From Boulder to Stockholm in 70 Years

The Psychological Record (2020) 70:659–670
https://doi.org/10.1007/s40732-020-00402-5

THEORETICAL ARTICLE

From Boulder to Stockholm in 70 Years: Single Case Experimental Designs in Clinical Research

Johan W. S. Vlaeyen[1,2] · Rikard K. Wicksell[3] · Laura E. Simons[4] · Charlotte Gentili[3] · Tamal Kumar De[5] · Robyn L. Tate[6] · Sunita Vohra[7] · Salima Punja[7] · Steven J. Linton[8] · Falko F. Sniehotta[9] · Patrick Onghena[5]

Published online: 3 June 2020

Abstract
With the objective of increasing the magnitude of treatment effects in behavioral health, there is steadily growing interest in tailoring assessments and interventions to better match individual needs. This aligns with the central idea that behavior can be adequately understood by considering the unique characteristics of the individual and context. Thus, data collected at an individual level provides critical evidence that can be used to inform health care decisions, improve treatment, or refine theories. Yet, the majority of research in behavioral health is based on group-level analyses. Recent developments in the field of single-case experimental design (SCED) has provided new opportunities to utilize individual data. The present article provides a state-of-the art overview regarding key aspects of SCED, including a historical background to why and how SCED emerged, declined, and recently reemerged as well as methodological aspects such as design issues, challenges related to reliability and validity of repeated observations, innovations in visual and statistical analyses of individual data, strategies to deal with missing values, methodology to examine effect size, and approaches to summarize data from a large number of SCEDs using multilevel models and meta-analyses of replication data. Finally, the article discusses key concerns and actions needed to move the field forward.

Keywords Single case experimental design · Idiosyncratic assessment · Treatment outcome · Effectiveness research

Introduction

Psychology is the science of the individual, yet the bulk of available research data is derived from groups. Indeed, randomized controlled trials (RCT) are considered the gold standard for studying the effects and mechanisms of complex interventions using standardized and validated instruments. It remains difficult to discern to what extent the results of these large-scale studies are useful for individual health care users and providers. RCTs can determine the "average" result for a given outcome variable but are not suited for determining the functional relationships between a treatment and observed change for one individual. As the gerontologist John Grimley Evans remarked, "Managers and trialists may be happy for treatments to work on average; patient's doctors expect to do better than that" (Evans, 1995). Fortunately, there is an efficient and robust alternative. Single-case experimental designs (SCEDs)[1] are experimental designs in which a single unit (e.g., a client, group of clients, classroom, ward, hospital) is repeatedly observed for a predetermined period of time at various levels of at least one manipulated variable (e.g., the treatment). SCEDs have played an important role in the

Johan W. S. Vlaeyen and Rikard K. Wicksell contributed equally to this work.

Authors Note:
"Replication, replication, replication" (Morley, 2018). This article is dedicated to Stephen Morley (1950–2017), a wonderful friend and colleague who made significant contributions to the applications of single-case experimental design (SCED). It also summarizes the contributions of the Stockholm symposium, "Small is Beautiful," held October 2018, on single-case experimental designs. The objective of the meeting was to ignite discussions among researchers and clinicians actively using or interested in applying this methodology in their work, and to discuss strategies to disseminate, implement, and further develop the SCED approach.

✉ Johan W. S. Vlaeyen
johannes.vlaeyen@kuleuven.be

Extended author information available on the last page of the article

[1] It should be noted that there is an ongoing discussion within the field regarding the definitions of and relationships among terms used to label study designs and methods utilizing singe case data. In biomedical research, the term "N-of-1 trial" is commonly used for a multiple crossover evaluation performed in a single individual (Guyatt & Jaeschke, 1990). N-of-1 trials can be considered a subset of SCEDs.

development of new interventions in various subspecialties of psychology and medicine (Vlaeyen, de Jong, Geilen, Heuts, & van Breukelen, 2001). They are not only relatively easy to implement in the practice of health care, they also help us understand the circumstances and contexts under which interventions work. Repeated exposure to interventions over time and context allow testing for context by intervention interactions and time trends in a way that classical RCTs do not. SCEDs offer ongoing feedback on the progress of a person or a group, visualize progress during an intervention, and the results are available in *real time*. There is now adequate and appropriate open-source software available for both descriptive and inferential statistics (Onghena, Michiels, Jamshidi, Moeyaert, & Van den Noortgate, 2018). Finally, they elegantly bridge the gap between research and practice. In this article we start with the historical context in which SCED originally emerged, then declined, and reemerged more recently. We then present recent developments, revisit both visual and statistical analysis of single person data, and present methods to assess design and reporting quality. We end this review with some of the opportunities and challenges for further implementation of single-case design methodology in clinical behavior analysis.

Historical Background

The Boulder Conference

About 70 years ago, in the midst of a postwar period, the demand for mental health care providers substantially rose given the number of people who suffered various trauma-related disorders. In order to adequately respond to this urgency, the Committee on Training in Clinical Psychology, supported by the American Psychological Association, the United States Public Health Service, and the Veterans Administration organized the seminal Boulder Conference, bringing together representatives of the various domains of psychology. Participants convened for more than 2 weeks addressing various issues shaping the identity of the profession: needs, training, ethics, certification and licensure, and relations with other professions. At the end of the Boulder Conference, a consensus was reached, and the clinical psychologist was defined as a "scientist–practitioner" who was trained for both practice and research with equal emphasis on each. A major change was the idea that the psychologist is not just a provider of interventions, but also a researcher who systematically generates data from his or her own clinical "laboratory" enabling to improve his or her own professional practice to the benefit of the rest of the scientific community (Barlow, Nock, & Hersen, 2009). One year later, the final report expressed the expectancy that this definition was likely to affect major policies at training institutions globally for many years to come (Baker & Benjamin, 2000).

Looking back 70 years later, this wish has been only partially fulfilled. The tasks of psychologists in most countries prioritize practice, leaving little room for research (McQuaid, Aosved, & Belanger, 2018). The few clinical laboratories predominantly reside in academia, and are often disconnected with the national health care systems. The same scientist–practitioner gap can also be observed in health professions other than psychology (Smith & Wilkins, 2018). The question of why the ambitions of the Boulder Conference have only been partially achieved is an intriguing one, and here we present three potential obstacles to the clinicians' participation in research, each of them originating from different areas in science. These are (1) the interest in universal laws in behavioral science, (2) the primacy of classification over categorization in knowledge acquisition and communication and (3) the emphasis on the basic science strand of translational science.

The Interest in Universal Laws in Behavioral Science

The traditional randomized clinical trial (RCT) creates robust evidence for the effectiveness of treatments on groups, but has limited utility for health-care providers trying to make decisions for the individual in everyday practice. Nevertheless, the scientific study of single individuals has roots deep in the history of physiology and psychology. In the 19th century, great scientists reported on major discoveries that were based on studies with only a few subjects, with a focus on the dynamics and variability within each individual (Barlow et al., 2009). One of the best-known examples is Ivan Pavlov, whose far-reaching findings were based on single organisms, and validated via replication with subsequent subjects (Pavlov, 1927).

