Using Single-Case Experiments to Support Evidence-Based Decisions: How Much Is Enough?

Marc J. Lanovaz

Université de Montréal

John T. Rapp

Auburn University

Author Note

Marc J. Lanovaz, École de Psychoéducation, Université de Montréal; John T. Rapp, Department of Psychology, Auburn University.

Correspondence concerning this article should be addressed to Marc J. Lanovaz,
École de Psychoéducation, Université de Montréal, C.P. 6128, succ. Centre-Ville,
Montreal, QC, Canada, H3C 3J7.

Email: marc.lanovaz@umontreal.ca

The final, definitive version of this paper has been published in *Behavior Modification*, 45/3, May 2017 published by SAGE Publishing, All rights reserved.

https://doi.org/10.1177/0305735616657408

Abstract

For practitioners, the use of single-case experimental designs (SCEDs) in the research literature raises an important question: How many single-case experiments are enough to have sufficient confidence that an intervention will be effective with an individual from a given population? Although standards have been proposed to address this question, current guidelines do not appear to be strongly grounded in theory or empirical research. The purpose of our paper is to address this issue by presenting guidelines to facilitate evidence-based decisions by adopting a simple statistical approach to quantify the support for interventions that have been validated using SCEDs. Specifically, we propose the use of success rates as a supplement to support evidence-based decisions. The proposed methodology allows practitioners to aggregate the results from single-case experiments in order to estimate the probability that a given intervention will produce a successful outcome. We also discuss considerations and limitations associated with this approach.

Keywords: empirically supported treatments, evidence-based practice, external validity, replication, single-case experimental designs

Using Single-Case Experiments to Support Evidence-Based Decisions: How Much Is Enough?

Ever since its formal definition in the field of medicine by Sackett, Rosenberg, Gray, Haynes, and Richardson (1996), the concept of evidence-based practice has gradually become one of the dominant issues in the fields of psychology and education in the 21st century (American Psychological Association [APA], 2005; Chwalisz, 2003; Kazdin, 2008; Slavin, 2002). Evidence-based practice aims to promote the use of effective and empirically supported interventions by practitioners (e.g., psychologists, educators, behavior analysts, allied health professionals) that best meet the needs of individuals receiving their services (e.g., clients, students, patients). To meet this objective, practitioners must integrate research evidence, their expertise, and the person's personal values to provide the most effective intervention possible (Sackett et al., 1996; Spring, 2007). Even though there has been some debate as to whether expertise and personal values should be integrated within its definition, practitioners, researchers, and professional associations generally agree that evidence-based practices should be based on empirically supported assessments and interventions (Spring et al., 2005).

The use of evidence-based practices thus leads to an important question for the fields of psychology and education: How can a practitioner reliably identify empirically supported interventions? To this end, various groups have developed criteria and guidelines that assist practitioners and researchers in the identification of empirically supported interventions from studies that employed group designs (e.g., Briss et al., 2000; Chambless & Ollendick, 2001; Hadorn, Baker, Hodges, & Hicks, 1996; Harbour & Miller, 2001; Southam-Gerow & Prinstein, 2014) and single-case experimental designs (SCEDs; e.g., Horner et al., 2005; Kratochwill et al., 2010). This paper will focus on guidelines for SCEDs.

Single-case experimental designs involve using repeated measurements and within-subject replications of the effects of interventions for individual participants (Barlow, Nock, & Hersen, 2009; Kazdin, 2011). Because SCEDs evaluate changes in individuals and are often less costly than between-group alternatives (e.g., Bulkeley, Bundy, Roberts, & Einfeld, 2013; Hawkins, Sanson-Fisher, Shakeshaft, D'Este, & Green, 2007), many researchers have adopted their use to provide empirical support for interventions used by practitioners. In this vein, researchers and practitioners have recently called for more extensive training with SCEDs for graduate students in clinical sciences, psychology, and education (e.g., Onken, Carroll, Shoham, Cuthbert, & Riddle, 2014; Shoham et al., 2014; Smith, 2012).

The application of SCEDs in education, psychology, and related health-service disciplines has expanded considerably in the past decade (e.g., Kazdin, 2011; Novotny et al., 2014). Focusing just on psychology, Smith (2012) has identified 409 studies that used SCEDs in the PsycINFO® database and were published between 2000 and 2010. Surprisingly, Smith's search did not include commonly used descriptors of SCEDs such as reversal design, ABAB design, and multielement design. In 2008 alone, Shadish and Sullivan (2011) identified 113 studies involving single-case experiments when using a more comprehensive set of keywords. As such, it is likely that the number of SCED studies identified by Smith was a substantial underestimation.

During the past three years, at least four journals have dedicated special issues to SCED methodology. Specifically, articles published in *Journal of Applied Sport Psychology* (Barker, Mellalieu, McCarthy, Jones, & Moran, 2013), *Journal of Behavioral Education* (Burns, 2012), *Neuropsychological Rehabilitation* (Evans, Gast, Perdices, & Manolov, 2014), and *Journal of School Psychology* (Shadish, 2014) address topics such as the experimental merits of SCEDs, statistical analysis of data collected within SCEDs, and meta-analysis of results from multiple

SCEDs. In addition, a recent issue of *Remedial and Special Education* (Maggin & Chafouleas, 2013) published several commentaries on the What Works Clearinghouse (WWC; Kratochwill et al., 2010) criteria (also known as the 5-3-20 criteria [see below]; Kratochwill et al., 2013, see also Horner et al., 2012) for determining evidence standards for SCEDs to inform the development of policy and practice in education and psychology.

