Behavior Research Methods
2008, 40 (2), 467-478
doi: 10.3758/BRM.40.2.467

An R package for single-case randomization tests

Isis Bulté and Patrick Onghena
Katholieke Universiteit Leuven, Leuven, Belgium

Randomization tests are nonparametric statistical tests that obtain their validity by computationally mimicking the random assignment procedure that was used in the design phase of a study. Because randomization tests do not rely on a random sampling assumption, they can provide a better alternative than parametric statistical tests for analyzing data from single-case designs. In this article, an R package is described for use in designing single-case phase (AB, ABA, and ABAB) and alternation (completely randomized, alternating treatments, and randomized block) experiments, as well as for conducting statistical analyses on data gathered by means of such designs. The R code is presented in a step-by-step way, which at the same time clarifies the rationale behind single-case randomization tests.

When designing an experiment, researchers almost automatically conceptualize the design in terms of a group approach, which can include between-group or within-group comparisons. This is not surprising, because only this type of study is covered in most statistical and methodological courses. Furthermore, most journal articles and statistics handbooks are based on this traditional, large-*n* research methodology (see, e.g., Moore & McCabe, 2006; Ramsey & Schafer, 2002).

Although this approach is very useful, group designs also have their limitations. In some instances, single-case experimental designs can provide a viable alternative or supplement to group designs. For example, single-case designs may be preferred when generating pilot data in the early stages of a larger group study; when the research concerns rare types of experimental subjects; when examining such questions as "does this treatment work for this particular individual?"; when testing the generalizability of an average group effect to individual subjects; and, of course, when research funds are scarce, making it impossible to obtain enough subjects for a large-scale group study (Barlow & Hersen, 1984; Edgington & Onghena, 2007; Franklin, Allison, & Gorman, 1997; Kazdin, 1982).

In the form of *n*-of-1 clinical trials, single-case designs can also help bridge the gap between theory and practice (Lundervold & Belwood, 2000). Here, the patient undergoes different treatments in a randomized order, and thus acts as his or her own control (Guyatt et al., 1988; Guyatt et al., 1986). In this way, the optimal therapy for a particular patient can be identified quickly; an unnecessary prolonged treatment can be avoided; new information about the effects of treatment can be gathered systematically; and the patient can gain a sense of empowerment and control by being part of a study, potentially improving adherence to the therapy (Avins, Bent, & Neuhaus, 2005; Guyatt et al., 1990; Nikles, Clavarino, & Del Mar, 2005; Wegman et al., 2005; Wegman et al., 2003).

But what exactly are these "single-case research designs"? Christensen (2001) defines them as "designs that use only one participant or one group of individuals to investigate the influence of some experimental treatment condition" (p. 279). They should not be confused with case studies, which are a form of descriptive research characterized by the absence of a designed intervention (Backman & Harris, 1999; Kazdin, 1981a). Single-case designs are experiments with some kind of deliberate manipulation of the independent variable. Since such designs involve only one entity, all levels of the independent variable are administered to this entity, and repeated measures are taken (Onghena, 2005). The underlying rationale of single-case designs is thus similar to that of group designs: The effects of different levels of an independent variable on a dependent variable are studied (Kazdin, 2003).

A major advantage of experiments is that they allow us to infer causal relations, whereas nonexperimental research only allows for determining associations between variables (Onghena & Edgington, 2005). In a true experiment, however, not only should there be a deliberate manipulation of the independent variable, it should also be possible to draw valid inferences about treatment effects. Single-case research focuses on what Campbell (1957) called the *internal validity* of the study: The aim is to make inferences about a treatment's effect in a specific experiment with a specific subject. A researcher removes threats to internal validity by eliminating all competing explanations, and for this reason the incorporation of some kind of randomization is crucial (Todman & Dugard, 2001). In contrast to group studies, randomization cannot be accomplished by randomly assigning entities to treatments, because single-case designs include just one entity. However, measurement occasions instead can be randomly assigned to treat-

I. Bulté, isis.bulte@ped.kuleuven.be

ments. This removes history and maturation, the major threats to the internal validity of a study, by controlling for both known and unknown extraneous variables, thus eliminating possible time-related rival hypotheses (Edgington, 1975, 1996; Onghena & Edgington, 2005).

To improve external validity, systematic replications of single-case experiments are needed (Hayes, 1981). External validity is an issue not only in group studies, but also in single-case designs, because identifying interventions that extend across different conditions (different subjects, different settings, etc.) is important. The problems that are encountered in group research, such as the selection of subjects and the research conditions, are also relevant to single-case studies (Kazdin, 1981b). Although they are not within the scope of this article, two kinds of single-case replication studies are noteworthy: *simultaneous* and *sequential* replication designs (Onghena & Edgington, 2005). In simultaneous replication designs, replications are carried out at the same time—for example, in a multiple-baseline design across subjects, in which 2 or more subjects are involved simultaneously. This design can strengthen the internal as well as the external validity of a study (Ferron & Sentovich, 2002; Hayes, 1981; Koehler & Levin, 1998, 2000). In sequential replication designs, the replications are carried out one after the other and can be analyzed using meta-analytical procedures (see, e.g., Van den Noortgate & Onghena, 2003a, 2003b).

Major Types of Single-Case Experimental Designs

Two major types of single-case experimental designs exist—phase and alternation designs—into which the random assignment of measurement occasions to treatments can be incorporated. The choice between the two will depend, among other things, on the research questions and on practical feasibility.

Phase designs. In phase designs, comparisons are made within a time series (Hayes, 1981), and the subject's performance is evaluated over time across baseline (A) and intervention (B) phases (Kazdin, 2003). The baseline phase serves the same function as a no-treatment control group in group studies, with the difference that, in single-case designs, the comparisons are made within the same individual (Lundervold & Belwood, 2000). Phase designs can vary as a function of different factors, such as the order of the phases, the number of phases, and the number of conditions (Kazdin, 1982).

The simplest type of phase design is an *AB design*, in which all baseline measurements precede all treatment measurements (Edgington, 1996). The researcher starts by measuring the behavior of the subject repeatedly during the baseline phase. An intervention is then introduced, and the researcher keeps on recording the target behavior during the treatment phase. With the gathered information, the researcher examines the possible relationship between the intervention and the target behavior (Zhan & Ottenbacher, 2001). If the target behavior changes when the treatment is introduced, it becomes plausible that the intervention was responsible for this change (Lundervold & Belwood, 2000). However, this design can result in rather weak conclusions, since it is subject to many confounding variables (Barlow & Hersen, 1984).

In an *ABA withdrawal* (or *reversal*) *design*, the treatment is administered between two baseline phases (Edgington, 1996). By adding this third phase, the impact of the intervention can be determined with greater certainty, because it is unlikely that possible confounding factors will lose their effect following withdrawal of the intervention (Zhan & Ottenbacher, 2001). However, rejection of the null hypothesis in this design could also indicate an effect of withdrawal or a combined effect of treatment and withdrawal, rather than a treatment effect (Todman & Dugard, 2001). Another concern with this design involves an ethical issue: In clinical contexts, if the experiment ends with a baseline phase, the patient cannot benefit fully from the treatment, and in some instances can even be left in an undesirable state (Barlow & Hersen, 1984).

The addition of a second treatment phase (*ABAB design*) can help to deal with the problems of the withdrawal design. Because the experiment ends with—hopefully a successful—intervention, the client can profit from the benefits of the treatment. Also, because there are two occasions for demonstrating the treatment effect, stronger conclusions can be made. With each change in the data pattern, it becomes less likely that extraneous variables are the reason for the observed effect (Barlow & Hersen, 1984). As an illustration, consider the hypothetical example of an ABAB design presented by Onghena (1992). A graphical representation of his data set is shown in Figure 1.