At the beginning of the 20th century, inspired by Darwin's evolutionary ideas, a new development emerged, characterized by an interest in the measurement of individual differences, and errors or deviations from the "average" individual. This new interest shifted the scientific attention towards the methodology and statistics of group-based research. Sir Ronald Aylmer Fisher worked at the Rothamsted agricultural station, where the main interest was in large crop plots rather individual plants. He not only developed sophisticated statistical tests for the comparison of groups but he was also concerned about how one could infer knowledge about a large population from a small sample. Fisher clearly marked an important step towards a nomothetic approach to science (from Greek "nomos," meaning "law"), using sample-wise generation of general laws and knowledge about a large population (Fisher, 1925). Fisher was so successful with his statistical innovations that psychological scientists who at the time were aiming for general statements about human behavior, embraced his new statistical techniques embedded within this nomothetic orientation. As a result, the previous focus on variations within a single subject decreased, instead favoring

group designs. Over time, the group-based RCT became the gold standard for testing the effectiveness of new interventions (Grossman & Mackenzie, 2005).

The Primacy of Classification over Categorization

There are two ways of acquiring and transmitting knowledge: categorization and classification (Jacob, 2004). Categorization is the context-dependent creative synthesis of bits of information based on their perceived similarity. Categorization is a bottom-up learning process that leads to the creation of concepts needed to better understand the world around us, and that facilitates the recognition of salient events in the environment (Dunsmoor, Martin, & LaBar, 2012). In categorization, the individual builds concepts that are based on previously acquired knowledge in different but similar situations. Categorization also happens in science, for example, when breaking fresh grounds and creating new ideas. Typical for categories is that they can have fuzzy boundaries with elements potentially belonging to more than one category, depending on the context. For example, "spoon" is an object that can belong to either the category "utensils" or "toys," depending on the context ("dinner table" vs. "playground"). With the accumulation of more knowledge, and the need to share knowledge in a parsimonious way, experientially based categories may gradually vacillate into "supervised categories" (Pothos & Chater, 2005) or well-defined "classes" (Jacob, 2004). Classification, in contrast, is a top-down process, where elements are systematically organized based on predefined necessary characteristics. As a result, classes are mutually exclusive, nonoverlapping, and have fixed boundaries. For example, in DSM-V, depression is defined as experiencing five or more symptoms during the same 2-week period, and at least one of the symptoms should be either (1) depressed mood or (2) loss of interest or pleasure. These classification criteria are predetermined, usually context-independent, and all members of a class are considered equally representative for that particular class. In other words, if we know that someone belongs to a certain class (e.g., depressed persons), we attribute the characteristics proper to the class to that person, even though not all of the features may be equally well-represented, or some members of the class may even possess additional features.

Although classification is helpful to summarize and share relevant information, it is also known to create two perceptual biases: *accentuation* and *assimilation*. Accentuation refers to the overestimation of between-class differences: individuals belonging to different classes (e.g., depression and anxiety) may be perceived as more different than they are. Assimilation concerns the underestimation of within-class differences: individuals belonging to the same class may be perceived as more equal than they actually are, just by virtue of being assigned to a class. Accentuation and assimilation are well-known and have been studied in various domains, including vision (Tajfel & Wilkes, 1963), perception of auditory stimuli (Campbell, 1956), social perception (Rubin & Badea, 2012), interoception (Petersen, Van Staeyen, Vogele, von Leupoldt, & Van den Bergh, 2015), and pain (van der Meulen, Anton, & Petersen, 2017). In the context of health care, assimilation bias holds the risk that it may draw attention away from idiosyncratic features of the individual that may be important for effective personalized health care. In sum, the use of classification systems may unwittingly have moved the focus of clinicians towards the study of groups representing a certain predefined class, and away from the unique characteristics and features of the individual.

Translational Science

In search of more human benefits from basic research, interest in *translational science* has risen significantly. In general, translational science means to effectively "translate new knowledge mechanisms and techniques generated by advances in basic science research into new approaches for prevention, diagnosis and treatment of disease" (Fontanarosa & DeAngelis, 2002, p. 1728). At least two kinds of translational research exist depending on the scientific outcome that is being "translated." The first is located within basic science labs where scientists with cutting-edge technology use lab results to create new assessment and treatment devices. The second strand of translational science takes place in clinical science labs, where novel insights from basic science labs are translated towards interventions that are implemented and evaluated in real-world settings (Sung et al., 2003). Although both strands of translational efforts are essential, it seems acceptable to conclude that a disproportionate amount of public resources are spent on the first, with potentially effective treatments never reaching the individuals who may benefit from them. This is concerning as evidence suggests that the health-care system might gain more from investing in how to better deliver existing treatments than producing new ones (Morris, Wooding, & Grant, 2011). Individuals (or individual units such as hospitals) create so much data that new scientific insights about effects and mechanisms could come from applied single-case analyses feeding into basic science. In this context, the renewed interest in SCED may be both logical and timely.

The Way Forward

Despite their popularity, traditional RCTs have come under scrutiny as well (Kaptchuk, 2001). Scientists have increasingly realized that the "average" person barely exists, and that results of large RCTs cannot readily be generalized to the individuals within the sample, and even less so to the population. In other words, the external validity of RCTs may have been overestimated. In addition, complex cases are usually excluded from RCTs, as they reduce homogeneity required for sufficient statistical power in group comparisons. Also, a concern that has been expressed is that most of our treatments have only medium effect sizes, with high individual variation in effectiveness, which raises the pertinent question, "What works for whom?" (Vlaeyen & Morley, 2005). In fact, improving effect sizes are likely better addressed by tailoring interventions than developing new ones. However, this requires a better understanding of the individual variation in outcome and change mechanisms. Traditional RCT methodology will not be the preferred tool to answer this complex question, and SCEDs have been suggested instead. Shapiro, drawing on the work of Bernard and Sidman argued that the site of the processes of change is the individual, and that observations based on group averages might be misleading (Bernard, 1957; Shapiro, 1966; Sidman, 1952). Other influential sources have been written by colleagues later on (Barlow et al., 2009; Guyatt & Jaeschke, 1990; Kazdin, 1983; Kratochwill & Levin, 1992; Morley, 2018; Todman & Dugard, 2001). Two recent commentaries in the journal *Nature* presented similar arguments in favor of single-case methods, but from different angles. One critically observes that the most prescribed drugs in the United States help only a fraction (between 4% and 25%) of the people who take them, and suggests that in order to enhance personalized medicine, everyday clinical practice should be transformed into N-of-1 trials (Schork, 2015). Likewise, another commentary argues that "When treatments are given on many occasions for a chronic or recurring condition, N-of-1 studies are a good way of establishing the scope for personalized medicine" (Senn, 2018, p. 621). An increasing number of reasons point in the direction of pursuing these recommendations. Not only has the SCED methodology become more sophisticated (Vohra & Punja, 2019), appropriate statistical methods have been developed (Onghena & Edgington, 2005), and are now widely available through digital tools (Bulté & Onghena, 2008). Also, guidelines have been created for both the quality of single-case experiments (Tate et al., 2013), as well as the reporting of them (Tate et al., 2016a).