In addition to documenting the number of published SCEDs studies, Smith (2012) summarized various standards for evaluating and consolidating results from SCEDs, including those from the APA Divisions 12 and 16 task forces, the National Reading Panel, and WWC. Although the Division 12 criteria may be the most cited, the WWC criteria are arguably the most specific for studies employing SCEDs. To be categorized as “well established” based on the Division 12 Task Force, an intervention must have empirical support from at least two independent research teams for a total of nine successful experiments (Chambless et al., 1998). Chambless et al. (1998) did not specify a minimum number of studies, but we may infer a minimum of two as they required two independent research teams. Moreover, the demonstrated effects in these studies must be superior to a placebo or at least equivalent to other well-established interventions. The studies must also rely on treatment manuals and clearly specify the population sample.

By comparison, Kratochwill et al. (2010, 2013) proposed that a given intervention had “strong evidence of a causal relation” when five or more studies published by three or more independent research groups contained 20 or more successful experiments (i.e., the 5-3-20 criteria). Kratochwill et al. (2010) provided more specific guidelines regarding the demonstration of experimental control with each type of SCED than Chambless et al. (1998). Both articles emphasize detection of effects via visual analysis. In addition, there are at least two published scales with guidelines for evaluating the quality of individual SCED studies: one in pediatrics (Logan, Hickman, Harris, & Heriza, 2008) and another in neuropsychological rehabilitation (Tate, Perdices, McDonald, Togher, & Rosenkoetter, 2014). However, neither scale directly informs practitioners about the number of successful experiments required to determine if an intervention has sufficient evidence to be implemented with a specific person with a given problem.

While a reasonable argument can be made for (a) outlining specific criteria for experimental control with each SCED and (b) requiring two or more independent studies with multiple successful experiments, there appears to be neither a theoretical nor empirical basis for the number of successful experiments recommended in the criteria above. This issue raises an important question for evidence-based decision-making: How much is enough? To address this question, we propose the use of success rate as a supplement to the aforementioned criteria. Our proposed approach provides a quantitative model for the number of successful experiments and addresses the need to consolidate both the positive and negative findings when determining whether a given intervention has sufficient empirical support. What follows are our guidelines and rationale for using success rate with SCEDs to support evidence-based decisions.

Guidelines to Support Evidence-Based Decisions

To support practitioners, we propose specific guidelines to inform decision-making based on single-case experiments. The purpose of these guidelines is to assist practitioners in identifying interventions that have sufficient support from the research literature. When aiming to use an empirically supported approach, practitioners should consider interventions that have been the subject of studies meeting the following guidelines:

1) These experiments are well-designed (see Kratochwill et al., 2010), the characteristics of the participants are provided, and the description of the procedures allows their replication.
2) Two or more independent research teams have observed clinically or socially significant changes (i.e., replication across more than one site and research team).
3) A sufficient number of SCEDs were conducted so that the observed success rate can be estimated with a confidence interval (CI) range of 40% or less.
4) The aggregated results indicate that:
    a) the success rate is higher or falls within the CI of other available interventions meeting guidelines 1 to 3 (if available);

       OR
    b) the lower bound of the CI falls within the minimal acceptable success rate for the practitioner given the target problem (if other interventions are not available).

The previous guidelines should assist practitioners who must make informed decisions about intervention selection in applied settings. However, practitioners must understand the assumptions underlying these guidelines in order to apply them proficiently. To this end, we provide detailed explanations for the adoption of the proposed guidelines in the subsequent sections.

Characteristics and Analysis of Well-Designed Single-Case Experiments

Although a full review of SCEDs is beyond the scope of this paper, it is important to examine the characteristics and analysis of well-designed experiments. Commonly used SCEDs include multiple baseline designs (across participants, behaviors, or settings), reversal designs, multielement or alternating treatments designs, changing criterion designs, and combination designs (e.g., multiple baseline across participants with embedded multielement design for each participant; see Kazdin, 2011 for a review). Within these designs, the experimenter systematically introduces, withdraws, or modifies the parameters of the intervention to demonstrate that the effects are indeed due to the intervention and not to some extraneous variable. By carefully and systematically manipulating the presence of the intervention or its parameters, the experimenter can attribute the observed changes to the intervention itself.

For each type of SCED, experimental control of a dependent variable by the independent variable (i.e., intervention) is determined via three demonstrations of a predicted effect at three points in time within the design (Horner et al., 2012; Kratochwill et al., 2010, 2013). For example, the demonstration of experimental control requires at least three phase changes (e.g., ABAB) for reversal designs, at least six data points (i.e., three baseline and three intervention sessions) for multielement designs, and at least three parameter changes for changing criterion designs. In each of the previous designs, a single participant is sufficient to demonstrate experimental control.

Consistent with this “three-point” guideline, many researchers have suggested using at least three tiers (i.e., participants, behaviors, or settings) to demonstrate experimental control when using multiple baseline designs (Christ, 2007; Kazdin, 2011; Kratochwill et al., 2010; Novotny et al., 2014). When using multiple baseline designs, a single demonstration of experimental control can be made with as few as one participant, as with a multiple baseline design across three or more behaviors or settings, and with as many as three or more participants, as with a multiple baseline design across participants. In this way, a SCED may demonstrate experimental control with one participant, but may require three or more. By demonstrating the effects within or across participants three or more times, the experimenter minimizes the likelihood that the results were produced by extraneous variables (e.g., maturation, history, anecdotal event). Thus, the number of demonstrations of experimental control is *not* necessarily synonymous with number of participants in SCED studies. The extent to which experimental control is demonstrated via visual analysis in each SCED is categorized dichotomously (Horner et al., 2012). That is, a given SCED either does or does not depict experimental control of a dependent variable by an independent variable (i.e., intervention).

