1.00 for Chang et al.'s (2011) data; 2) consult the set of criteria established: Parker and Vannest's (2009) visual analysis led to the following benchmarks: 0–.65 (small), .66–.92 (medium), and .93–1.00 (large); 3) locate the value obtained in the range of values suggested: 0.91 is in the range suggesting medium effect, whereas 1.00 can be labelled as a "large" effect. If we used the criteria by Petersen-Brown and colleagues (2012), NAP=.91 would be less than a "small" effect and NAP=1.00 would be a large one.

Strengths. Visual analysis allows taking into consideration several aspects of the data, such as the type of effect (abrupt vs. progressive, immediate vs. delayed, sustained vs. temporary change), as well as the amount of variability and presence of outliers that can influence the quantifications. Moreover, SCED researchers are used to performing this kind of analysis (Parker & Brossart, 2003) and some even use it as gold standard (Petersen-Brown et al., 2012; Wolery, Busick, Reichow, & Barton, 2010).

Limitations. The main drawback of this approach is the evidence that visual analysts do not show sufficient agreement when analysing the same data (e.g., Danov & Symons, 2008; DeProspero & Cohen, 1979; Ninci, Vannest, Willson, & Zhang, 2015; Ottenbacher, 1993), leading to potentially unreliable results. Thus, it is likely that different analysts reach

different conclusions about the same data. Another potential drawback is closely related to a concern expressed regarding making decisions only on the basis of the graphical display of the data and not taking into account the complete context in which these data where gathered (Brossart, Parker, Olson, & Mahadevan, 2006). Thus, if only the visual representation of the data is used for categorizing the magnitude of effect, the useful information about the clients, the type of behaviours treated, and the context of the intervention is not taken into account. This in-depth knowledge is one of the strengths of SCED and should not be omitted from the assessment of intervention effectiveness. Moreover, unless all methodologists state explicitly the steps of the process followed (no such information is provided by Parker and Vannest, 2009, and Parker et al., 2009), it would not be transparent enough. Regarding the procedure followed by Petersen-Brown et al. (2012) it makes clear that the distinction between small vs. large effect is not close to being perfect, as the NAP values identified show a specificity of .81-.83 and a sensitivity of .73-.78.

Alternative 5: Objective Clinical Criteria

Ferguson (2009) is an author proposing interpretive guidelines that differ from Cohen's, but also underlining that these values do not necessarily imply practical significance. In that sense, potentially more useful benchmarks would benefit from well-accepted clinical cut-off scores (Durlak, 2009), such as ones available for the Beck Depression Inventory (Beck, Steer, & Brown, 1996) or the Montreal Cognitive Assessment (Nasreddine et al., 2005). Another option for interpretation in practical terms is quantifying the gains observed in national normative samples (Hill, Bloom, Black, & Lipsey, 2008). An example of this latter option is the comparison, in an educational context, of the effect of an intervention to a typical year of natural growth in students. Another option proposed by Hill and colleagues (2008) is to compare the effect of an intervention to the existing differences among subgroups of students.

Methodologists: establishing benchmarks. Establishing benchmarks in such a way requires an intensive process of data collection from large representative samples in order to gather the normative date or to construct a solid basis for the correlation between test scores and real-life functioning in the domain tested. Once such large-scale data are available, the raw scores (or the summary measure) obtained by the participant(s) in a SCED can be compared to the cut-off points. The effect size could be compared to the gaps quantified between subgroups based on age or other relevant demographic characteristics.