In the second half of the current article we highlight emerging innovations in SCEDs. After presenting recent developments, we revisit both visual and statistical analysis of single person data, and present methods to assess design and reporting quality. We end this review with some of the opportunities and challenges for further implementation of SCED methodology.

Methodological Aspects of Single-Case Experimental Designs (SCEDs)

SCEDs: What Do They Offer in an Era of Big Data?

SCEDs refer to research designs that are applied to experiments in which one entity is observed repeatedly during a certain period of time, under different levels ("treatments") of at least one independent variable. The essential characteristics of such single-case experiments are (1) that only one entity is involved (single-case), and (2) that there is a manipulation of the independent variable(s) ("experiment"; Onghena & Edgington, 2005). SCEDs (including N-of-1 trials) overcome the major limitation of parallel group-based RCTs, which are often criticized for their lack of relevance to clinical practice, due to the exclusion of individuals with comorbidities or concurrent therapies (Rothwell, 2005). Considered level 1A evidence, equivalent to a systematic review of RCTs, randomized N-of-1 trials and internally valid SCEDs represent an innovation in clinical trial methodology by balancing individual needs with methodological rigor, and provide an individualized approach to assessing treatment effects (Group, 2011). Indeed, SCEDs are relevant across a number of conditions and populations (Punja, Bukutu et al., 2016a). For example, large-scale RCTs are challenging for pediatric illnesses or rare conditions, because it is difficult to achieve adequate sample sizes (Walburn et al., 2017). There are arguments in favor of applying SCEDs in individuals with comorbid conditions, and those using concurrent therapies. They may be used to evaluate therapeutic effectiveness in chronic conditions: in an ideal situation the symptom or outcome being assessed should occur with sufficient frequency to ensure a reasonable length of trial. SCEDs are preferably used for treatments with rapid onset/termination of effect, as there is a potential for carry-over effects (DEcIDE & Panel, 2014). They enhance safety by reducing ineffective polypharmacy, thereby reducing adverse events as well as cost, and they are relevant in the evaluation of complementary therapies, which is particularly helpful in the context of individualized therapies that are not amenable to study by RCT.

SCEDs are also relevant to help reduce therapeutic uncertainty. This can occur when there is a lack of confidence that a current treatment is providing a benefit, or if it is uncertain that a proposed treatment will work with a particular individual

(Kronish et al., 2019). Other examples include: the individual insisting on taking a treatment that the clinician thinks will not work or is harmful, a clinician suspecting that a side effect is from a treatment but is unsure, or a clinician is unsure of the optimal dose of a given treatment (Mirza, Punja, Vohra, & Guyatt, 2017).

What You See is What You Get: Visual Analysis of Single-Case Data

One of the challenges in treatment outcome studies is how to determine if reliable changes have occurred. In psychology, pioneers of SCED have argued for the visual presentation of outcomes complemented with simple descriptive and interpretative statistics (Baron & Perone, 1998; Morley, 2018; Sidman, 1960). In a typical AB phase design a comparison of the baseline (with no intervention, phase A) with what happens when the treatment (or other intervention, phase B) is introduced provides the basis for evaluation. Researchers hope that the introduction of a given treatment will result in dramatic improvements that are visible when plotting the data onto a graph. However, one question is how we might determine whether the change observed is actually useful.

A basic approach that has much to recommend it is systematic visual inspection. Although visual inspection may be complemented by statistical methods, there is great value in viewing how the results unfold. A graphical presentation of the data provides rich information, for example, about how much variability there was in the baseline, if any visible change occurs, when the change occurs, and how stable the change is over time. Such information is lost in group designs where only the mean scores of the pre-, post-, and follow-up values are provided. Although visual inspection may seem prone to bias, there are guidelines for visually assessing results (Kazdin, 2011; Kratochwill & Levin, 1992; Morley, 2018), and many experts on research design maintain that changes visible to the naked eye are convincing results. However, given the great variability in most treatment studies, drawing inferences about whether the changes observed are reliable and not simply due to chance fluctuations in the data may be challenging. There are tools to assist the researchers in valuating visual analysis of single-case data (Bulté & Onghena, 2012).

If results are not clear with visual inspection, the practitioner-scientist has two excellent options. First, descriptive and simple inferential statistics may be applied, to help to determine the size of the observed change. For example, providing a graphical line depicting the mean value for the baseline and treatment can highlight differences. Further, by comparing how many data points are above (or below) the baseline mean can be used to calculate effect sizes (Parker & Vannest, 2009). Second, the design of the study can be improved and repeated to provide more data (Sidman, 1960). The SCED, by virtue of being small, invites replication and improved methodology.

Three aspects are of particular importance in producing clear results. First, there is a need to select measures that are valid, reliable (robust to repeated use), and sensitive to change. Such measures may be actual behaviors or subjective ratings. In general, well-defined variables will tend to be more valid and reliable. Second, a common reason why visual inspection is difficult is that the baseline is too variable or too short. Establishing a stable baseline requires a reliable measure and sufficient time for the baseline to be verified. Finally, when the intervention is weak, changes are difficult to observe if the baseline is not perfectly stable. Thus, working theoretically to isolate mechanisms that should produce changes in specific variables is recommended. The use of visual inspection is useful in evaluating results presented in scientific articles as well as in assessing outcomes in the clinic. They provide a way of analyzing results on an individual level that provides a rich spectrum of information.

Forcing Round Pegs into Round Holes: The Statistical Analysis of Single-Case Experimental Data

In his influential article "Forcing Square Pegs into Round Holes: Some Comments on "an Analysis-of-Variance Model for the Intrasubject Replication Design", Hartmann (1974) criticized the use of naive ANOVA models for the analysis of single-case experimental data, but he also indicated the alternative direction in which the statistical analysis of single-case experimental data had to go. By now, the problems of naïve ANOVA models for the analysis of single-case experimental data are well documented and widely acknowledged: the assumptions of normality, homogeneity of variance, and independent residuals are implausible for this kind of data (Richards, 2019; Smith, 2012; Solomon, 2014). However, what is the present state of affairs in Hartmann's alternative direction?

It is beyond the scope of the present contribution to give a complete overview of all viable and valid statistical techniques for the analysis of single-case experimental data, but we want to propose a classification for making discussions about the analysis of single-case experimental data more perspicuous. Furthermore, we can direct the interested reader to instructive tutorials and easy-to-use software for state-of-the-art statistical analysis of single-case experimental data. For our proposed classification, the first important distinction that we have to make is the distinction between descriptive and inferential statistics (Moore & Notz, 2017). If description is the only purpose, then tables, graphs, and descriptive measures suffice. Visual analysis, as discussed in the previous section, is the most common and popular technique for analyzing single-case data (for a tutorial, see Kratochwill et al., 2010). In addition, each of the data aspects included in visual analysis

may be quantified using a descriptive measure or an effect size statistic (Tanious, De, Michiels, Van den Noortgate, & Onghena, 2019). The important point to make is that we do not have to make any statistical assumptions for descriptive analyses: we just have to be clear about the data aspect we are visualizing or describing. If statistics are used for testing hypotheses or delivering estimates, then there are two basic dichotomies. The first dichotomy refers to the distinction between model-based and design-based inference (Koch & Gillings, 1984). In model-based inference, a statistical model is fitted to the observed data, and the parameters and their standard errors are estimated from the data. In design-based inference, the data are analyzed in accordance with the way the data were collected; the only stochastic element entering the calculations refers to the sampling scheme or the assignment procedure. The second basic dichotomy, which is particularly important for inferential statistics regarding single-case experimental data, refers to the distinction between inference for one single-case experiment and inference for replications of single-case experiments. Crossing these two dichotomies results in a two-by-two classification of statistical techniques.