When imagining that the data in this example stand for some kind of undesirable behavior in a 5-year-old, a mere visual inspection of the data pattern can lead to the tentative conclusion that the unwanted behavior is less present during the treatment phases than during baseline. This is not true for all measurement times, however, and there is also a certain amount of variation in the scores within each phase.

Of course, phase designs can become far more complex than the types given above—for instance, by the addition of other treatments (e.g., ABACAB) or by the inclusion of an interaction term to allow for investigation of additive and nonadditive effects of the different treatments (e.g., Barlow & Hersen, 1984; Franklin, Allison, & Gorman, 1997; Kazdin, 1982).

Alternation designs. When a frequent succession of the different conditions is possible, an alternation design can be opted for. Such designs can be used to compare the effects of different treatments, as well as to compare performance in treatment versus no-treatment conditions (Zhan & Ottenbacher, 2001). For example, Baplu (2005) used an alternation design to compare the effects of a placebo with those of methylphenidatum in an adolescent with concentration difficulties. Figure 2 shows the summed scores of three items relating to concentration difficulties, each measured twice a day on a scale from 0 to 3 (where 0 stands for *no difficulties*). Visual inspection of the data patterns for both conditions suggests that slightly more concentration problems appear in the placebo condition than in the treatment condition. A clear distinction between the two would occur, however, if the line of one condition were always above that of the other condition.

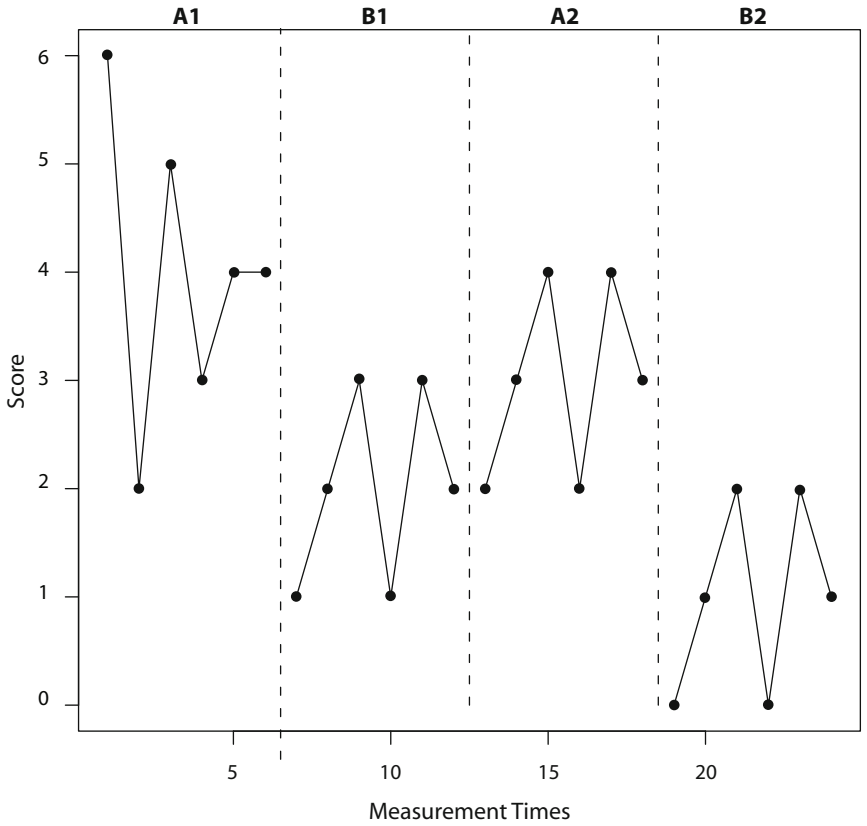

Figure 1. Hypothetical example of an ABAB design (Onghena, 1992).

In alternation designs, comparisons are made between time series for the different levels of the independent variable (Hayes, 1981). The basic strategy is the rapid alternation of two or more conditions within a single subject. However, rapid alternation does not necessarily mean frequent alternation or using short intermediate time periods; it can also mean that the subject/client potentially receives an alternative treatment every time he or she is seen by the researcher/therapist (Barlow & Hersen, 1984).

This design has some major advantages over phase designs: It does not require the withdrawal of treatment, which may result in a reversal of therapeutic benefits; there is no need for a baseline phase; and, because of the possibly very short "phases," comparisons can be made much more quickly. It also has some shortcomings, however. First, there is a risk of multiple-treatment interference, although random assignment can more or less overcome this problem (Barlow & Hayes, 1979; Zhan & Ottenbacher, 2001). Second, this type of design can only be used when a frequent alternation of the treatments is possible. It is therefore not useful for research on most types of psychotherapy (Onghena & Edgington, 2005).

As with phase designs, more complex alternation designs can also be constructed—for example, by combining the levels of two or more independent variables, which would result in a factorial single-case design (Edgington & Onghena, 2007).

Incorporation of Randomization in the Designs

Once the type of single-case design to use has been decided, some kind of randomization needs to be incorporated into it. The specific randomization schedules differ for the two main types of design.

Phase designs. In phase designs, the sequencing of phases is fixed, so randomization cannot be applied to the treatment order. The moment of phase change, however, can be randomly determined without altering the order of the treatments (Onghena & Edgington, 2005). For an experiment with k phases, $k - 1$ moments of phase change have to be selected. In an AB design with six measurement times, one intervention moment thus has to be determined randomly—the shift from baseline phase to treatment phase—resulting in the following possibilities:

AAAAAA	AABBBB
AAAAAB	ABBBBB
AAAABB	BBBBBB
AAABBB	

However, when we make no restrictions on the minimum number of measurement occasions per phase, we can end up with too few or no measurements for one of the phases. Therefore, it might be better to specify in advance the total number of measurement times and the minimum number

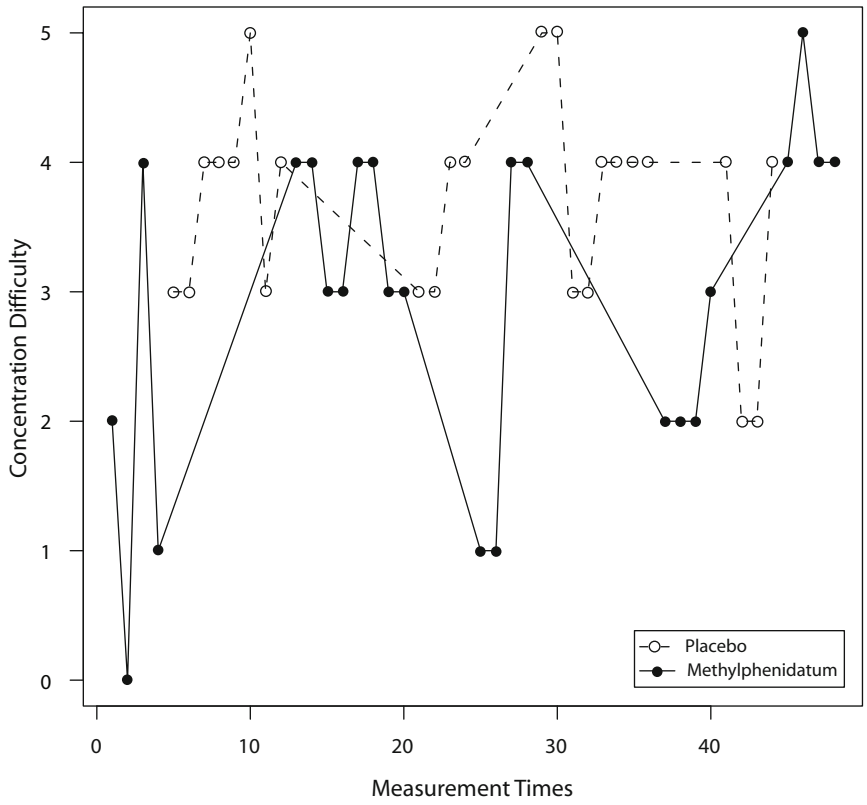

Figure 2. Alternation design with a treatment condition and a placebo condition to test the impact of methylphenidatum on the concentration difficulties of an adolescent (Baplu, 2005).