Applied researchers: using benchmarks. The steps are as follows: 1) compute a quantification for which there are clinically relevant criteria available: for the Labelle and Mihailidis (2006) data, the mean number of caregiver interventions provided as target behaviour can be the focus, instead of the NAP value. For Chang et al.'s (2011) study, the success rate computed in each session is a potentially meaningful measure. The use of raw measurements, as we do here, has been deemed useful for boosting interpretation whereas standardized measures (and also percentages and proportions such as NAP) enable comparability across studies (Cumming, 2012). 2) Consult the set of criteria established: Labelle and Mihailidis (2006) suggests the following categories: constant cuing—an average of more than 10 caregiver interventions per trial per phase; minimal cuing—an average of 5–10 caregiver interventions per trial per phase; occasional cuing—an average of fewer than 5 caregiver interventions per trial per phase. For Chang et al.'s (2011) an errorless completion of the task can be considered a clinically relevant result. 3) locate the value obtained in the range of values suggested: for the Labelle and Mihailidis (2006) study, the mean number of caregiver interventions ranges between 3.6 and 7.5 for the treatment phase, with an average of 5.3 and a median of 5.1. Thus, these values suggest that minimal cuing is required – a good but not optimal result. For Chang et al.'s (2011) data, a 100% success rate is achieved in the intervention phases, indicating a clinically relevant improvement, as the values were below 40% for one of the participants and around 60% for the other participant in absence of intervention.

Strengths. This is arguably *the* alternative to be used, if possible, as it relates the measurements and numerical summaries obtained in a study to criteria that speak the language of practical importance, without necessarily hiding behind statistical significance or the apparent magnitude of a numerical value.

Limitations. Cut-off scores and normative data are not available in all disciplines and research areas and definitely not for all outcomes (e.g., frequency or rate of a specific behaviour). This is more so in the SCED context, as SCEDs are sometimes used to treat rare situations and thus large-sample data are not likely to be obtained. Moreover, there might not be such a criterion as 100% achievement or elimination (0%) of an undesired behaviour for all types of raw measurements.

Alternative 6 (The Proposal): Using Primary Authors' Effectiveness Categories

Features. In order to have benchmarks that allow interpreting effect sizes in absolute terms (unlike Alternative 2), the current proposal does not require sorting the effect sizes before judging their potential importance. In contrast with Alternative 4, we advocate for relying on the primary authors' judgement when assigning a label to the effect observed in a study and not only on the graphical display of the data. We consider that each researcher carrying out and reporting a study is best-suited and responsible for the within-study interpretation of the effect. This is so, given that the knowledge that this researcher or practitioner has of the individual, the problematic behaviour, and the context is greater than the knowledge of an external reader focussing on the graphed data only. Moreover, taking into account the primary authors' explicitly expressed opinion on whether the effect is considered to be practically important or not, allows examining the possibility that a small effect size can be a substantively relevant result (Kelley & Preacher, 2012). Accordingly, Vannest and Ninci (2015) underline that the interpretation of effect sizes should be related to practical significance and not focus only on the apparent size of the numerical value. These authors state that "an ES is not small, medium, or large by itself, and should be described in relationship to a client's needs, goals, and history, as well as the intervention and setting used for the client" (p. 408). Our approach, unlike across-studies comparisons of effect sizes, is well-aligned with this idea, given that the aspects to be considered in the interpretation are best known by primary researchers.

Methodologists: establishing benchmarks. The steps are as follows: 1) identify a set of studies dealing with the same type of problem addressed (e.g., brain injury, developmental disabilities, education) or to the same type of intervention (e.g., behaviour modification interventions, cognitive therapy, technology-based interventions)[2]; 2) read thoroughly the articles and the primary authors' descriptions and evaluative comments on intervention effectiveness and code them into categories (e.g., negative impact, no improvement, unclear or small or weak improvement, improvement, moderate improvement, substantial improvement); 3) retrieve the data from the graphical and/or tabular information available in the report and compute one or several effect size measures (an option is to stick with

[2] It has been suggested that a different set of criteria is necessary for different areas (Hemphill, 2003) and any generalizations across areas are potentially oversimplifications and possibly flawed.

the effect size reported in the study, without having to retrieve raw data); 4) compute the effect size of interest for all data sets; 5) group the values of the effect size according to the label assigned in step 2; 6) for each of the effectiveness categories, compute the key percentiles; 5) present the key percentiles representing the effect sizes values that has been labelled as "negative impact"; 6) repeat step 5 for all subsequent labels (no improvement, unclear or small improvement, improvement, moderate improvement, substantial improvement).