A prototypical example of *model-based inference* for one single-case experiment is interrupted time series analysis (Tarlow & Brossart, 2018). For example, testing the effects of an exposure treatment in patients with chronic low back pain with increased pain-related fear, a sequential cross-over A-B phase design was applied (Vlaeyen et al., 2001). In two patients, exposure was followed by graded activity as the control condition, and in two other patients the order of treatments was reversed. The authors fitted an autoregressive time series model to each subject for each outcome measure, adjusting for possible background trend and autocorrelation. Data demonstrated that exposure (and not graded activity) made each individual confront rather than avoid specific physical movements, which resulted in reduced self-reported disability. Prototypical examples of *design-based inference* for single-case experiments are randomization tests and confidence intervals based on randomization test inversion (Michiels, Heyvaert, Meulders, & Onghena, 2017). A typical example is the outcome study of the therapist-aided exposure treatment for women with lifelong vaginismus (Ter Kuile et al., 2009). A replicated single-case A-B–phase design was used. A no-treatment baseline period was contrasted with exposure, with random determination of the start of the exposure phase. The randomization tests on the daily measures revealed that, compared with the baseline period, the intervention was successful in increasing the frequency of intercourse in 9 of the 10 women treated. The use of *multilevel models* (Moeyaert, Ferron, Beretvas, & Van den Noortgate, 2014) and *between-case standardized effect size analyses* (Hedges, Pustejovsky, & Shadish, 2012; Solmi & Onghena, 2014) can be considered as examples of model-based inference for replications of single-case experiments. A multilevel approach was applied using sequential AB-phase designs in 27 youth with chronic pain receiving exposure-based cognitive-behavioral treatment (Simons et al., 2020).Treatment phase was superior to the no-treatment randomized baseline phase for avoidance, pain acceptance, and pain intensity, whereas fear and pain catastrophizing did not improve until the follow-up phase. Finally, combination methods have been proposed that can be considered as examples of design-based inference for replications of single-case experiments (Levin, Ferron, & Gafurov, 2018; Solmi & Onghena, 2014). The point to make here is that the four approaches represent four different perspectives on statistical inference, each with their own statistical assumptions and inferential ambitions.

Because the number of viable and valid statistical techniques for the analysis of single-case experimental data might be overwhelming, it is convenient to have accessible tutorials and easy-to-use software available (Houle, 2009; Manolov & Moeyaert, 2017; McGill, 2017). An overview of user-friendly software, with special reference to recently developed free Shiny webapps, is maintained by Rumen Manolov of the University of Barcelona at the Open Science Framework (see https://osf.io/t6ws6/). In conclusion, it turns out that Hartmann's round holes need round pegs, and that the round pegs are now practicable and easily available in free online statistical software.

Striving for Level 1 Evidence in the Conduct and Report of Single-Case Experimental Designs

The frequency of single-case intervention studies in clinical psychology has increased exponentially over the past 50 years. In the neurorehabilitation field alone, the number of publications has grown from a smattering of studies prior to the 1970s to more than 1,550 today (Tate, McDonald, Moseley, Perdices & Togher, 2019). Evidence suggests, however, that many single-case studies are inadequate, both in terms of their scientific quality (Maggin, Briesch, & Chafouleas, 2013; Reichow, Barton, & Maggin, 2018; Tate, Perdices, McDonald, Togher, & Rosenkoetter, 2014) and adequacy of reporting (Shamseer et al., 2015; Tate et al., 2016b). Because well-designed single-case studies have the potential to provide Level 1 evidence, poorly designed and poorly reported studies miss the opportunity to furnish such a high level of evidence.

Some two decades ago, a landmark article was published that was to become a model for the reporting of clinical trials (Begg et al., 1996). The authors observed "a wide chasm between what a trial should report and what is actually reported in the literature" (p. 637). Thus, the CONSORT[2] Statement was born. It was taken up immediately and quickly applied to

[2] CONSORT=CONsolidated Standards Of Reporting Trials

other research methodologies. Since then, two reporting guidelines, developed using CONSORT procedures, are now available for SCEDs: the CONSORT Extension for N-of-1 Trials (CENT; Shamseer et al., 2015; Vohra et al., 2015) and the Single-Case Reporting guideline In BEhavioral interventions (SCRIBE; Tate et al., 2016a) developed for the wider range of SCEDs used in the behavioral sciences. Both guidelines are similarly structured. For example, the SCRIBE checklist contains 26 items in six domains and covers all aspects of a report: title and abstract ($n = 2$ items), introduction ($n = 2$), method ($n = 14$), results ($n = 3$), discussion ($n = 3$), and documentation ($n = 2$). The SCRIBE has been endorsed by the American Psychological Association in its journal article reporting standards (Appelbaum et al., 2018).

Prior to publication of the CENT and SCRIBE, authorities in SCED methodology had raised concerns about the scientific quality of single-case studies. In order to provide a high level of evidence, SCEDs need to control for threats to internal validity (e.g., risk of bias). A seminal article delineated a set of 21 "quality indicators" (Horner et al., 2005), and an important next step drew upon those quality indicators to produce standards of design and evidence (Kratochwill et al., 2013). Critical appraisal scales offer a convenient method to evaluate the scientific quality of a study, and a number of instruments are designed for single-case methods (Tate & Perdices, 2019). An example of a critical appraisal tool for single-case designs is the Risk of Bias in N-of-1 Trials Scale (RoBiNT; Tate et al., 2013). This 15-item scale contains two subscales relating to internal validity (seven items, such as randomization, blinding, interrater reliability) and external validity and interpretation (eight items, such as baseline characteristics, therapeutic setting, data analysis, replication). Items are rated on a three-point scale, the internal validity items consistent with current design standards. Psychometric properties of the scale are good, with evidence of construct (discriminant) validity, and high interrater reliability (ICC range = .87–.93). The scale can be used to critically appraise a published report as well as designing a study at the planning stage. The SCRIBE/CENT checklist and RoBiNT scale are complementary tools to assist users of single-case methods to plan, conduct, and report work at a high standard.

Limitations

SCEDs also have their limitations. They are less appropriate when acute and directed care is needed such as in emergency departments. Also, when the researcher is interested in effectiveness of a treatment over the long term (e.g., 1-year follow-up) it is more challenging to establish experimental contrasts in a SCED. Perhaps most important, each trial is only applicable to that particular individual. However, it is possible to combine data from replicated designs in different individuals and enhance external validity, which can be conceived as a gradient of successful replications of a particular therapeutic approach for a particular health problem (Punja, Xu et al., 2016b). Moreover, it is striking to observe sometimes better outcomes in the SCEDs as compared to RCTs. This may be due to the idiosyncratic outcome measures employed, which may be more sensitive to change than the generic measures used in the RCTs (Vlaeyen, Morley, Linton, Boersma, & De Jong, 2012).