of observations per phase (Edgington, 1975; Onghena & Edgington, 2005). By including a minimum of two measurement occasions for each phase in the previous example, we have fewer possibilities, but the assignments make more sense: AABBBB, AAABBB, and AAAABB.

Alternation designs. The randomization schemes in alternation designs are easier to picture, because they resemble random assignment in the corresponding large-group designs. Instead of randomly assigning subjects to different groups, randomization is incorporated into the design by randomly assigning measurement times to different conditions (since there is only 1 subject). This actually comes down to randomly determining the treatment order.

The simplest randomization schedule in single-case alternation designs is a *completely randomized design*. The random assignment procedure here mirrors the one used when randomly assigning subjects to different groups for an independent-samples *t* test (or a one-way ANOVA) in a multisubject design (Edgington, 1996). As an example, let us consider six measurement occasions, of which three have to be assigned to Treatment A and three to Treatment B. The random assignment procedure then comes down to randomly choosing three measurement times for Treatment A, leaving the remaining three times for Treatment B. This equates to

$$\binom{6}{3} = \frac{6!}{3!3!} = 20$$

possible assignments:

AAABBB	BBBAAA
AABABB	BBABAA
AABBAB	BBAABA
AABBBA	BBAAAB
ABAABB	BABBAA
ABABAB	BABABA
ABABBA	BABAAB
ABBAAB	BAABBA
ABBABA	BAABAB
ABBBAA	BAAABB

It is not difficult, however, to imagine that some of those assignments will be rather undesirable—for example, AAABBB. A single-subject analogue of a *randomized block design* can provide a solution. In a conventional multisubject randomized block design, the subjects are divided into blocks, followed by a random assignment of the subjects within each block to the different conditions (Edgington, 1996). In the single-case version of the randomized blocks, adjacent treatment times are grouped together in blocks, and the conditions are assigned randomly

within each block. The order of the treatment conditions is consequently randomized independently within each block of treatment times (Todman & Dugard, 2001). In the case of two conditions, this means that the treatments are presented in pairs and that the order of the two members of the pair is determined randomly and separately for each block (Onghena & Edgington, 2005). Suppose that these two treatments, A and B, are administered in three blocks of one treatment pair each, so that the following $2^3 = 8$ assignments are possible:

AB AB AB	BA BA BA
AB AB BA	BA BA AB
AB BA AB	BA AB BA
AB BA BA	BA AB AB

If one only wants to prevent the same treatment from being administered on several consecutive measurement times, one can use an *alternating treatments design*. This design prevents the temporal clustering of treatments by ensuring that the randomization does not permit more than a specified number of successive time blocks to have the same treatment (Onghena & Edgington, 1994). As an example, we will again use a total of six measurement times to be divided equally over two treatments, A and B. To avoid sequences of consecutive administrations of the same treatment that are too long, we constrain the design to a maximum of two consecutive administrations of a treatment. This results in the following 14 possible assignments:

AABABB	BBABAA
AABBAB	BBAABA
ABAABB	BABBAA
ABABAB	BABABA
ABABBA	BABAAB
ABBAAB	BAABBA
ABBABA	BAABAB

This larger number of possible assignments, as compared with a randomized block design, can be very useful, because the smallest possible p value that can be obtained with a randomization test is the inverse of the number of possible randomizations (see below) (Onghena & Edgington, 2005). In addition, a set of possible assignments that is too small will lead to a statistical test with too little power (Ferron & Onghena, 1996; Ferron & Sentovich, 2002; Ferron & Ware, 1995).

Data Analysis in Single-Case Designs

Visual analysis. The most commonly used method of data analysis in single-case research is probably still visual inspection. This consists of the examination of a graphical display of the data, usually with a measure of time plotted on the abscissa and the dependent variable on the ordinate. Although this may seem a completely subjective procedure, the data are visually inspected according to specific criteria (Kazdin, 1984). Franklin, Gorman, Beasley, and Allison (1997) distinguish three general interpretive principles for visual inspection: the *central location* within phases, as well as any changes in central location between phases; *variability* in the data, including changes in the variation over time; and *trend location* within and between different phases of data collection. Despite these principles, there are some major problems with this kind of data analysis. First, the lack of concrete decision rules permits subjectivity and inconsistency in the interpretations. In addition, there is a high risk of Type I errors. Finally, visual analysis requires a specific pattern of data (e.g., a stable baseline) that often does not emerge (Crosbie, 1993; Kazdin, 1982; Matyas & Greenwood, 1990; Zhan & Ottenbacher, 2001).

Statistical analysis. In situations in which visual inspection is difficult to apply, statistical tests can be a valuable supplement (Lundervold & Belwood, 2000). The parametric statistical tests that are mostly used in group research, such as t tests and ANOVAs, are often inappropriate in single-case designs, because they require assumptions about the data that are met rather infrequently in practice. Although these tests appear to be robust against the violation of such assumptions as normality and homogeneity of variances, this is not the case when group sizes are very small. In particular, the assumption that the residual errors should be independent is problematic in single-case data, because serial dependency is often present in the data. If the residuals are autocorrelated, the results from t and F tests can be seriously biased. Variations of those tests, such as the analysis proposed by Gentile, Roden, and Klein (1972), also appear to be inappropriate (Crosbie, 1993). Therefore, the use of parametric tests on single-case data should be discouraged, unless it can be convincingly shown that the necessary assumptions are met (see, e.g., Gorman & Allison, 1997; Hooton, 1991; Kazdin, 1982; Ludbrook, 1994; Recchia & Rocchetti, 1982; Todman & Dugard, 2001).

The proposed alternatives to these parametric tests include time series analysis and nonparametric rank tests (e.g., the Mann–Whitney U test and the Kruskal–Wallis test). Because these are suitable for the analysis of data when serial dependency is present, they may be a good alternative to parametric tests (Siegel, 1957). However, there are also some difficulties here. Rank tests lack sensitivity to real treatment effects when they are based on only a few subjects, and information may be lost because the scores are discarded once ranks are determined. For time series analysis, a large number of observations are required (various authors have suggested at least 50 data points, which is much more than are typically available in applied research), and such analysis is a rather complex procedure that involves multiple steps (see, e.g., Gorman & Allison, 1997; Kazdin, 1982, 1984; Todman & Dugard, 2001).

Because of the difficulties with the aforementioned methods, we believe that randomization tests provide a strong alternative and may be preferred for analyzing single-case data. *Randomization tests* are statistical tests whose validity is based on the random assignment of units to treatments. By permutation of the order of the data, it is determined whether the same results would have been ob-

tained if the data were assigned to rearranged placements (Busse, Kratochwill, & Elliott, 1995).