Applied researchers: using benchmarks. The steps are as follows: 1) compute an effect size for the data gathered: NAP=0.91 for Labelle and Mihailidis (2006) and NAP=1.00 for Chang et al. (2011); 2) consult the effectiveness categories: we performed a preliminary study[3], following the steps for methodologists presented above, on all 38 participants from the SCED studies included in Jamieson et al.'s (2014) meta-analysis, using the NAP values they computed, and we obtained the following values for percentiles 25, 50, and 75 for the different effectiveness categories: no effect (.50, .59, .67), small effect (.75, .81, .87), moderate effect (.81, .97, 1.00), large effect (.84, .91, 1.00). Although these values cannot be considered definitive, they are useful for the current illustration; 3) locate the value in the effectiveness categories: the values presented in the previous step reflect one of the potential limitations of this alternative: the overlap between two of the categories. A NAP value such as 0.91 has been found to represent a moderate effect in some studies and a large effect in others; the case for NAP=1.00 is similar.

Strengths. First, this alternative allows respecting the decisions about magnitude made from the inside, by the primary authors, instead of imposing an external criterion on the basis of a limited amount of information about actual improvement (e.g., only a visual display or a numerical summary of the results). Ideally, primary authors would take into account their knowledge of client and context and the amount of change in everyday life functioning, apart from assessing the visual and quantitative summary of the data. Second, the information obtained is different from the one that stems from across-studies comparisons, assessing magnitude in absolute rather than relative terms. Third, establishing benchmarks according to Alternative 6 can be efficient, as it can be done in the context of a meta-analysis – the empirical distributions for the different effectiveness categories can be obtained for the same set of studies for which a relevant research question is being answered by means of a weighted average, heterogeneity test, moderator analysis, etc. Moreover,

[3] More details are available from the authors upon request.

meta-analysis sometimes include studies from grey literature (e.g., not published in peer-review journals), which helps reducing the possibility of publication bias overestimating the interventions effects.

Limitations. We present our proposal here as an alternative and not as a definitive solution. Several limitations need to be noted. First, the reasoning process that primary researchers follow is not always transparent, that is, it may not be clear which exactly the basis of a judgment (or a label) is. For instance, if a primary researcher calls an effect "large" after implicitly using Cohen's benchmarks, then actually Alternative 1 is being used, whereas if such a label is based on the fact that the difference between conditions is visually clear, then Alternative 4 is being followed. Nevertheless, despite the fact that multiple criteria plus client and context information may be not always be used when establishing effectiveness categories, at least combination of visual and statistical tools is apparently common among SCED researchers (Perdices & Tate, 2010). Second, it is possible that not all authors provide enough information (especially adjectives) regarding their assessment of the degree of improvement, making difficult a fine distinction between magnitudes of effect. Moreover, the methodologists would have the task to identify synonyms among these adjectives (e.g., that substantial, large, and marked improvement all refer to the same underlying idea). Third, it is possible that procedure leads to overlap between the empirical distributions of values for the different effectiveness categories making difficult to decide which of two labels would be more appropriate.

A Remark to Methodologists on Interpretative Benchmarks and Single-Case Designs

Parker and colleagues (Parker & Vannest, 2009; Parker et al., 2009) explicitly state that their interpretative benchmarks are based on assessing AB graphs. This gives rise to the following question: how should a methodologist willing to follow Alternative 6 proceed, if the primary authors assign a label to the effect observed in the whole study (i.e., with the common replications within and across participants; Shadish & Sullivan, 2011) and not to the difference observed in a comparison between only two phases? This question can also be raised regarding Alternative 4: if visual analysis of an ABAB or a multiple-baseline design suggests that the data show a "large" effect, how is a single quantification for such a design obtained? Following common meta-analytical practice for dealing with the dependence of outcomes obtained in a specific study, the possibilities for having a single quantification per study (if there are several