Current Opportunities and Challenges

The shortcomings of group-based approaches including RCT's, the growing attention and support for individual-level analyses, and recent developments in SCED contribute to a paradigm shift in behavioral health. However, dissemination, implementation, and further development of the approach are complex processes. We have identified a number of key opportunities and challenges that should be discussed, utilized, and addressed to enhance the use and utility of SCED.

Citizen Scientists

Today access to health information is overwhelmingly abundant. People can easily search online and find relevant information regarding their symptoms, conditions, and prognosis as well as the suggested treatment and its evidence base. In consequence, citizens of today can be active partners in designing and providing their health care. In fact, citizen involvement in health-care innovations is to an important extent considered critical, and frequently requested by funding agencies as well as patient organizations and hospitals. Moreover, self-help books and similar resources have already made the distinction between health care and self-care less clear. In some areas, the same (or highly similar) interventions as delivered by the health-care provider can be accessed through a book, an app, or a peer-support group. Moreover, individuals often try to test their own methods, inspired by scientific or other sources, which may be more or less likely to result in favorable outcomes. Similar to providing citizens with tools for behavior change, we can provide them a platform or method for testing, monitoring, and evaluating effects. This may be a powerful way of increasing the awareness and utility of a scientific approach. With this type of guidance, the individual is able to evaluate the benefits of interventions that may have more face validity than empirical support, and sometimes clear drawbacks in terms of costs and potential side effects.

With an easy-to-use SCED app for entering data and monitoring changes, the individual may improve his/her own understanding of the utility of that particular intervention. Technical developments facilitate these opportunities and provide new options to track symptoms, cognitions, and behavior in each time through ecological momentary assessment and

the now ubiquitous devices tracking location, movement, spatial orientation, sound, and light in phones and smart watches (Shiffman, Stone, & Hufford, 2008). These options enhance the scope incorporating aspects of prevention and public health (e.g., interventions to sleep better or move more) and they allow testing to what degree individual responses to interventions are dependent on context (Sniehotta, Presseau, Hobbs, & Araujo-Soares, 2012). Thus, the assumption is that the individual learns and develops because of the data collected, and the SCED may represent a powerful tool to support the citizen-scientist aspiration.

Engagement and Treatment Adherence

The SCED approach may provide a framework in which the individual is seen as "response-able," and even expected to contribute, to optimize the quality of the analysis and treatment. The health-care user is an agent who may find a purpose and meaning in being part of the investigation and decision process, both concerning the treatment plan and generating knowledge used to continuously improve knowledge and clinical routines. If such an approach is systematically used at a clinic, it may be relatively easy to initiate daily assessment prior to the start of treatment (baseline), and continue data collection, throughout treatment with this being perceived as meaningful. In other words, the psychological function of the participation and contribution may be that it adds value to the care and optimizes the quality and effect of the intervention. Thus, a candid invitation to self-monitor and collect idiosyncratic data may not only benefit research and health-care quality but may also foster meaningful involvement of health-care users that ultimately can enhance empowerment and health.

Accountability

Because of budgetary restrictions in most countries, there is increasing pressure on health-care providers by governments, regional health authorities, and health-care boards to meet certain goals or objectives. Even though these objectives are set through a collaborative process between governing bodies and concerned providers or organizations, they must be monitored on a regular basis (Denis, 2014). In order to provide useful feedback, such a monitoring of outcomes requires proper, adequate, and timely information. RCT's are commonly burdensome, and require the allocation of extensive financial resources, human capital, and logistics, which are beyond the capabilities of most health-care centers. As a consequence, trials that lack the optimal resources cannot be conducted or are performed with compromises at the expense of scientific rigor. Therefore, SCEDs offer a feasible alternative that will not only serve the accountability request from governmental bodies but will also provide helpful information within the health care user–provider collaboration. In our experience, individuals usually appreciate the ability to share information about the dynamics of their own beliefs, concerns, and life goals by employing idiographic assessment methods with credible and partly self-generated items.

Making Use of Aggregated Data: Creating Data Repositories

The SCED approach focuses on the individual and her uniqueness. As such, SCED shares the aspiration of personalized medicine to allow for individualized assessment with health-care decisions, practices, and interventions to be tailored to the individual based on their predicted response. Existing digital systems provide opportunities to collect several individuals' data in a data repository of use for both clinical and research purposes. The possibility of aggregating data from different individuals over time will provide opportunities to not only validate findings in specific treatments or studies, but also to clarify differences in trajectories and change mechanisms. Such repositories would also be instrumental in establishing the external validity of the treatment. Every successful replication adds evidence to the generalizability of a specific treatment approach. Hence, the SCED approach together with the digital tools available to achieve and store data provides new opportunities to refine theories and improve treatments. As with all registers, such a data repository may also be used to compile outcomes across individuals within a clinic or assess the relative quality and success of different clinics, with more detailed information allowing for more adequate comparisons of clinics, interventions, or cohorts of health-care users. Finally, meta-analytic procedures can be used for the aggregation of replicated single-case experiments, which could help generate overall conclusions about a certain intervention (Onghena et al., 2018).

Dissemination of SCED

The SCED approach can be readily applied in clinical research and practice, but unfortunately many are not aware of its existence or its benefits. Moreover, knowledge about the SCED approach is today largely present among researchers, thus bridging the gap between researchers and clinicians is an important challenge and key to dissemination. But even when knowledge exists, the scientist-practitioner may possess a general orientation towards the nomothetic approach. A potential obstacle is the widespread belief that most well-defined conditions require a standardized and evidence-based intervention "protocol," seemingly providing a sense of safety and security against potential therapeutic errors. For example, the psychological literature offers a wide range of handbooks on treatment protocols for almost all DSM diagnoses (Clark, 2011). As mentioned earlier, classification-based practices assume that differences between members of the same class are

minimal, and hence would benefit from the same treatment protocol. This may be at the cost of the idiosyncratic needs of the individual, something that may not always be recognized by health-care providers. In order to boost the implementation of SCEDs, the International Collaborative Network for N-of-1 Clinical Trials and Single-Case Experimental Designs (ICN) was established in 2017. The ICN is uniquely positioned to facilitate a range of activities to further the science of individualized research, and to date there are over 150 ICN members in 12 countries (Nikles & McDonald, 2017).

Training

To date, training in research methods for most, if not all, medical and health-care professions emphasizes group-level approaches such as the RCT. The students of today are the clinicians and researchers of tomorrow, which implies that by exclusively teaching traditional RCT methods, we perpetuate a lack of knowledge about the SCED approach. A key challenge to increase the scope of SCED in training is that education curricula are not easy to change. They are developed in alignment with guidelines from authorities to ascertain quality and standardization across universities. Moreover, teachers need to feel confident in describing the methods and have access to resources to provide examples and exercises that gradually increase the students' ability to understand and use the approach. There have been few available resources to guide teachers in setting up and running courses in SCED, but fortunately this is changing (Manolov, 2016). Building on this fledgling collection of resources is critical to increase uptake of training in this domain.