The most important advantage of randomization tests is that they are nonparametric, and consequently are not based on distributional assumptions or assumptions about the homogeneity of variances or the independence of residuals (Arndt et al., 1996; Hooton, 1991; Ludbrook, 1994; Recchia & Rocchetti, 1982; Wilson, 2007). Moreover, they are also free from the assumption of random sampling, an assumption on which the probability tables of parametric tests are built (Edgington, 1973; Edgington & Bland, 1993). Such random sampling, however, is usually an unrealistic ideal, and most experiments do not use randomly sampled subjects (Todman & Dugard, 2001). Another advantage is that serial dependency in the data does not affect the application of these tests (Kazdin, 1984). Also, randomization tests are fairly straightforward and intelligible, as well as easy to apply and extremely versatile, so that researchers can design their experiments as they like and then devise a randomization test that is suitable for their particular design (Edgington & Onghena, 2007).

Although randomization tests do not depend on parametric assumptions or on the assumption of random sampling, they do have one requirement: The experimenter has to designate certain times at which the treatment can be administered and then randomly assign each time to a treatment. This random assignment not only enhances the internal validity of the study, as indicated above, it also justifies the application of randomization tests without the need for random sampling (Edgington, 1980; Todman & Dugard, 1999). Ferron, Foster-Johnson, and Kromrey (2003) investigated the functioning of randomization tests with and without random assignment, and they concluded that "the absence of random assignment makes the legitimacy of using a randomization test more questionable" (p. 285) and that randomization tests need to be based on permutations that mirror the random assignment used in the experiment.

Software availability. With randomization tests, the computational burden can be very high. Although most of the examples used in this article can easily be calculated by hand, the number of possible combinations increases very rapidly with the number of observations. Therefore, conducting randomization tests usually requires the availability of a computer and suitable software (Edgington & Bland, 1993; Hooton, 1991; Recchia & Rocchetti, 1982; Todman, 2002). Most of the commonly used statistical software packages, however, present no readily available means for conducting single-case randomization tests. Todman and Dugard (2001) list some packages that do contain procedures for randomization tests in single-case designs (e.g., RANDIBM, StatXact, or SCRT), but most of these have disadvantages—for example, some do not provide tests for single-case designs for which there are no group alternatives; the cost is sometimes rather high; or they may run only under DOS and not be very flexible.

The open source implementation of the S-PLUS language, R, can provide a solution for this problem. R can be downloaded freely from the CRAN Web site (Comprehensive R Archive Network; cran.*r*-project.org) and runs on a variety of UNIX platforms, as well as on Windows and MacOS (Hornik, 2007). It is a powerful tool for statistical modeling and is extremely flexible, which enables it to cope with difficult and unusual data sets and problems, making it ideal for tailoring calculations to one's own specific statistical requirements (Crawley, 2005; Kelley, 2007). This makes R the perfect software environment for devising one's own randomization tests adapted to the design of a particular study. Another strength is the graphical possibilities of the system, which allows for very simple plots as well as for fine-grained control over the appearance of the graphical display of the data (Dalgaard, 2002). A potential disadvantage is that, unlike many other statistical systems, R is restricted to a command line interface, so working with R cannot be reduced to simple clicking in menus. However, various forms of graphical user interfaces will probably become available soon to make R accessible for more users (Maindonald & Braun, 2003). Besides the standard packages, which contain the basic functions and are automatically available in the installation, R has an extensive library of add-on packages. As far as we know, however, no R package of randomization tests for single-case experimental designs exists.

Rationale of Single-Case Randomization Tests

To fill the gap left by the nonexisting R package for single-case randomization tests, in the following section we provide R functions for this purpose. The R code needed to perform these commands can be obtained at no cost from ppw.kuleuven.be/cmes/SCRT-R.html. In an attempt to clarify and illustrate the rationale of randomization tests for single-case designs, a step-by-step procedure will be followed, in which the various stages, from designing an experiment to calculating the randomization test's *p* value, will be made clear. An overview of these steps, together with the R functions needed for each stage, is given in Figure 3. As an illustration, we will conduct, step by step, a randomization test for the hypothetical data set of Onghena (1992), displayed in Figure 1.

Step 1: Choice of design. On the basis of (among other things) the topic of the study, the research question, the stage of the research, and its practicability (available funds, subjects, time, etc.), one should decide whether a single-case design is the best option. If this is the case, a choice must be made between the two types of randomized single-case designs: alternation and phase designs. Although we will not focus on this topic in this article, one should also decide whether replications are needed, and if so, which type (simultaneous or sequential) will be best. In our illustrative example, we used an ABAB phase design.

Step 2: Null hypothesis, alternative hypothesis, and test statistic. The null hypothesis and the alternative hypothesis must be formulated before the start of the study. The *null hypothesis* tested in single-case randomization tests is one of no treatment effect—that is, that responses are independent of the treatment/condition under which they are observed and that performance is a function of factors unrelated to the treatment (Edgington, 1975; Edgington & Onghena, 2007; Kazdin, 1984; Nichols & Holmes, 2001). The *alternative hypothesis*, on

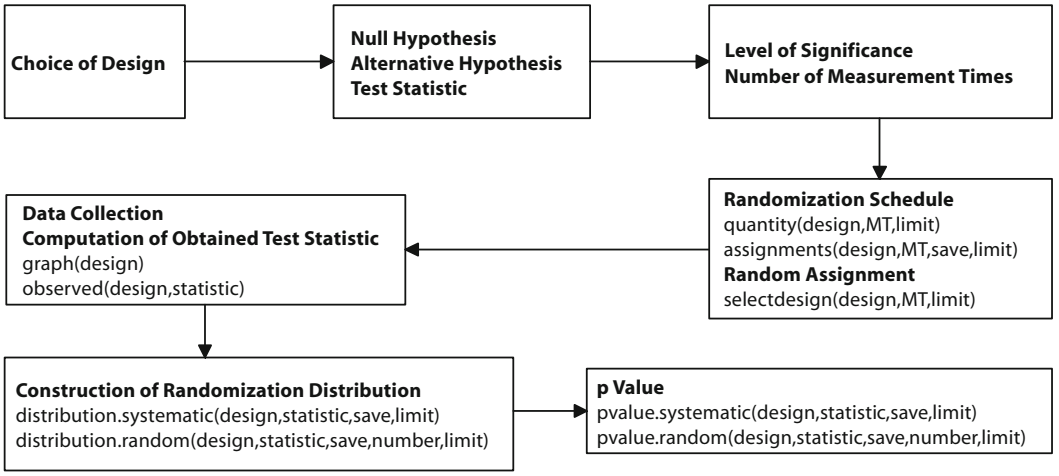

Figure 3. Step-by-step procedure for conducting a single-case randomization test, including R functions.

the other hand, can be decided by the researcher, on the basis of the type of treatment effect that is expected. As in group studies, the hypothesis can be either directional or nondirectional. In general, if the direction of the effect can be specified, it is better to choose the one-tailed option, because the test will then be more powerful. However, if the observed difference is not in the predicted direction, in that case no treatment effect will be discovered (Onghena, 1992).