multiple-baseline designs [Boman, Bartfai, Borell, Tham, & Hemmingsson, 2010] or several ABAB designs [e.g., Coker, Lebkicher, Harris, & Snape, 2009]) are: obtaining the average of the effect sizes in a study or picking one of those at random or due to a substantive reason (Borenstein, Hedges, Higgins, & Rothstein, 2009). If the label (provided by primary authors [Alternative 6] or by visual analysts [Alternative 4]) refers to the whole study with all data gathered, picking only one AB-comparison or even only one ABAB data does not seem justified. Regarding the way in which a single quantification is obtained per design, there have been several options proposed and followed: averaging using the mean or median (e.g., Kokina & Kern, 2010; Maggin et al., 2011), using only the first AB comparison (e.g., Parker et al., 2011; Strain, Kohler, & Gresham, 1998), or comparing the initial baseline to the final intervention phase (e.g., Heinicke & Carr, 2014; Olive & Smith, 2005). As before, we consider that if a label is put on the basis of all data (e.g., for an ABAB design), averaging is justified to a greater degree than picking only part of the data. Accordingly, regarding Alternative 2 methodologists should also keep in minc how the effect from a study is computed, before including it in the sorted empirical distribution, and before identifying the key percentiles. That is, they should be consistent in the rules the follow so that all values represent the same type of data (e.g., averages or comparisons of initial A to final B phase).

A Remark to Applied Researchers on Interpretative Benchmarks and Single-Case Designs

Interpretative benchmarks such as the ones proposed by Cohen for the standardized mean difference or for the point biserial correlation, the proposals of Parker and colleagues for their nonoverlap indices NAP and IRD, and the criteria by Scruggs and Mastropieri (1998) for PND refer to and/or were obtained from a comparison between a pair of conditions[4] (usually, control/baseline vs. treatment). Thus, a label indicating the size of the difference (or the degree of intervention effectiveness) can be obtained for each AB-comparison. However, an appropriate SCED should entail at least three replications of the AB sequence to allow establishing causal effects (Kratochwill et al., 2010; Tate et al., 2013). Therefore, as there would be several labels per design and several labels per study, how can

[4] There are of course interpretative benchmarks for omnibus comparisons (η^2) and for relationships between multiple variables (an R^2 based on multiple regression), but the former is not directly applicable to $SCED_S$, whereas the latter can actually be obtained from a regression model for an AB design, such as the one proposed by Allison ad Gorman (1993).

such labels be combined to represent the effect observed in the whole study? One option is to use a criterion that parallels the one employed by the WhatWorks Clearinghouse Standards (Kratochwill et al., 2010) when assessing the degree of evidence of an effect: at least three demonstrations of effect for "strong evidence", whereas a "moderate evidence" allows for one or more demonstrations of no-effect, still accompanied by three demonstrations of effect. Thus, a relatively stringent criterion can be the one used by Petersen-Brown et al. (2012) for the effects detected by visual analysts – if at least 75% of the replications suggest "large" effect, the global effect is judged to be "large" (if there are less than four replications a 100% is required); otherwise, the assessment continues with the lower categories as follows. If at least 75% of the replications suggest "medium" effect, this is the label attached to the global effect; idem with "small". A mix of labels such as 50% of the replications showing "small" effect and 50% "medium" are to be assessed conservatively (i.e., as global "small" effect in this case). This remark is applicable to all Alternatives discussed above.

Discussion

The field of SCED methodology and data analysis has received a lot of attention in recent years, as demonstrated by the re-edition of key textbooks (Barlow, Nock, & Hersen, 2009; Gast & Ledford, 2014; Kazdin, 2011; Kratochwill & Levin, 2014) and by the amount of Special Issues dedicated to the topic: Journal of Behavioral Education (Volume 21, Issue 3) in 2012, the 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 (Volume 52, Issue 2) in 2014, Neuropsychological Rehabilitation (Volume 24, Issues 3-4) in 2014, Journal of Counseling and Development (Volume 93, Issue 4) in 2015 and an upcoming special issue in Developmental Neurorehabilitation. Nevertheless, one of the topics that have not received sufficient attention is how to interpret the numerical values yielded by the variety of analytical techniques proposed. Although the effect size indices are useful as objective numerical summaries both for communicating results and for making possible meta-analysis, it would also be valuable to be able to assess, on a solid basis, the degree of effectiveness on the basis of the these numerical summaries. The current paper fills this gap by reviewing several alternatives for establishing empirical benchmarks that help interpreting effect size measures and making a new proposal. The present work also reflects the need for looking for alternatives to Cohen's benchmarks (Institute of