Concluding Remarks

This review summarized how SCEDs may be an approach better suited for clinical research, and offer a valuable approach to treatment that may close the long-existing gap between practitioners and scientists. In recent years, SCED technology has made immense progress, in terms of design possibilities, various visual and statistical analyses are available, and users may benefit from quality guidelines and guidelines for the reporting of studies using SCED. Given these developments, SCEDs have never been so accessible. Having mentioned this, RCTs will remain to be of importance in health-care research. Data from group-based RCTs may be used as a "place to start" in the treatment of one individual (Barlow et al., 2017). In addition, governmental bodies may not be primarily interested in individual's outcomes but more general outcomes, e.g., whether a certain treatment deserves to be supported or not, and if so, if it is cost-effective. It should not be noted that, multiple replications, if conducted systematically, will aggregate data from multiple SCEDs, yielding several important outcomes: accountability to the individual seeking care, accountability to governmental bodies developing health-care policies, ecologically valid data for applied researchers, and data that can be used by educational and training centers to develop or refine their curricula.

Acknowledgements We wish to acknowledge the people who were instrumental for and during the symposium: the participants in the strategic meeting during day 1 (in addition to the authors of the article): Maria Tillfors, Ida Flink, Katja Boersma, Mike Kemani, Lance McCracken, Julia Glombiewski, and Allen Finley, as well as the attendants of the open meeting during day 2.

Availability of Data and Materials NA/This is a theoretical review article.

Compliance with Ethical Standards

Conflicts of Interest None

Research involving human participants and/or animals NA

Informed Consent NA

References

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *The American Psychologist, 73*(1), 3–25. https://doi.org/10.1037/amp0000191.

Baker, D. B., & Benjamin Jr., L. T. (2000). The affirmation of the scientist-practitioner. A look back at Boulder. *The American Psychologist, 55*(2), 241–247 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/10717972.

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

Barlow, D. H., Farchione, T. J., Bullis, J. R., Gallagher, M. W., Murray-Latin, H., Sauer-Zavala, S., et al. (2017). The unified protocol for transdiagnostic treatment of emotional disorders compared with diagnosis-specific protocols for anxiety disorders: A randomized clinical trial. *JAMA Psychiatry, 74*(9), 875–884. https://doi.org/10.1001/jamapsychiatry.2017.2164.

Baron, A., & Perone, M. (1998). Experimental design and analysis in the laboratory study of human operant behavior. In K. A. Lattal & M. Perone (Eds.), Handbook of research methods in human operant behavior (pp. 45–91). Plenum Press.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., et al. (1996). Improving the quality of reporting of randomized controlled trials: The CONSORT statement. *Journal of the American Medical Association, 276*(8), 637–639. https://doi.org/10.1001/jama.276.8.637.

Bernard, C. (1957). An introduction to the study of experimental medicine. Courier Corporation Dover. (Original published 1865)

Bulte, I., & Onghena, P. (2008). An R package for single-case randomization tests. *Behavior Research Methods, 40*(2), 467–478 https://doi.org/10.3758/BRM.40.2.467.

Bulté, I., & Onghena, P. (2012). When the truth hits you between the eyes: A software tool for the visual analysis of single-case experimental data. *Methodology, 8*(3), 104–114.

Campbell, B. A. (1956). The reinforcement difference limen (RDL) function for shock reduction. *Journal of Experimental Psychology, 52*(4), 258–262 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/13367349.

Clark, D. M. (2011). Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: The IAPT experience. *International Review of Psychiatry, 23*(4), 318–327. https://doi.org/10.3109/09540261.2011.606803.

DEcIDE, & Panel, M. C. N.-o.-G (2014). Design and implementation of n-of-1 trials: A user's guide (AHRQ Publication No. 13(14)-EHC122-EF). Retrieved from http://www.effectivehealthcare.ahrq.gov/N-1-Trials.cfm., Issue.

Denis, J. L. (2014). Accountability in healthcare organizations and systems. Healthc Policy, 10(Spec issue), 8–11. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/25305384

Dunsmoor, J. E., Martin, A., & LaBar, K. S. (2012). Role of conceptual knowledge in learning and retention of conditioned fear. *Biological Psychology, 89*(2), 300–305. https://doi.org/10.1016/j.biopsycho.2011.11.002.

Evans, J. G. (1995). Evidence-based and evidence-biased medicine. *Age and Ageing, 24*(6), 461–463. https://doi.org/10.1093/ageing/24.6.461.

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver & Boyd.

Fontanarosa, P. B., & DeAngelis, C. D. (2002). Basic science and translational research in JAMA. *Journal of the American Medical Association, 287*(13), 1728. https://doi.org/10.1001/jama.287.13.1728.

Grossman, J., & Mackenzie, F. J. (2005). The randomized controlled trial: Gold standard, or merely standard? *Perspectives in Biology and Medicine, 48*(4), 516–534. https://doi.org/10.1353/pbm.2005.0092.

Group, O. L. o. E. W (2011). The Oxford levels of evidence 2. Retrieved from www.cebm.net/index.aspx?o=5653.

Guyatt, G. H., & Jaeschke, R. (1990). N-of-1 randomized trials: Where do we stand? West J Med, 152(1), 67–68. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/2309475

Hartmann, D. P. (1974). Forcing square pegs into round holes: some comments on "An analysis-of-variance model for the intrasubject replication design.". *Journal of Applied Behavior Analysis, 7*(4), 635–638. https://doi.org/10.1901/jaba.1974.7-635.

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods, 3*(3), 224–239. https://doi.org/10.1002/jrsm.1052.

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*(2), 165–179.

Houle, T. T. (2009). Statistical analyses for single-case experimental designs. In D. H. Barlow, M. K. Nock, & M. Hersen (Eds.), Single case experimental designs: Strategies for studying behavior change (Vol. 3; pp. 271−305). Allyn & Bacon.

Jacob, E. K. (2004). Classification and categorization: A difference that makes a difference. *Library Trends, 52*(3), 515–540.

Kaptchuk, T. J. (2001). The double-blind, randomized, placebo-controlled trial: Gold standard or golden calf? *Journal of Clinical Epidemiology, 54*(6), 541–549 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/11377113.

Kazdin, A. E. (1983). Single-case research designs in clinical child psychiatry. *Journal of the American Academy of Child Psychiatry, 22*(5), 423–432 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/6630801.

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

Koch, G. G., & Gillings, D. B. (1984). Inference, design based vs. model based. In N. L. Johnson & S. Kotz (Eds.), Encyclopedia of statistical sciences (Vol. 4; pp. 84–88). New York, NY: Wiley.

Kratochwill, T. R., & Levin, J. R. (1992). Single-case research design and analysis: New directions for psychology and education. Lawrence Erlbaum Associates.

Kratochwill, T. R., Hitchcock, J. R., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). Single-case designs technical documentation. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/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*(1), 26–38.

Kronish, I. M., Cheung, Y. K., Shimbo, D., Julian, J., Gallagher, B., Parsons, F., & Davidson, K. W. (2019). Increasing the precision of hypertension treatment through personalized trials: A pilot study. *Journal of General Internal Medicine, 34*(6), 839–845. https://doi.org/10.1007/s11606-019-04831-z.

