Evaluating new gerontechnologies: Proof of concept is necessary, but not sufficient

William. D. Kearns PhD[a]
James L. Fozard PhD[b]

[a]Department of Child and Family Studies, College of Behavioral and Community Sciences, University of South Florida, Tampa, Florida 33612, USA, E: kearns.william@gmail.com;
[b]School of Aging Studies, College of Behavioral and Community Sciences, University of South Florida, Tampa, Florida 33612, USA, E: james.fozard@verizon.net

W.D Kearns, J.L. Fozard. Evaluating new gerontechnologies: Proof of concept is necessary, but not sufficient. Gerontechnology 2016;14(3):139-145; doi:10.4017/gt.2016.14.3.007.00 **Background** At a minimum, evaluation of a gerontechnology requires two steps: (i) a proof of concept demonstration that the product can operate as intended; and (ii) a demonstration that targeted aging and aged users can, want to, and will use the product as intended to achieve the ends as advertised. The second step is the more difficult; the present article describes some practical procedures for measuring the effect of the product on the individual's behavior. **Methods** Single Case Experimental Designs (SCED) allow for efficient evaluation of a product by a small number of users; SCED are especially useful when one or very few prototypes are available for evaluation. **Results** The basic elements of SCEDs include a no-treatment baseline periods alternating with periods in which the gerontechnology is introduced, with the baseline phase representing a period in which the feature of the product being evaluated is disabled; measurable changes in user behavior with vs. without the product feature enabled constitutes the evaluation. Examples illustrating some ways the basic elements can be implemented are provided. **Conclusions** SCEDs provide a good alternative to the more expensive and complex Randomized Controlled Trials (RCTs) generally considered the 'gold standard' for evaluations.

Keywords: Single Case Experimental Designs, research methods, single subject designs

The first step in the evaluation of a new gerontechnology is a demonstration that the prototype device CAN work - the engineering proof of concept. The second is a demonstration that the device DOES work as intended - a more complex process. The two steps are the essential components of what can be a complex evaluation process ranging from the determination of perceived user need of, and interest in a product - sometimes by means of focus groups - to a study of the benefits of a product for various target groups. A full evaluation requires specification of the target population of users, control of extraneous factors, e.g., health status or use of mobility aids, and the specification of a comparison or control group. The Randomized Control Trial (RCT) is considered the best procedure for evaluation of new medical treatments, drugs and devices because it addresses all of those requirements.

A full RCT is not always necessary or appropriate to evaluate consumer products including most wearable or personal gerontechnologies. The core question of whether the product works as intended requires a demonstration that individual users will achieve the end for which it has been designed. The answer requires a demonstration that product use changes the targeted behavior, and that the user will voluntarily continue to use the product. An alternative to RCT is most appropriate when technical considerations include the availability and costs of the prototype to be evaluated and whether the prototype is new or a competitor of an existing product. The purpose of the present paper is to identify effective and relatively inexpensive approaches to evaluating new gerontechnologies - Single Case Experimental Design (SCED, also called Single Case Design - SCD)

Background and uses of SCEDs

To move beyond a proof of concept engineering prototype, hard evidence justifying further development and commercialization of a technology is needed. The hard evidence is a demonstration that the technology effectively changes user behavior in the anticipated direction. In the 1950's behaviorists such as B.F. Skinner and his colleagues advanced the radical notion that group statistical designs describe modal characteristics of individuals, but fit no one individual in particular[1]. They argued that to do truly relevant research on how to alter problem behaviors in real life situations, the focus must be on changing the behavior of the individual, and that it was essential to demonstrate that one had behavioral control by showing one could systematically undo a change in behavior just as efficaciously and

reliably as it was induced[1]. This was a radical concept for psychologists at the time to accept, since group design studies were considered as the only legitimate means to conduct investigations. However, there was emerging a great need to be able to reliably demonstrate clinically significant and replicable behavior change in a single individual over time as a means to build an evidence base leading to large-scale clinical trials with high generalizability.

Demonstrating that the introduction of a given gerontechnology can produce a reliable quantifiable change in a behavior (dependent variable) of a single individual is the first step in gathering generalizable evidence. Kratochwil and colleagues, in a whitepaper published by the US federal government's Institute of Education Sciences, have detailed best practice technical specifications for SCED which present the minimum criteria for acceptable designs with high internal validity, and provide evidence standards, including statistical evaluation, for assessing outcomes, categorizing outcomes into 'strong evidence', 'moderate evidence', or 'no evidence'[2].