When using structured visual analysis or statistical analyses methods, many recent studies have shown that well-designed SCEDs produce Type I error rates below .05 and have power values above .80 (e.g., Bartlett, Rapp, & Henrickson, 2011; Fisher, Kelley, & Lomas, 2003; Heyvaert & Onghena, 2014; Manolov, Sierra, Solanas, & Botella, 2014; Parker, Vannest, & Davis, 2011; Solomon, 2013; Shadish et al., 2014; Solmi & Onghena, 2014; Swaminathan, Rogers, & Horner, 2014). Hence, the internal validity and statistical values yielded by well-designed SCEDs are increasingly accepted in the fields of medicine, psychology, and education (Hawkins et al., 2007; Horner et al., 2005; Lillie et al., 2011; Tate et al., 2014). The primary concerns with SCEDs are the generality or generalizability of the results. For practitioners, the problem is thus determining how many single-case experiments from how many published studies are enough to have sufficient confidence that an intervention will be effective with an individual from a given population.

Using Success Rate as a Decision-Making Tool

For practitioners, the purpose of identifying empirically supported interventions is to increase confidence that a given intervention implemented by trained professionals will generally produce desirable effects with individuals who share common characteristics with the sample population (Chambless & Hollon, 1998; Chambless & Ollendick, 2001). A discussion of what participant characteristics should be considered when determining whether an intervention is appropriate for a given person is beyond the scope of this paper; however, recent articles by Maggin and colleagues address this issue to some extent (e.g., Maggin & Chafouleaus, 2013; Maggin, O'Keeffe, & Johnson, 2011). When practitioners are attempting to select an intervention in an applied setting, they are asking themselves, "What is the likelihood that this given intervention will solve this person's problem?" The answer to this question is a matter of success rate. For practitioners, knowing the percentage of similar individuals with whom an intervention has been effective in the past is important.

Completion of a series of successful experiments by two or more independent research teams may provide a percentage of individuals with whom the intervention produced observable and desirable changes (i.e., success rate). Although there is growing interest in statistical analyses of data sets within SCEDs (e.g., Kazdin, 2011; Manolov et al., 2014; Shadish, 2014; Smith, 2012), it is likely that visual analysis remains the primary tool for SCED researchers and practitioners (e.g., Fisher & Lerman, 2014; Manolov et al., 2014; Ninci, Vannest, Willson, & Zhang, 2015). As a whole, SCEDs are generally best at evaluating moderate-to-large changes that are evident via visual analysis (Barlow et al., 2009; Cooper, Heron, & Heward, 2007; Kazdin, 2011). By contrast, their sensitivity is somewhat restricted with visual analysis when detecting small changes in the dependent variables (Fisher et al., 2003; Kazdin, 2011). This characteristic of SCEDs is actually an advantage when attempting to identify success rates because the designs will tend to detect primarily large changes with individuals who clinically or socially benefited from the intervention. Hence, the use of SCEDs will tend to yield fairly conservative values of success rates.

Using binomial statistics, we can readily compute the CIs for success rates observed for interventions validated using SCEDs[1]. The practitioner can be confident that the actual success rate falls within the CI, which makes it a useful tool for decision-making. Based on this approach, there are three important issues for identifying empirically supported interventions: the definition of what constitutes a successful experiment, the minimum success rate threshold at which an intervention should be categorized as "empirically supported", and the minimum number of experiments necessary to draw useful conclusions about the generality of the intervention effects.

The notion of what constitutes a successful experiment is central to calculating a success rate for a given intervention. In general, practitioners aim to produce socially or clinically important changes in clients, students, or patients (Baer, Wolf, & Risley, 1968; Jacobson, Follette, & Revenstorf, 1984). From a clinical standpoint, an intervention is a success when it produces changes that have a meaningful impact on the person's health, functioning, or integration. Practitioners are typically concerned with identifying interventions that will produce these socially or clinically important changes, which is why any definition of success should take into consideration this parameter. To determine whether a change is socially or clinically significant in a given experiment, we suggest that the practitioners use their goal or objective for the person as a benchmark. For example, a practitioner aiming to reduce a person's behavior to a specific target may only consider experiments that achieve this target as successful.

For the purpose of the guidelines, we propose considering an experiment as successful when (a) experimental control is demonstrated over the target, (b) the change is in the desired

[1] Note that at least seven methods are available to compute confidence intervals for binomial distributions (Newcombe, 1998). As recommended by Brown, Cai, and DasGupta (2001) for small sample sizes, we used the Wilson score method (Wilson, 1927) at a 95% confidence level to compute all the intervals reported in the current paper. Because manually calculating the Wilson score confidence interval can be complex, we recommend that practitioners use online calculators or spreadsheets to compute specific values (e.g., Herbert, 2011; Lowry, n.d.).

direction, and (c) the observed change is socially or clinically significant. Any experiment not meeting these criteria would be categorized as unsuccessful. As mentioned previously, the number of participants necessary to demonstrate experimental control varies based on design. For studies using multielement and reversal designs, each participant counts as a success if the aforementioned criteria are met. In contrast, at least three participants will be necessary to produce a single successful experiment for multiple baseline designs across participants.