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2018). Comparison of randomization-test procedures for single-case multiple-baseline designs. *Developmental Neurorehabilitation, 21*(5), 290–311. https://doi.org/10.1080/17518423.2016.1197708.

Maggin, D. M., Briesch, A. M., & Chafouleas, S. M. (2013). An application of the What Works Clearinghouse Standards for evaluating single-subject research: Synthesis of the self-management literature base. *Remedial and Special Education, 34*(1), 44–58.

Manolov, R. (2016). List of single-case data analysis software tools. Open Science Framework. https://osf.io/t6ws6/#!

Manolov, R., & Moeyaert, M. (2017). How can single-case data be analyzed? Software resources, tutorial, and reflections on analysis. *Behavior Modification, 41*(2), 179–228. https://doi.org/10.1177/0145445516664307.

McGill, R. J. (2017). Single-case design and evaluation in R: An introduction and tutorial for school psychologists. *International Journal of School and Educational Psychology, 5*, 39–51. https://doi.org/10.1080/21683603.2016.1173610.

McQuaid, E. L., Aosved, A. C., & Belanger, H. G. (2018). Integrating research into postdoctoral training in health service psychology: Challenges and opportunities. *Training & Education in Professional Psychology, 12*(2), 82–89.

Michiels, B., Heyvaert, M., Meulders, A., & Onghena, P. (2017). Confidence intervals for single-case effect size measures based on randomization test inversion. *Behavior Research Methods, 49*(1), 363–381. https://doi.org/10.3758/s13428-016-0714-4.

Mirza, R. D., Punja, S., Vohra, S., & Guyatt, G. (2017). The history and development of N-of-1 trials. *Journal of the Royal Society of Medicine, 110*(8), 330–340. https://doi.org/10.1177/0141076817721131.

Moeyaert, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2014). From a single-level analysis to a multilevel analysis of single-case experimental designs. *Journal of School Psychology, 52*(2), 191–211. https://doi.org/10.1016/j.jsp.2013.11.003.

Moore, D. S., & Notz, W. I. (2017). Statistics: Concepts and controversies (9th ed.). Freeman.

Morley, S. (2018). *Single case methods in clinical psychology: A practical guide*. New York, NY: Routledge.

Morris, Z. S., Wooding, S., & Grant, J. (2011). The answer is 17 years, what is the question: Understanding time lags in translational research. *Journal of the Royal Society of Medicine, 104*(12), 510–520. https://doi.org/10.1258/jrsm.2011.110180.

Nikles, J., & McDonald, S. (2017). International Collaborative Network (ICN) for N-of-1 Clinical Trials and Single-Case Experimental Designs. https://www.nof1sced.org/.

Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *The Clinical Journal of*

Pain, 21(1), 56–68 discussion 69–72. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15599132.

Onghena, P., Michiels, B., Jamshidi, L., Moeyaert, M., & Van den Noortgate, W. (2018). One by one: Accumulating evidence by using meta-analytical procedures for single-case experiments. *Brain Impairment, 19*(1), 33–58.

Parker, R. I., & Vannest, K. (2009). An improved effect size for single-case research: nonoverlap of all pairs. *Behavior Therapy, 40*(4), 357–367. https://doi.org/10.1016/j.beth.2008.10.006.

Pavlov, I. (1927). Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex. (Translated and edited by G.V. Anrep ed.). New York. Dover Publications.

Petersen, S., Van Staeyen, K., Vogele, C., von Leupoldt, A., & Van den Bergh, O. (2015). Interoception and symptom reporting: disentangling accuracy and bias. *Frontiers in Psychology, 6*, 732. https://doi.org/10.3389/fpsyg.2015.00732.

Pothos, E. M., & Chater, N. (2005). Unsupervised categorization and category learning. *Quarterly Journal of Experimental Psychology A, 58*(4), 733–752 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/16104104.

Punja, S., Bukutu, C., Shamseer, L., Sampson, M., Hartling, L., Urichuk, L., & Vohra, S. (2016a). N-of-1 trials are a tapestry of heterogeneity. *Journal of Clinical Epidemiology, 76*, 47–56. https://doi.org/10.1016/j.jclinepi.2016.03.023.

Punja, S., Xu, D., Schmid, C. H., Hartling, L., Urichuk, L., Nikles, C. J., & Vohra, S. (2016b). N-of-1 trials can be aggregated to generate group mean treatment effects: A systematic review and meta-analysis. *Journal of Clinical Epidemiology, 76*, 65–75. https://doi.org/10.1016/j.jclinepi.2016.03.026.

Reichow, B., Barton, E. E., & Maggin, D. M. (2018). Development and applications of the single-case design risk of bias tool for evaluating single-case design research study reports. *Research in Developmental Disabilities, 79*, 53–64. https://doi.org/10.1016/j.ridd.2018.05.008.

Richards, S. B. (2019). Single subject research: Applications in educational settings. Cengage.

Rothwell, P. M. (2005). External validity of randomised controlled trials: "To whom do the results of this trial apply?". *Lancet, 365*(9453), 82–93. https://doi.org/10.1016/S0140-6736(04)17670-8.

Rubin, M., & Badea, C. (2012). They're all the same! … but for several different reasons: A review of the multicausal nature of perceived group variability. *Current Directions in Psychological Science, 21*(6), 367–372.

Schork, N. J. (2015). Personalized medicine: Time for one-person trials. *Nature, 520*(7549), 609–611. https://doi.org/10.1038/520609a.

Senn, S. (2018). Statistical pitfalls of personalized medicine. *Nature, 563*(7733), 619–621. https://doi.org/10.1038/d41586-018-07535-2.

Shamseer, L., Sampson, M., Bukutu, C., Schmid, C. H., Nikles, J., Tate, R., et al. (2015). CONSORT extension for reporting N-of-1 trials (CENT) 2015: Explanation and elaboration. *BMJ, 350*, h1793. https://doi.org/10.1136/bmj.h1793.

Shapiro, M. B. (1966). Generality of psychological processes and specificity of outcomes. *Perceptual and Motor Skills, 23*(1), 16. https://doi.org/10.2466/pms.1966.23.1.16.

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. Annu Rev Clin Psychol, 4, 1–32. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/18509902

Sidman, M. (1952). A note on functional relations obtained from group data. *Psychological Bulletin, 49*(3), 263–269 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/14930162.

Sidman, M. (1960). *The tactics of scientific research*. New York, NY: Basic Books.

Simons, L., Vlaeyen, J. W. S., Declercq, L., Smith, M. A., Beebe, J., Hogan, M., Li, E., et al. (2020). Avoid or engage? Outcomes of graded exposure in youth with chronic pain using a sequential replicated single-case randomized design. *Pain, 161*(3), 520–531. https://doi.org/10.1097/j.pain.0000000000001735.

Smith, J. D. (2012). Single-case experimental designs: a systematic review of published research and current standards. *Psychological Methods, 17*(4), 510–550. https://doi.org/10.1037/a0029312.

Smith, L. S., & Wilkins, N. (2018). Mind the gap: Approaches to addressing the research-to-practice, practice-to-research chasm. *Journal of Public Health Management and Practice, 24*(Suppl. 1), S6–S11. https://doi.org/10.1097/PHH.0000000000000667.