Linked with the alternative hypothesis is the choice of an appropriate *test statistic*, because this choice can also be made on the basis of the expected treatment effect. An advantage of randomization tests over parametric tests is that the test statistics of the former can be much simpler, because they do not have to include an estimate of the variability, but only need to reflect the expected type of treatment effect. Therefore, the denominator of parametric test statistics can frequently be left out, keeping only the numerator (Edgington & Onghena, 2007). Furthermore, in nonparametric statistical testing, the researcher can choose any test statistic considered appropriate. In contrast, in parametric statistics, only test statistics whose sampling distribution under the null hypothesis is known can be used (Ferron & Sentovich, 2002; Maris & Oostenveld, 2007; Maris, Schoffelen, & Fries, 2007; Nichols & Holmes, 2001). Briefly, any test statistic that is sensitive to the predicted treatment effect can be used with randomization tests. Here we focus on an expected difference in level, which can be reflected by a difference between means (the numerator of the Student's *t* test). For a nondirectional test statistic, we can use the absolute difference between the means, whereas an expected direction can be tested by measuring a corresponding directional difference (Onghena, 1992; Onghena & Edgington, 2005).

The null hypothesis in our example states that the baseline and the treatment conditions do not differ. Because we expect the undesirable behavior to be less frequent in treatment phases than in baseline phases, we will use a one-tailed alternative hypothesis, concerning a difference in level between the A and the B phases. As a test statistic, following Onghena (1992), we will use the difference between the sums of means for the baseline and treatment phases:

$$T = \left(\bar{A}_1 + \bar{A}_2\right) - \left(\bar{B}_1 + \bar{B}_2\right).$$

Step 3: Determination of the level of significance and the number of measurement times. Before gathering the data, the level of significance, α, and the number of measurement occasions should also be determined. In deciding the number of measurement times, a cost–benefit analysis must be performed: An assessment needs to be made that balances the theoretical optimum and practical feasibility. Measurement occasions in single-case designs can be equated to the subjects in group designs. With just a few subjects in a group design, valid and reliable statements in no way can be made about the eventual treatment effects. The same is true for single-case designs with just a few measurement occasions. But, of course, the higher the number of measurement times, the higher will be the costs (in money and in time) of conducting the experiment.

Linked with the number of measurement times is the significance level. In most studies, the significance level, α, is set to .05 or .01. The smallest possible *p* value that can be obtained with randomization tests, however, equals the inverse of the number of possible assignments (Onghena & Edgington, 2005). This quantity depends, among other things, on the number of measurement occasions and on the selected design. Therefore, an α of .01, or even .05, will not always be feasible, because with too few possible data divisions, rejecting the null hypothesis will never be attainable. For the example, the total number of measurement occasions equals 24, and the significance level is .05.

Step 4: Randomization schedule and selection of the assignment. A final matter that needs to be dealt with before data collection is the randomization of an aspect of the design. When this matter is decided, the design of the experiment can be specified in detail by determining how the different conditions or treatments will be divided over the measurement times. Therefore, we have to randomly select one data arrangement among all theoretically pos-

sible permutations. Because it can become rather laborious work to first list all the possible data arrangements and then choose one, it can be helpful to have an R function that does the job instead. For the ABAB design in our example, this means that we should specify the minimum number of measurement times for each phase. Suppose that we set this minimum to four for each of the four phases.

To determine how many possible data arrangements there are for a specific design, the function `quantity(design,MT,limit)` can be used. This function has the arguments `design` (`AB`, `ABA`, `ABAB`, `CRD` (completely randomized design), `RBD` (randomized block design) or `ATD` (alternating treatments design)), `MT` (number of measurement times), and `limit` (equal to the minimum number of observations per phase, for phase designs, or the maximum number of consecutive administrations of the same condition, for alternating treatment designs). In the case of our example, this means that the following should be entered in the R console: `quantity(design="ABAB",MT=24,limit=4)`. The result of this computation is 165, which means that for an ABAB design with 24 measurement times and a minimum of four measurement occasions per phase, the total number of possible data arrangements is 165.

When there are as many possibilities as in our example (or even many more), it would take a lot of time and effort to enumerate them by hand. Therefore, we created the R function `assignments(design,MT,save,limit)`. It has the same arguments as the `quantity` function, plus one extra: With the `save` argument (either `"yes"` or `"no"`), the user can indicate whether he or she wants to save the possible assignments to a file (`save="yes"`) or just see it as output in the R console (`save="no"`). If one chooses to save the assignments in a text file, a window will pop up in which one can indicate where to save it. It can be saved in an existing file, or a new file can be created by inputting a file name with the .txt extension. In the latter case, R will ask for confirmation ("The file does not exist yet. Create the file?"). For our example, suppose that we do not want to save the possible assignments. The suitable R command is then `assignments(design="ABAB",MT=24,save="no",limit=4)`. As output, we will get the 165 possible assignments, which are listed at ppw.kuleuven.be/cmes/SCRT-R.html.

Instead of numbering all possible assignments and then randomly choosing one by means of a random number generator, the function `selectdesign(design,MT,limit)` can be used. For our example, the command `selectdesign(design="ABAB",MT=24,limit=4)` could generate the following sequence of measurement times: A1 A1 A1 A1 A1 A1 B1 B1 B1 B1 B1 B1 A2 A2 A2 A2 A2 A2 B2 B2 B2 B2 B2 B2.

Step 5: Data collection and computation of the observed test statistic. Now the experiment can be conducted and the data collected, according to the randomly chosen design. The data from Onghena (1992) are shown in Table 1.

To make sure that the R functions provided in this article work well, we suggest that researchers follow a few guidelines when creating the text (.txt) file containing the data. The data frame can easily be made in a text editor (e.g., EditPad or Notepad) or in Excel, with the file saved as `text (tab delimited)`. The data file should consist of two columns (if made in a text editor, separated by a tab); the first should contain the condition labels (A and B for alternation and AB designs; A1, A2, B1, and B2 for phase designs with more than two phases), and the second, the obtained scores. Each row will contain one measurement occasion. It is important not to label the rows or the columns.

A graphical presentation of the data, as in Figure 1, can be obtained with the function `graph(design)`. In response to this command, R will open a window to ask for the name of the file in which the data to be graphed can be found. For alternation designs, after the plot is drawn, the user will have to indicate where the legend should be put by clicking within the graph with the cursor.

When performing a randomization test, one first has to calculate the observed test statistic from the obtained raw data (Adams & Anthony, 1996). Because of the small number of scores involved in our example, this can easily be calculated by hand:

$$T = \left(\bar{A}_1 + \bar{A}_2\right) - \left(\bar{B}_1 + \bar{B}_2\right) = (4+3) - (2+1) = 4.$$

For more (or more complex) data, the R function `observed(design,statistic)` can be convenient. The second argument refers to the test statistic that should be used. For alternation designs and AB phase designs, there are three built-in possibilities: `A-B`, `B-A`, and `|A-B|`, which stand for the difference (or absolute value of the difference) between the condition means. For phase designs with more than two phases, six more options are available: `PA-PB`, `PB-PA`, and `|PA-PB|` refer to the (absolute value of the) difference between the means of the phase means, and `AA-BB`, `BB-AA`, and `|AA-BB|` represent the (absolute value of the) difference between the sums of the phase means. Of course, by making small adjustments to the code, other test statistics could easily be adopted. In response to the command, R will ask for the file in which the data can be found. To calculate the test statistic in our example, the command would be `observed(design="ABAB",statistic="AA-BB")`, which gives the same result as our calculation by hand (i.e., 4).