Essential to any evaluation of the effects of a given gerontechnology is gathering longitudinal baseline data on the behavior under study for a sufficiently long interval to ascertain its stability. According to Kratochwil et al.[2] a minimum of 5 data points are required to meet this standard. The presence of a trend in the data requires extending the baseline period until the trend's long-term characteristics are fully understood before transitioning from the baseline phase (A) to the treatment phase (B). For example, if a gerontechnology designed to increase the frequency of a walking is introduced in phase B when the amount of walking already is clearly increasing during the baseline phase A, it threatens the internal validity of the design and may result in erroneous conclusions being drawn about the effects of the introduction of the gerontechnology on the behavior. Conversely, if the gerontechnology is intended to suppress walking and the long-term baseline trend is stably increasing, the suppressive effect of the gerontechnology on the upward trend can be assessed with somewhat greater confidence insomuch as the long-term trend is opposite to the predicted effect of the gerontechnology treatment introduced in the B phase. Similarly, a stable baseline with no trend means the gerontechnology can be introduced and the impact on the dependent variable assessed.

Reliability testing

Kratochwil et al. have strongly argued that for a 'Reversal' SCED (one in which the treatment is systematically alternated with a baseline phase) a sufficiently long baseline (A) phase (>4 data points), followed by at least 5 data points in the treatment (B) phase subsequent to a return to baseline (A) for >4 data points, and ending with a reintroduction of phase B for >4 data points is the minimum requirement for demonstrating a reliable effect of the treatment condition. Kratochwil et al. argue the following standards of evidence be used to assess the reliability of change in a measured dependent variable across SCED designs[3]:

(i) Change in slope: The dependent variable assumes a persistent change in slope from the baseline condition;

(ii) Change in level: The means of the distributions of data points for the baseline and gerontechnology conditions are visibly different;

(iii) Number of overlapping data points in the baseline and gerontechnology conditions: Few if any observations in the baseline (A) and gerontechnology (B) treatment conditions should overlap. The data from the baseline and gerontechnology conditions represent different distributions that presumably differ as a result of introducing the gerontechnology.

This last point in #iii deserves repeating insomuch as the three criteria are necessary but not sufficient evidence for presuming causality. Meeting the three criteria in a single introduction of a gerontechnology may be due entirely to a chance combination of circumstances. Kratochwil[2] has argued that minimally three alternating conditions are required to demonstrate behavioral control in that subject. Note the demonstration of control in one subject does not necessarily extend to a second subject; it says the obtained results coincided with the introduction of the treatment for that subject and that they were repeatable when a second instance of the B phase commenced. Confidence in the reliability of the treatment effect is increased when the gerontechnology is introduced and withdrawn, and each time the same response is achieved. Confidence in the generalizability of the results to other individuals is further increased through replication of the protocol across a total of 3 subjects while obtaining similar results in each case[2].

It should be noted that SCED might be considered highly controlled experiments where extraneous sources of variation are minimized. As a result, the data from SCEDs are very clean, and with sufficient replications a small group design repeated measures analysis of variance could validly be applied to statistically analyze the differences between the conditions. In those cases where substantive individual differences exist, the data may be transformed to remove that

source of variation through normalization, and render the effects of the gerontechnology treatment condition more apparent.

The application of a repeated measures analysis of variance to a single subject's data is inappropriate due to non-independence of observations, which can inflate the type 1 error rate. Parker et al.[4;p129] have argued for the use of non-parametric statistics to evaluate within subject sequential data in what they term a phase 'non-overlap analysis' and present no fewer than 8 methods for assessing the impact of a treatment on an individual subject's behavior. The non-overlap indices, they argue, are robust non-parametric approaches that make no assumptions about the distribution of the data. Depending upon the source, SCED data can be non-normal, skewed, and may exhibit inconstant score variance over time. The non-parametric approach, they argue, is more appropriate than computing means and standard deviations for such measures. It is beyond the scope of this paper to review all 8 methods; the reader is directed to Parker et al.'s excellent chapter for a full discussion. Two of the 8 methods presented by Parker and colleagues, the 'Non Overlap of All Pairs' (NAP) and the 'TauU Method' evaluate the relationship of each data point in the baseline phase against corresponding data points in the treatment phases, and by using the Mann Whitney Test and the Kendall Rank Correlation statistic, respectively, each approach generates two-tailed probabilities for evaluation of the experimental treatment.