Education Sciences, 2015; Sun et al., 2010) as it is intended to fuel the discussion on the topic by means of the review of alternatives and the new proposal.

Recommendations and Implications for Applied Researchers

When writing a report, applied researchers have already available a set of recommendations in relation to what analyses to perform and how to report results (Wilkinson & APA Task Force on Statistical Inference, 1999). Additionally, in different areas of research there are also more general guidelines about the content of the whole report, not only the data analysis section (Moher et al., 2010; Shamseer et al., 2015; and in the near future for SCED, Tate et al., 2012). In the current paper we make two further recommendations: (a) that applied researchers clearly state the degree of intervention effectiveness (or the magnitude of difference between conditions) using one of the commonly employed adjectives (small/unclear, medium/moderate, large/strong) or to state why they prefer not to use any adjective; and (b) to clearly state why they used the adjective as a label: on the basis of maintenance of the effect, generalization to other settings, normal functioning in everyday life, client satisfaction or well-being, a comparison to a cut-off point or a normative sample, magnitude of the effect size measure, the characteristics of the graphically displayed data, etc., or they are using a set of previously proposed benchmarks. When using benchmarks, applied researchers should be aware of the meaning of the label, for instance, that the effect is expected to be greater than 75% of effects reported in the research domain or for this type of design (Alternative 2) or that the effect is similar to the ones that visual analysts tend to refer to as "large" (Alternative 4). Such precision in the use of the terms would help avoiding confusions and misunderstandings and it would also make possible establishing benchmarks on the basis of the current proposal (Alternative 6).

In case an applied researcher decides that there is no evidence supporting the use of a specific label beyond the distinction "effective" or "not effective" or no need for using such labels, we recommend taking into account and reporting two types of quantifications. On the one hand, raw measures such as kilograms of weight loss, percentage of time exhibiting on-task behaviour (Cumming, 2012; Valentine & Cooper, 2003) can help evaluating the clinical or practical significance of the behavioural change. On the other hand, standardized measures can also be used to allow comparisons across studies, meta-analytical integrations, and sample size planning for future studies (Kelley & Preacher, 2012).

Focussing specifically on applied researchers working in the SCED field, the current paper is intended to inform that interpretative benchmarks have been proposed for several nonoverlap indices (including the promising NAP and IRD), but that the bases of these benchmarks are different in nature (especially considering the currently criticized PND). Moreover, following Alternative 2, it is possible to interpret the values of another nonoverlap measure, Tau. Another warning made here is regarding the use of Cohen's benchmarks for interpreting *d*-statistics computed from SCED data. It can only be argued that the proposals made by Hedges et al. (2012, 2013) and potentially Pustejovsky, Hedges, and Shadish (2014) are directly comparable to between-group designs *d*-statistics, although larger effects may still be expected in the SCED context due to the use of individually tailored interventions. In contrast, direct applications of classical standardized mean difference indices are more problematic (Beretvas & Chung, 2008). Actually, even the *d*-statistic obtained from generalized least squares regression analysis (Maggin et al., 2011) cannot be considered to be comparable to between-groups *d* and thus Cohen's benchmarks would still be inappropriate. Finally, applied researchers should keep in mind that the size of an effect is not the only relevant aspect when assessing intervention effectiveness, as an appraisal of the methodological quality of the study is also necessary (Tate et al., 2013).