Sniehotta, F. F., Presseau, J., Hobbs, N., & Araujo-Soares, V. (2012). Testing self-regulation interventions to increase walking using factorial randomized N-of-1 trials. *Health Psychology, 31*(6), 733–737. https://doi.org/10.1037/a0027337.

Solmi, F., & Onghena, P. (2014). Combining p-values in replicated single-case experiments with multivariate outcome. *Neuropsychological Rehabilitation, 24*(3–4), 607–633. https://doi.org/10.1080/09602011.2014.881747.

Solomon, B. G. (2014). Violations of assumptions in school-based single-case data: Implications for the selection and interpretation of effect sizes. *Behavior Modification, 38*(4), 477–496. https://doi.org/10.1177/0145445513510931.

Sung, N. S., Crowley Jr., W. F., Genel, M., Salber, P., Sandy, L., Sherwood, L. M., et al. (2003). Central challenges facing the national clinical research enterprise. *Journal of the American Medical Association, 289*(10), 1278–1287. https://doi.org/10.1001/jama.289.10.1278.

Tajfel, H., & Wilkes, A. L. (1963). Classification and quantitative judgement. *British Journal of Psychology, 54*, 101–114 Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/13980241.

Tanious, R., De, T. K., Michiels, B., Van den Noortgate, W., & Onghena, P. (2019). Assessing consistency in single-case A-B-A-B phase designs. *Behavior Modification*. https://doi.org/10.1177/0145445519837726.

Tarlow, K. R., & Brossart, D. F. (2018). A comprehensive method of single-case data analysis: Interrupted Time-Series Simulation (ITSSIM). *School Psychology Quarterly, 33*(4), 590–603. https://doi.org/10.1037/spq0000273.

Tate, R. L., & Perdices, M. (2019). Single-case experimental designs for clinical research and neurorehabilitation settings. Routledge. London and New York.

Tate, R. L., Perdices, M., Rosenkoetter, U., Wakim, 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*(5), 619–638. https://doi.org/10.1080/09602011.2013.824383.

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*(3–4), 315–331. https://doi.org/10.1080/09602011.2013.875043.

Tate, R. L., Perdices, M., Rosenkoetter, U., Shadish, W., Vohra, S., Barlow, D. H., Horner, R., Kazdin, A., Kratochwill, T., McDonald, S., Sampson, M., Shamseer, L., Togher, L., Albin, R., Backman, C., Douglas, J., Evans, J. J., Gast, D., Manolov, R., Mitchell, G., Nickels, L., Nikles, J., Ownsworth, T., Rose, M., Schmid, C. H., & Wilson, B. (2016a). The Single-Case Reporting Guideline In BEhavioural Interventions (SCRIBE) 2016 statement. *Archives of Scientific Psychology, 4*(1), 1–9. https://doi.org/10.1037/arc0000026

Tate, R. L., Perdices, M., Rosenkoetter, U., McDonald, S., Togher, L., Shadish, W., Horner, R., Kratochwill, T., Barlow, D. H., Kazdin, A., Sampson, M., Shamseer, L., & Vohra, S. (2016b). The Single-Case Reporting Guideline In BEhavioural Interventions (SCRIBE) 2016: Explanation and elaboration. Archives of Scientific Psychology, 4(1), 10–31. https://doi.org/10.1037/arc0000027.

Tate, R., McDonald, S., Moseley, A., Perdices, M., & Togher, L. (2019). NeuroBite: NeuroRehab Evidence Resource. https://neurorehab-evidence.com/web/cms/content/home

Ter Kuile, M. M., Bulte, I., Weijenborg, P. T. M., Beekman, A., Melles, R., & Onghena, P. (2009). Therapist-aided exposure for women with lifelong vaginismus: A replicated single-case design. *Journal of Consulting and Clinical Psychology, 77*(1), 149–159. https://doi.org/10.1037/a0014273.

Todman, J. B., & Dugard, P. (2001). Single-case and small-n experimental designs: A practical guide to randomization tests. Mahwah, N.J Lawrence Erlbaum Associates.

van der Meulen, M. A., Anton, F., & Petersen, S. (2017). Painful decisions: How classifying sensations can change the experience of pain. *European Journal of Pain, 21*(9), 1602–1610. https://doi.org/10.1002/ejp.1061.

Vlaeyen, J. W., & Morley, S. (2005). Cognitive-behavioral treatments for chronic pain: What works for whom? *The Clinical Journal of Pain, 21*(1), 1–8.

Vlaeyen, J. W., de Jong, J., Geilen, M., Heuts, P. H., & van Breukelen, G. (2001). Graded exposure in vivo in the treatment of pain-related fear: A replicated single-case experimental design in four patients with chronic low back pain. *Behaviour Research and Therapy, 39*(2), 151–166. https://doi.org/10.1016/s0005-7967(99)00174-6.

Vlaeyen, J. W., Morley, S., Linton, S., Boersma, K., & De Jong, J. (2012). Pain-related fear: Exposure-based treatment for chronic pain. IASP Press.

Vohra, S., & Punja, S. (2019). A case for n-of-1 Trials. *JAMA Internal Medicine, 179*(3), 452. https://doi.org/10.1001/jamainternmed.2018.7166.

Vohra, S., Shamseer, L., Sampson, M., Bukutu, C., Schmid, C. H., Tate, R., et al. (2015). CONSORT extension for reporting N-of-1 trials (CENT) 2015 statement. *British Medical Journal, 350*, h1738. https://doi.org/10.1136/bmj.h1738.

Walburn, J., Sarkany, R., Norton, S., Foster, L., Morgan, M., Sainsbury, K., et al. (2017). An investigation of the predictors of photoprotection and UVR dose to the face in patients with XP: A protocol using observational mixed methods. *BMJ Open, 7*(8), e018364. https://doi.org/10.1136/bmjopen-2017-018364.

Affiliations

Johan W. S. Vlaeyen[1,2] · Rikard K. Wicksell[3] · Laura E. Simons[4] · Charlotte Gentili[3] · Tamal Kumar De[5] · Robyn L. Tate[6] · Sunita Vohra[7] · Salima Punja[7] · Steven J. Linton[8] · Falko F. Sniehotta[9] · Patrick Onghena[5]

1 Health Psychology Research, Faculty of Psychology and Educational Sciences, KU Leuven, Tiensestraat 102, 3000 Leuven, Belgium

2 Experimental Health Psychology, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, Netherlands

3 Department of Clinical Neuroscience, Psychology Division, Karolinska Institutet, Stockholm, Sweden

4 Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University School of Medicine, Palo Alto, CA, USA

5 Methodology of Educational Sciences, Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium

6 John Walsh Centre for Rehabilitation Research, The Kolling Institute of Medical Research, Faculty of Medicine and Health, The University of Sydney, Camperdown, Australia

7 Department of Pediatrics, University of Alberta, Edmonton, Alberta, Canada

8 Center for Health and Medical Psychology, School of Law, Psychology and Social Work, Örebro University, Örebro, Sweden

9 Institute of Health & Society, Newcastle University, Newcastle upon Tyne, UK