Serial data from a single subject can be evaluated using time-series analysis, which evaluates the autocorrelation present in successive observations. Time series procedures test the autocorrelation present in successive observations gathered in the baseline conditions versus data from the same subject in the (experimental test) gerontechnology and provide confidence intervals regarding whether the observations derive from the same distribution. The gerontechnology condition in order to be effective should, in theory, reliably alter the autocorrelation of the data series gathered during the treatment condition compared to the data gathered during baseline. Time series analysis is quite powerful but requires numerous observations in each of the contrasting conditions to compute reliable estimates. The reader is directed to Gottman's excellent text on time series analysis of behavioral data for further information on this topic[5].

SCED's major strength

A major strength of SCED is the reduction of between-subjects variability through the use of 'subjects as their own control'; gathering data from the same subject over time as experimental conditions are systematically introduced and removed. SCED requires multiple observations over time under a minimum of two conditions, one of which is a baseline phase (A) during which observations of the dependent variable (our subject's behavior) are made under ambient or 'baseline' conditions and an 'experimental' or 'treatment' phase (B) wherein the effects of an independent variable (associated perhaps with some new gerontechnology) on the dependent variable are assessed.

Observations are required to be gathered systematically at regular intervals uninfluenced by the study condition in which they were recorded. Smith[6] in his 2012 review of 409 published SCED studies published across 134 journals concludes a minimum of 3 to 5 data points are required in each phase in order to evaluate the effect of an intervention. Fewer than 3 data points preclude a visual examination of the data for the presence of a trend in the direction of the hypothesized effect (an undesirable condition which can confound interpretation of the intervention results). The following examples illustrate the two basic variants of SCED and when their use is appropriate.

Examples

Example 1: Reversal SCED

A gerontechnologist creates a persuasive technology that s/he thinks will encourage walking (physical exercise), a behavior strongly correlated with improved health in older adults. The device, which is worn on a belt, can be switched to passively record walking, or can deliver encouraging prompts. Because of the great expense incurred to produce the device, only a single prototype exists. So how do we gain confidence that it will work as expected and increase walking?

In this instance a group RCT design is not feasible since the cost of replicating the device is too high to allow data from dozens or hundreds of subjects to be collected. The gerontechnologist decides to use a Reversal SCED and first seeks older individuals who fall in the at-risk group s/he is interested in – the primary user group s/he wishes to encourage to walk more. In the first baseline phase (A1) s/he assigns the prototype to a participant who is asked to wear it daily for 5 days while the total distance walked is measured, but no prompting intervention is provided to the subject.

After 5 days of recording, the gerontechnologist switches on the prototype's audio so that each episode of walking over the next 5 days receives an encouraging prompting message while walking data is gathered (B1). Next, the gerontech-

nologist switches off the encouraging prompting and records 5 final days of data as a second baseline (A2) to determine if the prompt is essential to maintaining the walking or if the behavior will 'extinguish' when the prompt is removed (the 'reversal' in the SCED design). After 5 days of recording in A2, the gerontechnologist switches on the prototype's audio so that each episode of walking over the next 5 days receives an encouraging prompting message while walking data is gathered (B2). In this example sufficient data for 3 subjects (and a potentially publishable study) could be gathered in 3x20=60 days or about 2 months; possibly in half that time if only a half day's walking data is gathered such that two subjects could generate walking data in the morning and afternoon on each day. The results of our hypothetical first subject's data appear in *Figure 1*.

The reader should note that day-to-day variability in walking during baseline A1 was low and that our subject was relatively sedentary, walking less than 3.0km/day. During the first treatment phase (B) the gerontechnology was introduced and the prompting resulted in an immediate increase to 4.0km/day with a peak at around 5.1km/day when the prompting was switched off and baseline conditions reinstated. During the second (follow-up) baseline when prompting was switched off, the participant's behavior changed significantly but remained high in the first day of the second baseline, indicating some carryover effects which diminished as the effects of prompting wore off over the next few days.
It should be noted that the mean of the gerontechnology phase (B1) is higher than the preceding and follow-up baseline (A2) and that the lines do not overlap, further supporting the notion that the gerontechnology had an immediate clinically significant effect on this subject's behavior. The change in slope and the lack of overlap with baseline phases are two accepted criteria used to visually determine if behavior change actually occurred as a result of the intervention[2]. The reintroduction of the gerontechnology treatment (B2) reveals an immediate sustained increase in walking behavior with no overlapping datapoints and supports the contention that the gerontechnology increases walking in this subject. Other criteria have been established to deal with overlapping data points of adjacent phases, for example Parker et al.[4] have advanced six additional criteria for assessing change that include measures such as the percent of data exceeding the median trend, the percent of treatment data exceeding the baseline median, etc. It is beyond the scope of the current paper to address each of these criteria in detail, and the reader is encouraged to consult their excellent chapter.