Recommendations and Implications for Methodologists

Methodologists carrying out quantitative integrations of studies are encouraged to dedicate extra effort in order for their meta-analyses to present not only a global summary of the effect of interest, but also to help establishing empirical benchmarks. Reporting the key percentiles of the ordered distribution of effect sizes in a specific area of research (Alternative 2) can be useful as a means of across-studies comparison. Moreover, paying attention to the description of the effects made by the primary authors can help identifying the effect size values associated with strong interventions (Alternative 6). In that way, Cohen's benchmarks can be used only when appropriate for group designs or correlational studies and for power analysis instead of being universally applied across designs and domains.

A recommendation to methodologists proposing analytical techniques is to describe explicitly the process followed and the justification of the benchmarks they suggest (if any). Taking into account the possibility that such benchmarks may become routinely and automatically applied, methodologists should make sure that their basis is strong.

Limitations and Future Research

It is noteworthy that the current paper provides only a discussion of potential strengths and limitations of the different ways of establishing empirical benchmarks for SCED effect sizes, but that we did not carry out a formal comparison via a field test with real data or by simulation. Moreover, the pros and cons provided here could be an object of debate from other researchers, but we consider such discussion on the topic is necessary if consistent and solid interpretation of effect sizes is to be achieved. Future research could focus not only on such formal comparison between alternatives for establishing interpretative benchmarks, but also on studying the effect size reporting and interpreting practices of SCED research, as done by Sun et al. (2010) in a more general context. Dealing with this topic both academically/methodologically and in relation to actually published reports could help bringing both worlds together.

RESUMEN

Discusión de diferentes maneras de establecer criterios interpretativos empíricos para tamaños del efecto de diseños de caso único. El presente trabajo responde a la necesidad expresada de desarrollar criterios interpretativos para los índices de tamaño del efecto, repasando diferentes maneras para conseguirlo. El objetivo de los criterios es proporcionar herramientas a los analistas para que éstos puedan valorar si el efecto observado en su estudio es más bien "pequeño", "mediano" o "grande". El contexto en el cual tiene lugar la discusión son los diseños de caso único, para los cuales se ha propuesto una gran variedad de técnicas analíticas cuya base diferente (e.g., grado de solapamiento versus diferencia de medias estandarizada) supone un reto para la interpretación. Para cada una de las alternativas que se comentan, se destacan las ventajas e inconvenientes. Adicionalmente, se comenta cómo estos criterios pueden ser obtenidos, una tarea propia de los metodólogos, y cómo pueden ser utilizados por investigadores aplicados que desean disponer de más evidencias sobre la magnitud del efecto observado, más allá de decidir si el efecto existe o no. Una de las alternativas es una propuesta que se realiza en el marco del presente artículo. A pesar de que también presenta desventajas, como todas las alternativas, consideramos que es necesario discutir esta alternativa y todas las demás con la finalidad de avanzar en la interpretación de tamaños del efecto, en un momento en el cual el hecho de calcular y reportar sus valores numéricos es (o se supone) habitual.

REFERENCES

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

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

Beck, A. T., Steer, R. A., & Brown, G. K. (1996*). Manual for the Beck Depression Inventory-II*. San Antonio, TX: Psychological Corporation.

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

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.

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.

Citrome, L. (2014). Quantifying clinical relevance. *Innovations in Clinical Neuroscience, 11*, 26-30.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

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

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

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.

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

Costa, P.T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEOFFI) professional manual*. Odessa, FL: Psychological Assessment Resources, Inc.

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

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.

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analyses of intersubject data. *Journal of Applied Behavior Analysis, 12*, 573-579.

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology, 34*, 917-928.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40*, 532-538.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141*, 2-18.

Gast, D. L., & Ledford, J. R. (2014). *Single subject research methodology: Applications in special education and behavioral sciences* (2nd ed.). London, UK: Routledge.

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

Grissom, R. J., & Kim, J. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology, 79*, 314-316.

Haase, R. F., Wachter, D. M., & Solomon, G. S. (1982). How significant is a significant difference? Average effect size of research in counseling psychology. *Journal of Counseling Psychology, 20*, 58-65.

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.

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist, 58*, 78-80.