The second issue that arises is one of setting a success rate at which we would consider an intervention to be sufficiently supported. For example, an intervention that increases the *survival rate* (i.e., success rate) of individuals with a life-threatening disease by 5% may have sufficient support to be implemented on a large scale, particularly if other empirically supported options are not recognized (Rosenthal, 1990). However, a behavioral intervention to reduce stress that is effective with only 10% of individuals with whom it is implemented would most likely not be recommended. Any value set will invariably depend highly on the type of intervention being evaluated. The acceptable success rate will also be relative to other interventions available in the research literature. For example, let’s assume that the most effective intervention for a given problem has a success rate of 30%. In this case, an intervention with a success rate similar to or above 30% would probably be considered as empirically supported. In contrast, if a relatively new intervention had a success rate of 60%, but that the most effective known intervention for the same population had a success rate of 80%, the former would probably not be recommended by practitioners and researchers. Using this relative approach, an empirically supported intervention is defined based on the evidence available for interventions for the same problem within a similar population (e.g., Slocum, Spencer, & Detrich, 2012). This approach ensures that practitioners consider interventions that are most likely to produce beneficial changes first.

The third issue with using success rates is the minimum number of experiments required to have confidence in the generality of the collective results. As mentioned previously, SCEDs have strong internal validity; researchers and practitioners can have high levels of confidence that changes observed during well-designed single-case experiments were the result of the intervention and not extraneous variables. In order to control for potential bias, the experiments should have been conducted by two or more independent research teams, with each team conducting at least one successful experiment (i.e., an experiment that produces clinically or socially significant changes). However, having few experiments will produce a disproportionally large CI range. Let's assume that two research teams have each conducted two experiments (four total experiments) and three of these experiments produced improvements. In this case, the success rate would be 75%, but the CI would range from 30% to 95%. That is, the range would cover most of the possible success rates. In this situation, the extent to which the findings from the three experiments will be applicable to similar individuals with a comparable problem is not clear.

Within the context of service delivery, we argue that to be considered useful, the CI range should be narrower (i.e., 40% or less). As part of our guidelines, we propose a maximum range of 40% because it is narrow enough to provide a useful estimate to practitioners selecting interventions in applied settings in which there is often considerable variability in treatment success while also ensuring that the lower bound of the estimate will never fall below 30% (for success rates of 50% or more). That said, practitioners needing more accurate estimates of success rates may set the maximum range at a lower value. It should be noted that the range is a proposed maximum: As more successful experiments are reported in the literature, the CI will generally narrow so that the estimates of success rates become more precise (Brown et al., 2001;

Newcombe, 1998). This method establishes a minimum number of replications across experiments that are necessary for an intervention to be considered empirically supported. One particularity of computing CIs based on the binomial distribution is that, given a constant sample size, the closer the observed value of the success rate is to 50% (i.e., the probability of the intervention success and failure is equivalent), the larger the CI range. When the observed success rate is 100%, only six single-case experiments are necessary to reduce the range of the CI to acceptable levels (i.e., 40% or less). In contrast, the minimum number single-case experiments that will produce a CI of 40% or less regardless of success rate is 20.

Table 1 shows the minimum number of single-case experiments to achieve a success rate of at least 50% with a CI range of 40% or less. Practitioners may use this Table to rapidly identify interventions that have a success rate above 50% and an adequate CI range (i.e., 40% or less). For example, if 11 of 12 experiments published on a particular intervention were successful, the estimate of success rate (i.e., 92%; CI [65%, 99%]) would be sufficiently accurate for the practitioner to be able to recommend and implement it with an individual with similar characteristics. If only 6 of 12 experiments were successful (50%; CI [25%, 75%]), the margin of error on the success rate would be too broad to draw useful conclusions about the generality of the intervention effects. In the latter case (i.e., success rate of exactly 50%), it would require at least 20 experiments for the range to be sufficiently narrow for the practitioner to have adequate confidence in the success rate predicted by the experiments. Table 1 also shows that if the number of successful experiments is above 12, the margin of error will always be 40% or less. It should be noted that the practitioner should look up the exact CI as more successful experiments are available, the narrower it will be (for success rates above 50%).

As an example of convenience (both authors conduct research on vocal stereotypy), we applied the guidelines stated above to SCED studies on using noncontingent matched stimulation to reduce vocal stereotypy in children with autism spectrum disorders (ASD). Noncontingent matched stimulation consists of providing continuous access to music or toys that produce auditory stimulation (Rapp, 2007). After reviewing the literature, we found 13 studies (Anderson & Le, 2011; Carroll & Kodak, 2014; Lanovaz & Argumedes, 2009; Lanovaz, Rapp, & Ferguson, 2012; Lanovaz et al., 2014; Lanovaz & Sladeczek, 2011; Lanovaz, Sladeczek, & Rapp, 2011, 2012; Love, Miguel, Fernand, & LaBrie, 2012; Rapp, 2007; Rapp et al., 2013; Saylor, Sidener, Reeve, Fetherston, & Progar, 2012; Taylor, Hoch, & Weissman, 2005) published by multiple independent research groups that included at least one well-designed single-case experiment, which examined the effects of noncontingent matched stimulation on vocal stereotypy in children between 4 and 12 years old. Across studies, there were 38 experiments for which 32 (from five independent research groups) successfully reduced engagement in vocal stereotypy for children with ASD. By using an online calculator (e.g., Herbert, 2011; Lowry, n.d), we can readily compute that the success rate for noncontingent matched stimulation is currently 84%, CI [70%, 93%]. As predicted by Table 1, the CI range is 40% or less as the number of single-case experiments exceeds 20. Based on these results, practitioners can be fairly confident that noncontingent matched stimulation will reduce engagement in vocal stereotypy for a majority of children with ASD between the ages of 4 and 12 years with whom it is correctly implemented.