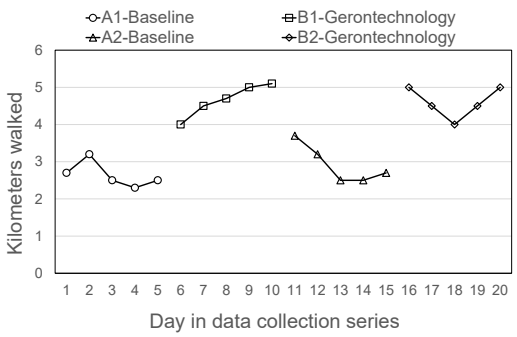

Figure 1. Single subject ABAB reversal design; Kilometers walked daily in four conditions on 20 consecutive days

One threat to the internal validity of many SCED studies is the lack of random assignment of subjects, treatments or phases. SCED methodologists have been more recently introduced randomization methods into their designs to strengthen their internal validity and broaden their acceptance by the academic community[7].

The previous illustrations provide a fairly simple set of conditions in which the effect of introducing a treatment is clear-cut. The need for obtaining a stable baseline prior to introducing a treatment is paramount. Trends in the baseline proceeding in the expected direction of the intervention must be controlled; one approach to controlling this source of potential confounding is to extend the baseline so the nature of any trend is understood (e.g. the trend may actually be a cyclical function that will return to initial levels given enough time, at which time the treatment may be applied). If the researcher suspects that the cyclical function is affecting treatment phase results, the treatment phase can be extended to determine if the natural cyclical behavior of the dependent measure will negate the effect of the intervention. If the influence of the cycle on the intervention data is minimal, the investigator may conclude the intervention was successful.

It should also be noted that the Baseline1-Treatment1-Baseline2-Treatment1 model can be expanded to include a repeat administration of the treatment (i.e. Baseline1-Treatment1-Baseline2-Treatment1-Baseline3, etc.), or some entirely different treatment may be tested (perhaps the gerontechnology discussed earlier can provide either a visual prompt or an auditory prompt and the investigator wishes to learn which of the two is more effective, or an entirely different gerontechnology could be tested during the second treatment phase labeled "C1", etc.).

Example 2A: Multiple baselines across behaviors
It is possible to gather information on the effect of interventions targeting different behaviors within the same individual. In this more elaborate instance of SCED, confidence grows with the repeated demonstration that a set of behaviors within a given person can be systematically modified, and power to generalize outcomes across individuals is gained by repeating the protocol across successive individuals. This type of SCED is appropriate when it is anticipated that the effects of the intervention are permanent, as may be the case when learning is involved.

Let us consider another example: our gerontechnologist has a single wearable device s/he thinks is capable of modifying health behaviors by persuasive prompting. The health behaviors s/he wants to increase in each of the 3 research subjects include walking, medication adherence, and attending clinical sessions. S/he gathers baseline data over several days on walking, medication compliance, and clinical session attendance and then enables prompting for only one of the behaviors s/he intends to modify (walking), while observing whether the prompted behavior increases, and more importantly whether the unprompted behaviors (medication adherence and session attendance) do not change. After a period equivalent to the baseline interval with at least 5 data points (In *Figure 2* as a mean value for clarity), the change in the first behavior under study is assessed. If the prompted behavior increases (defined as a clearly visible change in slope and/or a change in mean level that is determined to be significant using overlap analysis), prompting is enabled on the second behavior under study (and maintained on the first behavior) until the effects of the treatment on the second behavior are apparent and deemed significant using overlap analysis - the gerontechnologist evaluates medication compliance in the same manner as walking was assessed (presence of a change in slope and/or a change in mean level using overlap analysis). The process is then repeated for the third and final behavior, session attendance, while holding prompting constant for the other two behaviors. *Figure 2* presents hypothetical data for this scenario illustrating the staggered introduction of prompting for the individual subject. Confidence about the generality of the effect is increased by successively replicating the procedure on the remaining 2 subjects.