Step 6: Constructing the randomization distribution. The rationale of randomization tests is that they con-

Table 1
Hypothetical Data From Onghena (1992), Collected in an ABAB Design With 24 Measurement Times

Condition	A1A1A1A1A1A1	B1B1B1B1B1B1	A2A2A2A2A2A2	B2B2B2B2B2B2
Score	6 2 5 3 4 4	1 2 3 1 3 2	2 3 4 2 4 3	0 1 2 0 2 1

Table 2
Exhaustive Randomization Distribution for the Data From Onghena (1992)

1.0000	1.0500	1.0909	1.1833	1.2000	1.2409	1.3500	1.3909	1.4500	1.5000	1.5000
1.5000	1.5536	1.5833	1.6250	1.6548	1.6833	1.6833	1.8000	1.8214	1.8536	1.8770
1.8833	2.0000	2.0036	2.0333	2.1000	2.1136	2.1214	2.1250	2.1250	2.1548	2.1667
2.2167	2.2227	2.2500	2.2500	2.2500	2.2955	2.2964	2.3036	2.3333	2.3333	2.3536
2.3690	2.4000	2.4000	2.4036	2.4048	2.4048	2.4136	2.4167	2.4417	2.4500	2.4964
2.4964	2.4964	2.5045	2.5119	2.5119	2.5119	2.5227	2.5333	2.5393	2.5417	2.5500
2.5833	2.5833	2.6214	2.6250	2.6250	2.6310	2.6333	2.6417	2.6417	2.6500	2.6857
2.7000	2.7167	2.7286	2.7341	2.7429	2.7500	2.7500	2.7500	2.8000	2.8393	2.8393
2.8393	2.8393	2.8417	2.8500	2.8500	2.8556	2.8667	2.8690	2.8714	2.8714	2.9000
2.9286	2.9333	2.9500	2.9500	2.9583	2.9667	2.9750	2.9762	3.0222	3.0667	3.0714
3.0833	3.0833	3.0857	3.0857	3.1250	3.1250	3.1250	3.1393	3.1556	3.1667	3.1786
3.1786	3.2143	3.2143	3.2167	3.2750	3.2857	3.2917	3.2917	3.2976	3.3000	3.3691
3.3714	3.3714	3.3714	3.3750	3.3750	3.3833	3.3833	3.4000	3.4286	3.4667	3.5000
3.5198	3.5238	3.5333	3.5357	3.5476	3.6048	3.6056	3.6750	3.7143	3.7143	3.7143
3.7500	3.7556	3.8214	3.8571	3.9167	3.9500	3.9643	4.0000	4.1250	4.1250	4.2000

sider all possible recombinations of the data, given the randomization procedure that was used in the study (Todman, 2002). The null hypothesis states that no effect of the different conditions will occur. This means that, under the null hypothesis, the obtained responses will be the same as those that would have been obtained under any other random ordering (Edgington, 1975; Edgington & Onghena, 2007; Kazdin, 1984; Nichols & Holmes, 2001). To test this null hypothesis, the test statistic for every possible random division of the data that could occur under the null hypothesis has to be calculated (Adams, Gurevitch, & Rosenberg, 1997; Edgington, 1975; Solow, 1993). Therefore, the score obtained for each measurement time is kept fixed, but the conditions assigned to the measurement times are randomly shuffled according to the possible orderings. The test statistics are then sorted in ascending order, which forms the randomization distribution under the null hypothesis (an equivalent of the sampling distribution in parametric statistical testing).

With a large amount of data, it will not be feasible to compute the test statistics for all possible permutations. When a systematic randomization test, which calculates all of the test statistics, becomes computationally cumbersome, a random version, in which a random sample of the test statistics is calculated instead, can be chosen (Besag & Diggle, 1977). This "nonexhaustive randomization test" does not use the entire distribution, but only a simulated one. In this way, the enormous calculation capacity that would otherwise be required is reduced (Recchia & Rocchetti, 1982). Depending on whether all possible permutations (the "systematic" procedure) or only a random sample (the "random" procedure) is carried out, either an exhaustive or a nonexhaustive randomization distribution will be obtained. The R function for the systematic randomization distribution is `distribution.systematic(design, statistic,save,limit)`. In the random version, `distribution.random(design,statistic, save,number,limit)`, one additional argument, `number`, is needed: With it, one can indicate how many randomizations are required (e.g., 1,000). In response to the command, a pop-up window will appear asking for the file from which the data should be read. If the `save` argument is set to `"yes"`, a second window will open in which one can input the name of the file to which the randomization distribution should be saved.

In the example, the number of possible assignments was 165, which means that 165 different test statistics need to be calculated in order to obtain the exact randomization distribution. Since this does not take too much computational time, we will use the systematic version: `distribution.systematic(design="ABAB", statistic="AA-BB",save="no",limit=4)`. The obtained randomization distribution is given in Table 2.

Step 7: Assessing the *p* value. By locating the observed test statistic in the randomization distribution, the statistical significance of the outcome can be obtained (Murray, Varnell, & Blitstein, 2004; Strauss, 1982). The proportion of test statistics in the randomization distribution that exceed or equal the observed test statistic gives the randomization test's *p* value. When this value is less than or equal to the predetermined value of α, the null hypothesis is rejected, and the alternative hypothesis can be accepted (Potvin & Roff, 1993).

In line with the function for constructing the randomization distribution, there are also two versions of the function for calculating the *p* value: `pvalue.systematic (design,statistic,save,limit)` and `pvalue.random(design,statistic,save, number,limit)`. With the save argument, the user has the possibility of saving the randomization distribution, used to compute the *p* value, to a file, even though it is not printed as output in the R console. To calculate the *p* value for the hypothetical data set of Onghena (1992), `pvalue .systematic(design="ABAB",statistic ="AA-BB",save="no",limit=4)` has to be input as a command in the R console. This gives a *p* value of .0242, which is smaller than the predetermined significance level of .05. From this, the null hypothesis of no difference between the treatment and baseline phases can be rejected, a result that concurs with the tentative conclusion from our visual inspection of the data.

Discussion

Single-case designs and randomization tests to analyze the resulting data are rarely used by researchers. Part of

the reason for this lies in the fact that most statistical and methodological courses do not elaborate on this topic, but focus instead on large-scale group studies and classical *t* tests and ANOVAs. Therefore, many scientists do not know much about these methods and techniques or consider them to be inferior to large-*n* designs or parametric statistical tests. However, single-case randomization tests have their own merits, and in some instances should be preferred, or at least seen as a supplement. Another reason for the underutilization of those designs and tests is that most widely used statistical software programs do not include the possibility of conducting randomization tests for single-case experiments. Because of the high computational burden of randomization tests, it would be a hopeless task to calculate everything by hand.

In this article, we have tried to fill this lacuna. By outlining single-case phase and alternation experiments in more detail and describing the accompanying randomization procedures, we hope that we have called attention to the specificities and opportunities offered by these experimental designs. The presentation of the different steps that need to be taken when designing single-case experiments and then analyzing them using randomization tests was meant to clarify the rationale behind such tests. Finally, by supplementing these steps with R functions to execute the calculations, we have minimized the time and effort needed to conduct single-case randomization tests.

We ended our step-by-step explanation of these procedures with the calculation of the randomization test's *p* value, but this of course is never the ultimate goal or the endpoint of a statistical analysis. We agree with Wilson (2007) that the arbitrary significance level of .05 should be considered with caution and that a *p* value does not tell the whole story. Statistically significant results can be clinically meaningless, so statistical significance does not by definition entail clinical significance (Lundervold & Belwood, 2000; Turk, 2000). As Robinson and Levin (1997) have mentioned, besides indicating whether an observed effect is statistically significant, researchers should also pay attention to the magnitude or importance of the effect. For this reason and many others, the enormous flexibility of R gives it a huge advantage over other statistical software packages. For instance, R can be used to easily calculate effect size measures, percentages of nonoverlapping data, and other indicators of the importance of observed systematic effects. Also, researchers can draw graphical indicators of the magnitude of an effect, thanks to the very well-developed graphical possibilities of the system (Dalgaard, 2002; Kelley, 2007).