Based on the previous guidelines, the notion of empirical support is not absolute but rather dependent on the state of the knowledge at any given point in time. For example, an intervention with a 45% success rate may be deemed as empirically supported at first, but may eventually become obsolete if another intervention is eventually shown to have a success rate of

75%. From a service delivery perspective, this approach ensures that individuals are receiving the intervention most likely to improve their condition in relation to their problem as more evidence becomes available. When the lack of other interventions prevents comparisons, the practitioner should use clinical judgment by considering the nature and severity of the problem. As discussed earlier, interventions with lower success rates may be adequate for life-threatening conditions, but would not be recommended for psychological, behavioral, or educational conditions. The nature and impact of the person’s problem is thus central to setting a minimum acceptable success rate by the practitioner.

Other Important Considerations

Success rate is not the only variable that practitioners should consider when identifying potential interventions for an individual. If two interventions have similar support, practitioners should also integrate other variables such as side-effects, cost, preference, and expertise (e.g., Kazdin, 2008; Logan et al., 2008; Sackett et al., 2006). Interventions that produce fewer side-effects, are less costly, are preferred by the individual or in which the practitioner holds expertise should be recommended before others given similar success rates within populations with similar characteristics. Adopting this approach will guide practitioners in offering the best potential interventions for their clients, students, or patients. As recommended by Wilcynski (2012), “even when efficacious [empirically supported] interventions are selected, the only way to know if the intervention is effective for a given client is to collect data…using one of the many different single-subject [case] research designs….” (p. 308).

We emphasize that the guidelines are recommended for practitioners attempting to determine whether one intervention should be used for a particular individual. As such, it should not replace the use of meta-analyses in the research literature, which are generally more comprehensive but also more complex, time consuming, and impractical for practitioners (e.g., De Los Reyes & Kazdin, 2009; Ma, 2009; Rosenthal, Rosnow, & Rubin, 1999; Shadish, 2014). That said, systematically presenting success rates with CIs in meta-analyses would provide important information for practitioners identifying empirically supported interventions while guiding researchers for future investigations.

Conclusions

Using success rate as a decision-making tool for identifying empirically supported interventions is a potent alternative to using a one-size-fits-all criterion (e.g., nine experiments) when relying on single-case methodology. Depending on success rate and its CI, the minimum number of experiments necessary will vary between 6 and 20 to provide an estimate precise enough to support practitioners. Interestingly, these recommendations fall within the range of other standards and guidelines for single-case experiments (e.g., Chambless & Ollendick, 2001; Horner et al., 2005; Kratochwill et al., 2010).

One limitation of inferring success rates from single-case experiments is related to the publication bias. For example, a recent study by Sham and Smith (2014) found that effect sizes of single-case experiments published in peer reviewed journals were higher than those produced when other types of publications (i.e., thesis and dissertations) were included in the analyses. Put differently, it is possible that well-designed SCED studies that produced weak or negative outcomes are less likely to be published. Relatedly, SCED researchers may be less inclined to submit papers containing weak or negative findings for review. This problem is not limited to single-case experiments; it affects other types of designs and has been omnipresent in multiple fields of study (Easterbrook, Gopalan, Berlin, & Matthews, 1991; Ferguson & Brannick, 2012). One solution for practitioners is to also search for unpublished sources, but we believe that this would be impractical and too time consuming. As such, this problem may remain until peer review journals accept more failures to replicate within their publications.

A similar concern is that not all practitioners involved in the selection of interventions have access to peer-reviewed journals and specialized abstract databases (e.g., PsycInfo®) within their workplace. One potential solution is to use freely available databases such as PubMed® or Google Scholar ® to search abstracts and then access the articles through local university libraries. Professionals should also check with their licensing or certifying agency as some provide free access to certain journals and databases to members (e.g., Behavior Analyst Certification Board, 2014). If these options are not viable, we argue that it is an ethical obligation for psychologists, behavior analysts, and other allied health professionals to remain informed of the latest research literature. Therefore, these practitioners should advocate for access to peer-reviewed journals and specialized abstract databases in their workplace. Finally, our guidelines are based on the premise that SCEDs are best at identifying moderate-to-large improvements (Fisher et al., 2003; Kazdin, 2011). The proposed methodology may not be well suited to detect small changes in behavior. Because the practitioner's purpose is generally to produce noticeable improvements in functioning, it seems reasonable to aim for producing moderate-to-large changes; inclusion of potentially small changes in the success rate statistics would likely have a negligible impact on decision-making.

The current paper focused on the use of SCEDs for improving clinical and educational practices. For practitioners, knowing with how many individuals the intervention was successful is probably as, if not more, important than knowing the mean effects. Considering the importance of estimating the probability of successful outcomes for practitioners, using success rate to support decision-making tool has the potential for improving the quality of services provided to individuals who receive psychological, behavioral, and educational services.