Henson, R. K. (2006). Effect-size measures and meta-analytic thinking in counseling psychology research. *The Counseling Psychologist, 34*, 601-629.

Hill, C. J., Bloom, H. S., Black, A. R., Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*, 172-177.

Institute of Education Sciences. (2015). *Request for applications: Statistical and research methodology in education*. Retrieved October 16, 2015 from https://ies.ed.gov/funding/pdf/2016_84305D.pdf

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12-19.

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.

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.

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

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

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17*, 137-152.

Killeen, P. R. (2005). An alternative to null hypothesis statistical tests. *Psychological Science, 16*, 345-353.

Kokina, A., & Kern, L. (2010). Social Story™ interventions for students with autism spectrum disorders: A meta-analysis. *Journal of Autism and Developmental Disorders, 40*, 812-826.

Kotrlik, J. W., Williams, H. A., & Jabor, M. K. (2011). Reporting and interpreting effect size in quantitative agricultural education research. *Journal of Agricultural Education, 52*, 132-142.

Kratochwill, T. R., Hitchcock, J., 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., & Levin, J. R. (Eds.). (2014). *Single-case research designs and analysis: New directions for psychology and education* (2nd ed.). New York, NY: Routledge.

Labelle, K.-L., & Mihailidis, A. (2006). The use of automated prompting to facilitate handwashing in persons with dementia. *American Journal of Occupational Therapy, 60*, 442-450.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, art. 863.

Lambdin, C. (2012). Significance tests as sorcery: Science is empirical—significance tests are not. *Theory and Psychology, 22*, 67-90.

Lecoutre, B., Lecoutre, M.-P., & Poitevineau, J. (2010). Killeen's probability of replication and predictive probabilities: How to compute, use, and interpret them. *Psychological Methods, 15*, 158-171.

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.

Ma, H. H. (2010). Comparison of the relative effectiveness of different kinds of reinforcers: A PEM Approach. *The Behavior Analyst Today, 10*, 398-427.

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., & 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.

Manolov, R., Solanas, A., & Leiva, D. (2010). Comparing "visual" effect size indices for single-case designs. *Methodology, 6*, 49-58.

Manolov, R., Solanas, A., Sierra, V., & Evans, J. J. (2011). Choosing among techniques for quantifying single-case intervention effectiveness. *Behavior Therapy, 42*, 533-545.

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

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gotzsche, P. C., Devereaux, P. J.,… Altman, D. G. (2010). CONSORT 2010 explanation and elaboration: Update guidelines for reporting parallel group randomised trials. *BMJ, 340*, c869.

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., …, Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society, 53*, 695-699.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241-301.

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

Perdices, M., & Tate, R. L. (2010). Single-subject designs as a tool for evidence-based clinical practice: Are they unrecognised and undervalued? *Neuropsychological Rehabilitation, 19*, 904-927.

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.

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.

Rosch, E. (1978). Principals of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27-48). Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

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. (1998). Summarizing single-subject research: Issues and applications. *Behavior Modification, 22*, 221-242.

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

Shamseer, L., Sampson, M., Bukutu, C., Nikles, J., Tater, R., Jonston, B. C., … and the CENT Group. (2015). CONSORT extension for N-of-1 trials (CENT): Explanation and elaboration. *BMJ, 350*, h1753.

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

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.

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.

Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology, 102*, 989-1004.

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., on behalf of the SCRIBE Steering Committee (2012). Developing reporting guidelines for single-case experimental designs: The SCRIBE project. Paper presented at the 9th annual conference of the Special Interest Group in Neuropsychological Rehabilitation of the World Federation of NeuroRehabilitation, Bergen, Norway; abstract in *Brain Impairment, 13*(1), 135.

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education, 70*, 80-93.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6*, 371-386.

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology, 51*, 473-481.

Valentine, J. C. & Cooper, H. (2003). *Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes*. Washington, DC: What Works Clearinghouse.

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

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.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

(Manuscript received: 12 May 2015; accepted: 16 October 2015)