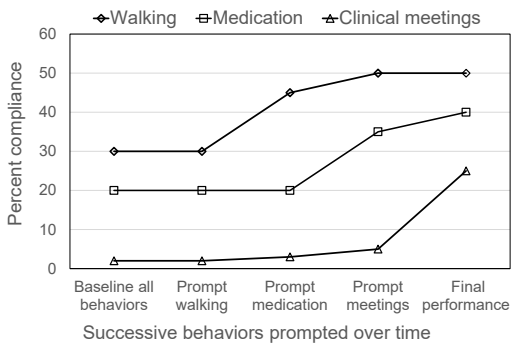

Figure 2. Multiple baseline across behaviors (Subject 1); Percentage compliance for three behaviors as an effect of staggered introduction of prompting across time

The multiple baselines across behaviors SCED can be expanded to as many behaviors as are practical to measure. This variant of SCED has great potential for smart home environments where an investigator may want to study a variety of behaviors in the same subject at the same time but resource availability limits testing to a single individual at a time. Note that the behaviors themselves must be independent. If while prompting walking, medication compliance changes significantly and does not return to baseline levels, the assumption of independence of behaviors may be violated and the design may not be appropriate since the two behaviors may be moderated by some third, unsuspected factor.

In multiple baselines across behaviors SCED, if the behavior change is expected to be irreversible (possibly due to learning), a reversal phase in which the treatment (in this case prompting) is withdrawn may be impractical. However, this does not preclude the adding of reversal phases if permanent behavior changes are not expected. The multiple baselines across behaviors SCED allows systematic examination of specific gerontechnologies' abilities to change multiple behaviors in a given individual, and is a very efficient design in terms of conservation of limited study resources and time.

Example 2B: Multiple baseline across subjects
One of the more difficult issues to deal with in SCED studies is the relatively permanent effect of learning, whose influences don't simply disappear when the intervention is switched off. Let us postulate for a moment that in our first example (ABAB SCED), turning on the prompt led to an immediate increase in walking in B1 that did not reverse when the gerontechnology was switched off in A2. How can our investigator be confident that the gerontechnology caused the increase in walking when its removal did not result in a decrease in the behavior? The first step is to extend the post-treatment baseline until the investigator is reasonably convinced that the increase in walking associated with the intervention really does not decrease back to the pre-intervention level given more time. This means that the

length of the post-intervention baseline may vary within the same subject or across subjects.

A solution to evaluating permanent effects of treatments is presented by the multiple baseline SCED, wherein the introduction of a treatment across subjects (or across behaviors in the same subject) is staggered. Consider the multiple baseline design in which only two phases (baseline and treatment) are used, but the introduction of the treatment for each successive subject is contingent on the completion of the treatment phase for the prior subject. A disadvantage of the multiple baseline design is that the baseline phases for each successive subject gets longer, and so gathering extended baselines of many weeks before the treatment is introduced may prove onerous for the later subjects. 'Multiple baseline across subjects' SCED's may prove impractical in cases where only one prototype of a monitoring device exists, especially if that device is required to collect data on baseline and treatment phases for all subjects simultaneously for the duration of the entire SCED study. However, if an alternate method can be substituted to gather baseline data, the effects of the single existing gerontechnology device can be assessed for its impact by introducing it to each subject in succession, yielding a clear picture of the efficacy of the device for changing walking behavior in the example described. Again, overlap analysis would provide nonparametric statistical evaluation of the observed change.

Drawer time

The obverse of the problem of subjects' failure to return to baseline after intervention withdrawal concerns abandonment of a gerontechnology occurring well after the formal evaluation period. This issue was briefly touched on earlier in this article, but merits further discussion. If one thinks a subject is not complying with the protocol and not wearing the gerontechnology, an inactivity measure – 'drawer time' (the duration each day the gerontechnology goes unused) may be obtained by affixing a small digital motion recorder to the gerontechnology under evaluation that measures the amount of time the gerontechnology remained stationary in the subject's desk drawer each day during the study. Progressive noncompliance appears as an increase in 'drawer time' over days as the gerontechnology was progressively abandoned during the behavior maintenance phase.