References

American Psychological Association. (2005, August). *Policy statement on evidence-based practice in psychology*. Retrieved from: http://www.apa.org/practice/guidelines/evidence-based-statement.aspx

Anderson, J., & Le, D. D. (2011). Abatement of intractable vocal stereotypy using an overcorrection procedure. *Behavioral Interventions*, *26*, 134-146. doi: 10.1002/bin.326

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis1. *Journal of Applied Behavior Analysis*, *1*, 91-97. doi: 10.1901/jaba.1968.1-91

Barker, J. B., Mellalieu, S. D., McCarthy, P. J., Jones, M. V., & Moran, A. (2013). Special issue on single-case research in sport psychology. *Journal of Applied Sport Psychology, 25*, 1-3. doi: 10.1080/10413200.2012.729378

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

Bartlett, S. M., Rapp, J. T., & Henrickson, M. L. (2011). Detecting false positives in multielement designs implications for brief assessments. *Behavior Modification*, *35,* 531-552. doi: 10.1177/0145445511415396

Behavior Analyst Certification Board. (2014, August). A new resource for BACB certificants. *BACB Newsletter – Special Issue on the ProQuest Benefit*. Retrieved from http://bacb.com/wp-content/uploads/2015/07/BACB_Newsletter_08-14.pdf

Briss, P. A., Zaza, S., Pappaioanou, M., Fielding, J., Wright-De Agüero, L., Truman, B. I., ... & Harris, J. R. (2000). Developing an evidence-based guide to community preventive services—methods. *American Journal of Preventive Medicine*, *18*, 35-43. doi: 10.1016/S0749-3797(99)00119-1

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science, 16*, 101-117. doi: 10.1214/ss/1009213286

Bulkeley, K., Bundy, A., Roberts, J., & Einfeld, S. (2013). ASD intervention in research in real world contexts. Refining single case designs. *Research in Autism Spectrum Disorders*, *7*, 1257-1264. doi: 10.1016/j.rasd.2013.07.014

Burns, M. K. (2012). Meta-analysis of single-case design research: Introduction to the special issue. *Journal of Behavioral Education*, *21*, 175-184. doi: 10.1007/s10864-012-9158-9

Carroll, R. A., & Kodak, T. (2014). An evaluation of interrupted and uninterrupted measurement of vocal stereotypy on perceived treatment outcomes. *Journal of Applied Behavior Analysis, 46*, 264-276. doi: 10.1002/jaba.118

Chambless, D. L., Baker, M. J., Baucom, D. H., Beutler, L. E., Calhoun, K. S., Crits-Christoph, P., ... & Woody, S. R. (1998). Update on empirically validated therapies II. *The Clinical Psychologist*, *51*, 3-16.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66*, 7-18. doi: 10.1037/0022-006X.66.1.7

Chambless, D. L., & Ollendick, T. H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Annual Review of Psychology, 52*, 685-716. doi: 10.1146/annurev.psych.52.1.685

Christ, T. J. (2007). Experimental control and threats to interval validity of concurrent and nonconcurrent multiple baseline designs. *Psychology in the Schools, 44*, 451-459. doi: 10.1002/pits.20237

Chwalisz, K. (2003). Evidence-based practice: A framework for twenty-first-century scientist-practitioner training. *The Counseling Psychologist, 31,* 497-528. doi: 10.1177/0011000003256347

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Upper Saddle River, NJ: Pearson Education.

De Los Reyes, A., & Kazdin, A. E. (2009). Identifying evidence-based interventions for children and adolescents using the range of possible changes model: A meta-analytic illustration. *Behavior Modification, 33,* 583-617. doi: 10.1177/0145445509343203

Easterbrook, P. J., Gopalan, R., Berlin, J. A., & Matthews, D. R. (1991). Publication bias in clinical research. *The Lancet, 337*, 867-872. doi: 10.1016/0140-6736(91)90201-Y

Evans, J. J., Gast, D. L., Perdices, M., & Manolov, R. (2014). Single-case experimental designs: Introduction to a special issue of Neuropsychological Rehabilitation. *Neuropsychological Rehabilitation, 24,* 305-314 doi: 10.1080/09602011.2014.903198

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, *17*, 120-128. doi: 10.1037/a0024445

Fisher, W. W., Kelley, M. E., & Lomas, J. E. (2003). Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. *Journal of Applied Behavior Analysis*, *36*, 387-406. doi: 10.1901/jaba.2003.36-387

Fisher, W. W. & Lerman, D. C. (2014). It has been said that, “There are three degrees of falsehoods: Lies, damn lies, and statistics.” *Journal of School Psychology, 52,* 243-248. doi: 10.1016/j.jsp.2014.01.001

Hadorn, D. C., Baker, D., Hodges, J. S., & Hicks, N. (1996). Rating the quality of evidence for clinical practice guidelines. *Journal of Clinical Epidemiology, 49*, 749-754. doi: 10.1016/0895-4356(96)00019-4

Harbour, R., & Miller, J. (2001). A new system for grading recommendations in evidence based guidelines. *BMJ: British Medical Journal*, *323,* 334-336. doi: 10.1136/bmj.323.7308.334

Hawkins, N. G., Sanson-Fisher, R. W., Shakeshaft, A., D'Este, C., & Green, L. W. (2007). The multiple baseline design for evaluating population-based research. *American Journal of Preventative Medicine, 33,* 162-168. doi:10.1016/j.amepre.2007.03.020

Herbert, R. (2011, March 24). Confidence interval calculator [Online calculator]. Retrieved from http://www.pedro.org.au/wp-content/uploads/CIcalculator.xls

Heyvaert, M., & Onghena, P. (2014). Analysis of single-case data: Randomisation tests for measures of effect size. *Neuropsychological Rehabilitation, 24,* 207-527. doi: 10.1080/09602011.2013.818564

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. doi: 10.1177/001440290507100203

Horner, R. H., Swaminathan, H., Sugai, G., & Smolkowski, K. (2012). Considerations for the systematic analysis and use of single-case research. *Education and Treatment of Children, 35,* 269-290. doi: 10.1353/etc.2012.0011

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15,* 336-352. doi: 10.1016/S0005-7894(84)80002-7

Kazdin, A. E. (2008). Evidence-based treatment and practice: New opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *American Psychologist, 63,* 146-159. doi: 10.1037/0003-066X.63.3.146

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

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 http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Kratochwill, T. R., Hitchcock, J., 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. doi: 10.1177/0741932512452794

Lanovaz, M. J. & Argumedes, M. (2009). Using the three-component multiple-schedule to examine the effects of treatments on stereotypy. *Journal on Developmental Disabilities, 15*(3), 64-68.