Although the example given in this article was based on an ABAB design, the same rationale could be applied to randomization tests for other phase designs and for single-case alternation designs. If we take, for example, the data of Baplu (2005) from the beginning of this article (see Figure 2) and assume that she used a completely randomized design, with no restrictions on the number of consecutive administrations of the same condition, the command `quantity(design="CRD",MT=48)` reveals more than 10^{13} possible data rearrangements. Because this many calculations would take a lot of computational time, an alternative option is to use a Monte Carlo version with 10,000 randomizations to compute the *p* value: `pvalue.random(design="CRD",statistic="A-B",save="no",number=10000)`. This gives a *p* value of .4926, on the basis of which we cannot reject the null hypothesis. Thus, the randomization test provides insufficiently convincing evidence for the efficacy of methylphenidatum in this adolescent. This is in line with a visual inspection of the data, which revealed a slightly, but not conclusively, positive effect of the medication, with somewhat fewer concentration difficulties in the treatment condition than with the placebo. Also, on the basis of Baplu's qualitative research, which consisted of interviews and observation sessions, no clear distinction between the conditions could be made. Because the methylphenidatum did not seem to have clinically or statistically significant effects on the adolescent's concentration difficulties, she decided that this medication was not the best solution for this boy's problems.

The R functions in this article are limited to AB, ABA, and ABAB phase designs and to completely randomized, alternating treatments, and randomized block alternation designs. An interesting direction for future research will be to use the generic capabilities of R to extend these functions to deal with other types of single-case experimental designs—for instance, interaction designs or designs with more than two conditions. It would also be interesting to develop more tools to efficiently analyze data from simultaneous and sequential replicated single-case experiments, to calculate different measures of effect size, and to incorporate more visual procedures for analyzing data.

AUTHOR NOTE

This study was funded by Grant G.0624.07 from the Research Foundation–Flanders (Belgium). The authors thank David C. Howell for his helpful comments on an earlier version of the manuscript and Wilfried Cools for his assistance with the R functions. Correspondence concerning this article should be addressed to I. Bulté, K.U. Leuven, Centre for Methodology of Educational Research, Andreas Vesaliusstraat 2, bus 3762, B-3000 Leuven, Belgium (e-mail: isis.bulte@ped.kuleuven.be).

REFERENCES

ADAMS, D. C., & ANTHONY, C. D. (1996). Using randomization techniques to analyse behavioural data. *Animal Behaviour*, **51**, 733-738.

ADAMS, D. C., GUREVITCH, J., & ROSENBERG, M. S. (1997). Resampling tests for meta-analysis of ecological data. *Ecology*, **78**, 1277-1283.

ARNDT, S., CIZADLO, T., ANDREASEN, N. C., HECKEL, D., GOLD, S., & O'LEARY, D. S. (1996). Test for comparing images based on randomization and permutation methods. *Journal of Cerebral Blood Flow & Metabolism*, **16**, 1271-1279.

AVINS, A. L., BENT, S., & NEUHAUS, J. M. (2005). Use of an embedded *N*-of-1 trial to improve adherence and increase information from a clinical study. *Contemporary Clinical Trials*, **26**, 397-401.

BACKMAN, C. L., & HARRIS, S. R. (1999). Case studies, single-subject research, and *N* of 1 randomized trials: Comparisons and contrasts. *American Journal of Physical Medicine & Rehabilitation*, **78**, 170-176.

BAPLU, A. (2005). *Rilatine, the smart drug. Een dubbelblind placebogecontroleerd N = 1 experiment bij een adolescent met concentratieproblemen.* Unpublished master's thesis, Katholieke Universiteit Leuven, Belgium.

BARLOW, D. H., & HAYES, S. C. (1979). Alternating treatments design: One strategy for comparing the effects of two treatments in a single subject. *Journal of Applied Behavior Analysis*, **12**, 199-210.

BARLOW, D. H., & HERSEN, M. (1984). *Single case experimental designs: Strategies for studying behavior change* (2nd ed.). New York: Pergamon.

BESAG, J., & DIGGLE, P. J. (1977). Simple Monte Carlo tests for spatial pattern. *Journal of the Royal Statistical Society: Series C*, **26**, 327-333.

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

CAMPBELL, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, **54**, 297-312.

CHRISTENSEN, L. B. (2001). *Experimental methodology* (8th ed.). Boston: Allyn & Bacon.

CRAWLEY, M. J. (2005). *Statistics: An introduction using R*. Chichester, U.K.: Wiley.

CROSBIE, J. (1993). Interrupted time-series analysis with brief single-subject data. *Journal of Consulting & Clinical Psychology*, **61**, 966-974.

DALGAARD, P. (2002). *Introductory statistics with R.* New York: Springer.

EDGINGTON, E. S. (1973). The random-sampling assumption in "Comment on component-randomization tests." *Psychological Bulletin*, **80**, 84-85.

EDGINGTON, E. S. (1975). Randomization tests for one-subject operant experiments. *Journal of Psychology: Interdisciplinary & Applied*, **90**, 57-68.

EDGINGTON, E. S. (1980). Validity of randomization tests for one-subject experiments. *Journal of Educational Statistics*, **5**, 235-251.

EDGINGTON, E. S. (1996). Randomized single-subject experimental designs. *Behaviour Research & Therapy*, **34**, 567-574.

EDGINGTON, E. S., & BLAND, B. H. (1993). Randomization tests: Application to single-cell and other single-unit neuroscience experiments. *Journal of Neuroscience Methods*, **47**, 169-177.

EDGINGTON, E. S., & ONGHENA, P. (2007). *Randomization tests* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.

FERRON, J., FOSTER-JOHNSON, L., & KROMREY, J. D. (2003). The functions of single-case randomization tests with and without random assignment. *Journal of Experimental Education*, **71**, 267-288.

FERRON, J., & ONGHENA, P. (1996). The power of randomization tests for single-case phase designs. *Journal of Experimental Education*, **64**, 231-239.

FERRON, J., & SENTOVICH, C. (2002). Statistical power of randomization tests used with multiple-baseline designs. *Journal of Experimental Education*, **70**, 165-178.

FERRON, J., & WARE, W. (1995). Analyzing single-case data: The power of randomization tests. *Journal of Experimental Education*, **63**, 167-178.

FRANKLIN, R. D., ALLISON, D. B., & GORMAN, B. S. (Eds.) (1997). *Design and analysis of single-case research.* Mahwah, NJ: Erlbaum.

FRANKLIN, R. D., GORMAN, B. S., BEASLEY, T. M., & ALLISON, D. B. (1997). Graphical display and visual analysis. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 119-158). Mahwah, NJ: Erlbaum.

GENTILE, J. R., RODEN, A. H., & KLEIN, R. D. (1972). An analysis-of-variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis*, **5**, 193-198.

GORMAN, B. S., & ALLISON, D. B. (1997). Statistical alternatives for single-case designs. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 159-214). Mahwah, NJ: Erlbaum.

GUYATT, G. H., KELLER, J. L., JAESCHKE, R., ROSENBLOOM, D., ADACHI, J. D., & NEWHOUSE, M. T. (1990). The *n*-of-1 randomized controlled trial: Clinical usefulness. Our three-year experience. *Annals of Internal Medicine*, **112**, 293-299.