This approach might also be employed to shed light on issues related to technology acceptance, especially in those cases where compliance may at first be strong and then progressively declines. One advantage of measuring 'drawer time' is that it offers a common metric to assess a wide variety of gerontechnologies. An example of technology abandonment is provided by hearing aids; modern digital hearing aids allow for selective amplification of frequencies, control for recruitment, and to some extent the suppression of background noise. However, earlier generations of hearing aids were more primitive; all sounds were amplified, resulting in a loud sound devoid of meaning. Users sought out the more primitive devices, but poor usability frequently resulted in abandonment. The issue of product abandonment is familiar to gerontologists performing longitudinal studies; because of selective dropout of research participants, the longitudinal and cross-sectional estimates of age-related changes in the variable of interest may diverge significantly over time.

Discussion

In this article we have argued that gerontechnology development has been held back by a dearth of studies demonstrating the proposed product's efficacy in changing older persons' behaviors in order to "help them achieve their life's ambitions". We have challenged the classical assumption that the 'gold standard' RCT is the only way to shed light on the viability of a new product, a classical assumption that is flatly incorrect.

Use of an RCT, given the constraint of a single expensive device, may require more than one lifetime to gather sufficient data to be acceptable to a journal's editor. The RCT may indeed be the gold standard, but a 'silver standard' is also warranted: SCED lend themselves well to the evaluation of gerontechnologies whose goal it is to change human behavior of an individual. SCED are robust, highly economical, address real world problems and have over a 50-year publication history in the behavioral sciences. Their simplicity and ease of interpretation have resulted in their acceptance by a growing body of researchers whose works number in the hundreds of publications in peer-reviewed ranked journals.

The choice to select a gerontechnology for further development into a commercial product should not be based only upon "it's a good idea and in theory it should work" but rather to provide empirical evidence that it actually does work. Given that this relatively straightforward and inexpensive methodology is available to gerontechnologists, is there a reason not to raise the bar and thus the stature of gerontechnology as a discipline by including empirical testing of our new products? The publication of empirical support for a novel technology benefits the developer by smoothing the way to commercialization, by addressing concerns of potential investors and thereby improving viability of gerontechnology programs worldwide.

Gerontechnology is a synthetic discipline involving the scientific foundations of both technology and gerontology. The differences in academic traditions and journals in gerontechnology's multiple disciplines create different professional career challenges for social scientists and engineers. When one's academic home is in the behavioral or social sciences, publication in peer-reviewed journals with high impact factors is required to become tenured at one's university. Engineers receive due credit from their academic institutions for publishing proof of concept articles where a single very expensive engineering prototype may be the result. However, for the behavioral or social scientist who wants to determine if the engineer's prototype can significantly impact people's lives (for example in the case of a one-off 'persuasive technology') how can a 'gold standard' Randomized Controlled Trial (RCT) large group design empirical study possibly be conducted using just a single expensive prototype device? Single case experimental designs offer a relatively low cost means for developers to build an evidence base demonstrating the effectiveness of their new gerontechnologies.

Acknowledgement
The authors wish to acknowledge the contributions of the reviewers of this manuscript whose insightful comments significantly improved the quality of the final work.

References

1. Ferster C, Skinner B. Schedules of reinforcement. Cambridge: B. F. Skinner Foundation;1957 (Reprinted 1997; ISBN 0-13-792309-0
2. Kratochwill TR, Hitchcock J, Horner RH, Levin JR, Odom SL, Rindskopf DM, Shadish WR. Single-case designs technical documentation. What Works Clearinghouse; 2010; http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=229; retrieved January 25, 2016
3. Kratochwill TR, Levin JR, editors, Introduction: An overview of single-case intervention research. In Kratochwill TR, Levin JR, editors, Single-case Intervention Research: Methodological and Statistical Advances. Washington: American Psychological Association; 2014; doi:10.1037/14376-001
4. Parker RI, Vannest KJ, Davis JL. Non-overlap analysis for single-case research. In: Kratochwill TR, Levin JR, editors, Single-case Intervention Research: Methodological and Statistical Advances. Washington: American Psychological Association; 2014; pp 127-151; doi:10.1037/14376-001
5. Gottman JM. Time-series analysis: A comprehensive introduction for social scientists. Cambridge: Cambridge University Press; 2009
6. Smith JD. Single-case experimental designs: A systematic review of published research and current standards. Psychological Methods 2012;17(4); doi:10.1037/a0029312
7. Kratochwill TR, Levin JR. Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. In: Kratochwill TR, Levin JR, editors, Single-case Intervention Research: Methodological and Statistical Advances. Washington: American Psychological Association; 2014; pp 53-89