Lanovaz, M. J., Rapp, J. T., & Ferguson, S. (2012). The utility of assessing musical preference before implementation of noncontingent music to reduce vocal stereotypy. *Journal of Applied Behavior Analysis*, *45*, 845-851. doi: 10.1901/jaba.2012.45-845

Lanovaz, M. J., Rapp, J. T., Maciw, I., Prégent-Pelletier, É., Dorion, C., Ferguson, S., & Saade, S. (2014). Effects of multiple interventions for reducing vocal stereotypy: Developing a sequential intervention model. *Research in Autism Spectrum Disorders*, *8,* 529-545. doi: 10.1016/j.rasd.2014.01.009

Lanovaz, M. J., & Sladeczek, I. E. (2011). Vocal stereotypy in children with autism: Structural characteristics, variability, and effects of auditory stimulation. *Research in Autism Spectrum Disorders*, *5*, 1159-1168. doi: 10.1016/j.rasd.2011.01.001

Lanovaz, M. J., Sladeczek, I. E., & Rapp, J. T. (2011). Effects of music on vocal stereotypy in children with autism. *Journal of Applied Behavior Analysis*, *44*, 647-651. doi: 10.1901/jaba.2011.44-647

Lanovaz, M. J., Sladeczek, I. E., & Rapp, J. T. (2012). Effects of noncontingent music on vocal stereotypy and toy manipulation in children with autism spectrum disorders. *Behavioral Interventions*, *27*, 207-223. doi: 10.1002/bin.1345

Lillie, E. O., Patay, B., Diamant, J., Issell, B., Topol, E. J., & Schork, N. J. (2011). The n-of-1 clinical trial: The ultimate strategy for individualizing medicine? *Personalized medicine, 8*, 161-173. doi: 10.2217/pme.11.7

Logan, L., Hickman, R., Harris, S., & Heriza, C. (2008). Single-subject research design: Recommendations for levels of evidence and quality rating. *Developmental Medicine & Child Neurology*, *50*, 99–103. doi: 10.1111/j.1469-8749.2007.02005.x

Love, J. J., Miguel, C. F., Fernand, J. K., & LaBrie, J. K. (2012). The effects of matched stimulation and response interruption and redirection on vocal stereotypy. *Journal of Applied Behavior Analysis*, *45*, 549-564. doi: 10.1901/jaba.2012.45-549

Lowry, R. (n.d.). The confidence interval of a proportion [Online calculator]. Retrieved from http://vassarstats.net/prop1.html

Ma, H. H. (2009). The effectiveness of intervention on the behavior of individuals with autism: A meta-analysis using percentage of data points exceeding the median of baseline phase (PEM). *Behavior Modification*, *33*, 339-359. doi: 10.1177/0145445509333173

Maggin, D. M. & Chafouleas, S. M. (2013). Introduction to the special series: Issues and advances of synthesizing single-case research. *Remedial and Special Education, 34,* 3-8. doi: 10.1177/0741932512466269

Maggin, D. M., O'Keeffe, B. V., & Johnson, A. H. (2011). A quantitative synthesis of methodology in the meta-analysis of single subject research for students with disabilities: 1985–2009. *Exceptionality, 19,* 109–135. doi: 10.1080/09362835.2011.565725

Manolov, R., Sierra, V., Solanas, A., & Botella, J. (2014). Assessing functional relations in single-case designs quantitative proposals in the context of the evidence-based movement. *Behavior Modification*, *38,* 878-913. doi: 10.1177/0145445514545679

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine, 17,* 857-872. doi: 10.1002/(SICI)1097-0258(19980430)17:8<857::AID-SIM777>3.0.CO;2-E

Ninci, J., Vannest, K. J., Willson, V., & Zhang, N. (2015). Interrater agreement between visual analysts of single-case data: A meta-analysis. *Behavior Modification*. Advanced online publication. doi:10.1177/0145445515581327

Novotny, M. A., Sharp, K. J., Rapp, J. T., Jelinski, J. D., Lood, E. A., & Steffes, A. K. (2014). False positives with visual analysis for nonconcurrent multiple baseline designs and ABAB designs: Preliminary findings. *Research in Autism Spectrum Disorders, 8,* 933-943. doi: 10.1016/j.rasd.2014.04.009

Onken, L. S., Carroll, K. M., Shoham, V., Cuthbert, B. N., & Riddle, M. (2014). Reenvisioning clinical science: Unifying the discipline to improve the public health. *Clinical Psychological Science*, *2,* 22-34. doi: 10.1177/2167702613497932

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification, 35*, 303-322. doi: 10.1177/0145445511399147

Rapp, J. T. (2007). Further evaluation of methods to identify matched stimulation. J*ournal of Applied Behavior Analysis, 40*, 73-88. doi: 10.1901/jaba.2007.142-05

Rapp, J. T., Swanson, G., Sheridan, S. M., Enloe, K. A., Maltese, D., Sennott, L. A., ... & Lanovaz, M. J. (2013). Immediate and subsequent effects of matched and unmatched stimuli on targeted vocal stereotypy and untargeted motor stereotypy. *Behavior Modification, 37*, 543-567. doi: 10.1177/0145445512461650

Rosenthal, R. (1990). How are we doing in soft psychology?. *American Psychologist, 45*, 775-777. doi: 10.1037/0003-066X.45.6.775

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (1999). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge University Press.

Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *BMJ: British Medical Journal, 312*, 71-72. doi: 10.1136/bmj.312.7023.71

Saylor, S., Sidener, T. M., Reeve, S. A., Fetherston, A., & Progar, P. R. (2012). Effects of three types of noncontingent auditory stimulation on vocal stereotypy in children with autism. *Journal of Applied Behavior Analysis*, *45*, 185-190. 10.1901/jaba.2012.45-185

Shadish, W. R. (2014). Analysis and meta-analysis of single-case designs: An introduction. *Journal of School Psychology, 52*, 109-122. doi: 10.1016/j.jsp.2013.11.009

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,* 528-553. doi: 10.1080/09602011.2013.819021

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. doi: 10.3758/s13428-011-0111-y

Sham, E., & Smith, T. (2014). Publication bias in studies of an applied behavior-analytic intervention: An initial analysis. *Journal of Applied Behavior Analysis*, *47*, 663-678. doi: 10.1002/jaba.146

Shoham, V., Rohrbaugh, M. J., Onken, L. S., Cuthbert, B. N., Beveridge, R. M., & Fowles, T. R. (2014). Redefining clinical science training: Purpose and products of the Delaware project. *Clinical Psychological Science, 2,* 8-21. doi: 10.1177/2167702613497931

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, *31,* 15-21. doi: 10.3102/0013189X031007015

Slocum, T. A., Spencer, T. D., & Detrich, R. (2012). Best available evidence: Three complementary approaches. *Education and Treatment of Children*, *35*, 153-181. doi: 10.1353/etc.2012.0013

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods, 17,* 510-550. doi: 10.1037/a0029312

Solmi, F., & Onghena, P. (2014). Combining p-values in replicated single-case experiments with multivariate outcome. *Neuropsychological Rehabilitation*, 24, 607-633. doi: 10.1080/09602011.2014.881747

Solomon, B. G. (2013). Violations of assumptions in school-based single-case data: Implications for the selection and interpretation of effect sizes. *Behavior Modification*, 38, 477-496. doi: 10.1177/0145445513510931

Southam-Gerow, M. A., & Prinstein, M. J. (2014). Evidence base updates: The evolution of the evaluation of psychological treatments for children and adolescents. *Journal of Clinical Child & Adolescent Psychology*, *43*, 1-6. doi: 10.1080/15374416.2013.855128

Spring, B. (2007). Evidence-based practice in clinical psychology: What it is, why it matters; what you need to know. *Journal of Clinical Psychology*, *63*, 611-631. doi: 10.1002/jclp.20373

Spring, B., Pagoto, S., Kaufmann, P. G., Whitlock, E. P., Glasgow, R. E., Smith, T. W…. Davidson, K. W. (2005). Invitation to a dialogue between researchers and clinicians about evidence-based behavioral medicine. *Annals of Behavioral Medicine*, *30*, 125-137. doi: 10.1207/s15324796abm3002_5

Swaminathan, H., Rogers, H. J., & Horner, R. H. (2014). An effect size measure and Bayesian analysis of single-case designs. *Journal of School Psychology, 52*, 213-230. doi: 10.1016/j.jsp.2013.12.002

Tate, R. L., Perdices, M., McDonald, S., Togher, L., & Rosenkoetter, U. (2014). The design, conduct and report of single-case research: Resources to improve the quality of the neurorehabilitation literature. *Neuropsychological Rehabilitation*, *24*, 315-331. doi: 10.1080/09602011.2013.875043

Taylor, B. A., Hoch, H., & Weissman, M. (2005). The analysis and treatment of vocal stereotypy in a child with autism. *Behavioral Interventions, 20*, 239-253. doi: 10.1002/bin.200

Wilczynski, S. M. (2012). Risk and strategic decision-making in developing evidence-based practice guidelines. *Education and Treatment of Children, 35*, 291-311. doi: 10.1353/etc.2012.0012

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*, 209-212. doi: 10.1080/01621459.1927.10502953

Table 1

Minimum number of successful experiments required to achieve a success rate of at least 50% and a CI range of 40% or less.

Total number of experiments	Minimum number of successful experiments
6	6
7	7
8	8
9	9
10	9
11	10
12	10
13	11
14	11
15	12
16	12
17	12
18	12
19	12

Note. Beyond 19 experiments, the CI range is always 40% or less regardless of success rate. The values in the table are based on a 95% confidence level using the Wilson score method without continuity correction and based on rounding to the nearest integer.

Author Biographies

Marc J. Lanovaz, Ph.D., BCBA-D is an Assistant Professor at the École de psychoéducation of the Université de Montréal. His research interests include the assessment and treatment of problem behavior, parental involvement and training, and the use of technology to facilitate the implementation of behavioral interventions.

John T. Rapp, Ph. D., BCBA-D is a Professor in the Department of Psychology at Auburn University. His research interests include the assessment and treatment of stereotypy and applications of single-subject experimental designs.