GUYATT, G., SACKETT, D., ADACHI, J., ROBERTS, R., CHONG, J., ROSENBLOOM, D., & KELLER, J. (1988). A clinician's guide for conducting randomized trials in individual patients. *Canadian Medical Association Journal*, **139**, 497-503.

GUYATT, G., SACKETT, D., TAYLOR, D. W., CHONG, J., ROBERTS, R., & PUGSLEY, S. (1986). Determining optimal therapy—Randomized trials in individual patients. *New England Journal of Medicine*, **314**, 889-892.

HAYES, S. C. (1981). Single case experimental design and empirical clinical practice. *Journal of Consulting & Clinical Psychology*, **49**, 193-211.

HOOTON, J. W. (1991). Randomization tests: Statistics for experimenters. *Computer Methods & Programs in Biomedicine*, **35**, 43-51.

HORNIK, K. (2007). *The R FAQ: Frequently asked questions on R*. Retrieved September 24, 2007, from CRAN.R-project.org/doc/FAQ/.

KAZDIN, A. E. (1981a). Drawing valid inferences from case studies. *Journal of Consulting & Clinical Psychology*, **49**, 183-192.

KAZDIN, A. E. (1981b). External validity and single-case experimentation: Issues and limitations (a response to J. S. Birnbrauer). *Analysis & Intervention in Developmental Disabilities*, **1**, 133-143.

KAZDIN, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York: Oxford University Press.

KAZDIN, A. E. (1984). Statistical analyses for single-case experimental designs. In D. H. Barlow & M. Hersen (Eds.), *Single case experimental designs: Strategies for studying behavior change* (2nd ed.; pp. 258-324). New York: Pergamon.

KAZDIN, A. E. (2003). *Research design in clinical psychology* (4th ed.). Boston: Allyn & Bacon.

KELLEY, K. (2007). Methods for the behavioral, educational, and social sciences: An R package. *Behavior Research Methods*, **39**, 979-984.

KOEHLER, M. J., & LEVIN, J. R. (1998). Regulated randomization: A potentially sharper analytical tool for the multiple-baseline design. *Psychological Methods*, **3**, 206-217.

KOEHLER, M. J., & LEVIN, J. R. (2000). RegRand: Statistical software for the multiple-baseline design. *Behavior Research Methods, Instruments, & Computers*, **32**, 367-371.

LUDBROOK, J. (1994). Advantages of permutation (randomization) tests in clinical and experimental pharmacology and physiology. *Clinical & Experimental Pharmacology & Physiology*, **21**, 673-686.

LUNDERVOLD, D. A., & BELWOOD, M. F. (2000). The best kept secret in counseling: Single-case ($N = 1$) experimental designs. *Journal of Counseling & Development*, **78**, 92-102.

MAINDONALD, J., & BRAUN, J. (2003). *Data analysis and graphics using R: An example-based approach*. Cambridge: Cambridge University Press.

MARIS, E., & OOSTENVELD, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, **164**, 177-190.

MARIS, E., SCHOFFELEN, J.-M., & FRIES, P. (2007). Nonparametric statistical testing of coherence differences. *Journal of Neuroscience Methods*, **163**, 161-175.

MATYAS, T. A., & GREENWOOD, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavioral Analysis*, **23**, 341-351.

MOORE, D. S., & MCCABE, G. P. (2006). *Introduction to the practice of statistics* (5th ed.). New York: Freeman.

MURRAY, D. M., VARNELL, S. P., & BLITSTEIN, J. L. (2004). Design and analysis of group-randomized trials: A review of recent methodological developments. *American Journal of Public Health*, **94**, 423-432.

NICHOLS, T. E., & HOLMES, A. P. (2001). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, **15**, 1-25.

NIKLES, C. J., CLAVARINO, A. M., & DEL MAR, C. B. (2005). Using *n*-of-1 trials as a clinical tool to improve prescribing. *British Journal of General Practice*, **55**, 175-180.

ONGHENA, P. (1992). Randomization tests for extensions and variations of ABAB single-case experimental designs: A rejoinder. *Behavioral Assessment*, **14**, 153-171.

ONGHENA, P. (2005). Single case designs. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 4; pp. 1850-1854). Chichester, U.K.: Wiley.

ONGHENA, P., & EDGINGTON, E. S. (1994). Randomization tests for restricted alternating treatments designs. *Behaviour Research & Therapy*, **32**, 783-786.

ONGHENA, P., & EDGINGTON, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *Clinical Journal of Pain*, **21**, 56-68.

POTVIN, C., & ROFF, D. A. (1993). Distribution-free and robust statistical methods: Viable alternatives to parametric statistics. *Ecology*, **74**, 1617-1628.

RAMSEY, F. L., & SCHAFER, D. W. (2002). *The statistical sleuth: A course in methods of data analysis* (2nd ed.). Pacific Grove, CA: Duxbury/Thomson Learning.

RECCHIA, M., & ROCCHETTI, M. (1982). The simulated randomization test. *Computer Programs in Biomedicine*, **15**, 111-116.

ROBINSON, D. H., & LEVIN, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher*, **26**, 21-26.

SIEGEL, S. (1957). Nonparametric statistics. *American Statistician*, **11**, 13-19.

SOLOW, A. R. (1993). A simple test for change in community structure. *Journal of Animal Ecology*, **62**, 191-193.

STRAUSS, R. E. (1982). Statistical significance of species clusters in association analysis. *Ecology*, **63**, 634-639.

TODMAN, J. (2002). Randomization in single-case experimental designs. *Advances in Clinical Neuroscience & Rehabilitation*, **2**, 18-19.

TODMAN, J. [B.], & DUGARD, P. (1999). Accessible randomization tests for single-case and small-*n* experimental designs in AAC research. *Augmentative & Alternative Communication*, **15**, 69-82.

TODMAN, J. B., & DUGARD, P. (2001). *Single-case and small-n experimental designs: A practical guide to randomization tests*. Mahwah, NJ: Erlbaum.

TURK, D. C. (2000). Statistical significance and clinical significance are not synonyms! (Editorial). *Clinical Journal of Pain*, **16**, 185-187.

VAN DEN NOORTGATE, W., & ONGHENA, P. (2003a). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly*, **18**, 325-346.

VAN DEN NOORTGATE, W., & ONGHENA, P. (2003b). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, & Computers*, **35**, 1-10.

WEGMAN, A. C. M., VAN DER WINDT, D. A. W. M., BONGERS, M., TWISK, J. W. R., STALMAN, W. A. B., & DE VRIES, T. P. G. M. (2005). Efficacy of temazepam in frequent users: A series of *N*-of-1 trials. *Family Practice*, **22**, 152-159.

WEGMAN, A. C. M., VAN DER WINDT, D. A. W. M., DE HAAN, M., DEVILLÉ, W. L. J. M., FO, C. T. C. A., & DE VRIES, T. P. G. M. (2003). Switching from NSAIDs to paracetamol: A series of *n* of 1 trials for individual patients with osteoarthritis. *Annals of the Rheumatic Diseases*, **62**, 1156-1161.

WILSON, J. B. (2007). Priorities in statistics, the sensitive feet of elephants, and don't transform data. *Folia Geobotanica*, **42**, 161-167.

ZHAN, S., & OTTENBACHER, K. J. (2001). Single subject research designs for disability research. *Disability & Rehabilitation*, **23**, 1-8.

(Manuscript received December 22, 2007;
accepted for publication January 4, 2008.)