Behavior Research Methods
2009, 41 (2), 477-485
doi:10.3758/BRM.41.2.477

Randomization tests for multiple-baseline designs: An extension of the SCRT-R package

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

Multiple-baseline designs are an extension of the basic single-case AB phase designs, in which several of those AB designs are implemented simultaneously to different persons, behaviors, or settings, and the intervention is introduced in a staggered way to the different units. These designs are well-suited for research in the behavioral sciences. We discuss the advantages and limitations for valid inferences, and suggest a statistical technique—randomization tests—for use with multiple-baseline data, to complement visual analysis. In addition, we provide an extension of our SCRT-R package (which already contained means for conducting randomization tests on single-case phase and alternation designs), for multiple-baseline AB data.

Multiple-baseline designs are variants of single-case designs well suited to behavioral research. In this article, we want to bring these designs to the attention of experimental psychologists and social and behavioral researchers in general, discuss such designs' advantages and limitations for valid inference in behavioral research, and suggest a statistical data-analytic technique to complement visual inspection, together with software to conduct those analyses.

A multiple-baseline design consists of a series of replicated single-case designs, in which the replications are carried out at the same time. They extend the basic single-case AB phase design by implementing several of those AB designs simultaneously to different persons, behaviors, or settings (Ferron & Scott, 2005; Onghena & Edgington, 2005). For convenience, these separate persons, behaviors, or settings will henceforth be called *units*.

The most characteristic feature of a multiple-baseline design is that the intervention is applied sequentially across the different units. By extending each A phase a little further than the previous one, the intervention is introduced in a staggered way. When a change in each unit takes place if, and only if, the intervention is introduced for that unit, researchers can be more confident to attribute the effects to this intervention instead of to extraneous effects (Baer, Wolf, & Risley, 1968; Barlow & Hersen, 1984; Hayes, 1981; Kazdin, 1982; Kinugasa, Cerin, & Hooper, 2004; Koehler & Levin, 2000).

As an example, Ziegler (1994) used a multiple-baseline design to determine the effectiveness of an attentional shift training program on the performance of targeted soccer skills. In soccer, one of the most important skills is the ability to attend and respond quickly and accurately to appropriate environmental cues. As subjects, Ziegler chose 4 male collegiate soccer players who scored low on a test of attentional shift. During baseline, she observed their ability to hit a target in four soccer drills (each athlete was allowed 3 attempts in each drill, a total of 12 attempts per testing session). After a stable baseline had occurred for the 1st subject (after Session 6), the attentional shift training intervention was introduced for him. For the other 3 subjects, intervention started after the 8th, 10th, and 12th session, giving rise to the typical staggered administration. During the treatment phase, the men continued to be observed. As indicated in Figure 1, the accuracy of execution of the experimental soccer drill improved after treatment. There is a marked increase in points scored after the intervention.

Besides applying the treatment sequentially to several subjects, as in the example of Ziegler (1994), the intervention in multiple-baseline designs can also be introduced sequentially to different behaviors within the same subject, or to several independent situations, settings, or time periods in a given subject (Barlow & Hersen, 1984; Kazdin, 1982; Kinugasa et al., 2004). The placement of the intervention points throughout the different units can be decided in several ways. Most traditionally, researchers use a response-guided assignment procedure, in which the intervention points are chosen on the basis of emerging patterns in the data (Ferron & Jones, 2006). In the example of Ziegler, this was done for the 1st subject by starting the treatment phase after baseline data were stable. For the other subjects, she decided to start the B phase on the basis of a systematic assignment schedule. Here the intervention points are decided a priori, whereby an even staggering across time is obtained. Another procedure that can be used to determine the intervention points throughout the different units is a random assignment approach, in which the placement is determined at random (Marascuilo & Busk, 1988; Onghena, 1992; Wampold & Worsham, 1986).

I. Bulté, isis.bulte@ped.kuleuven.be

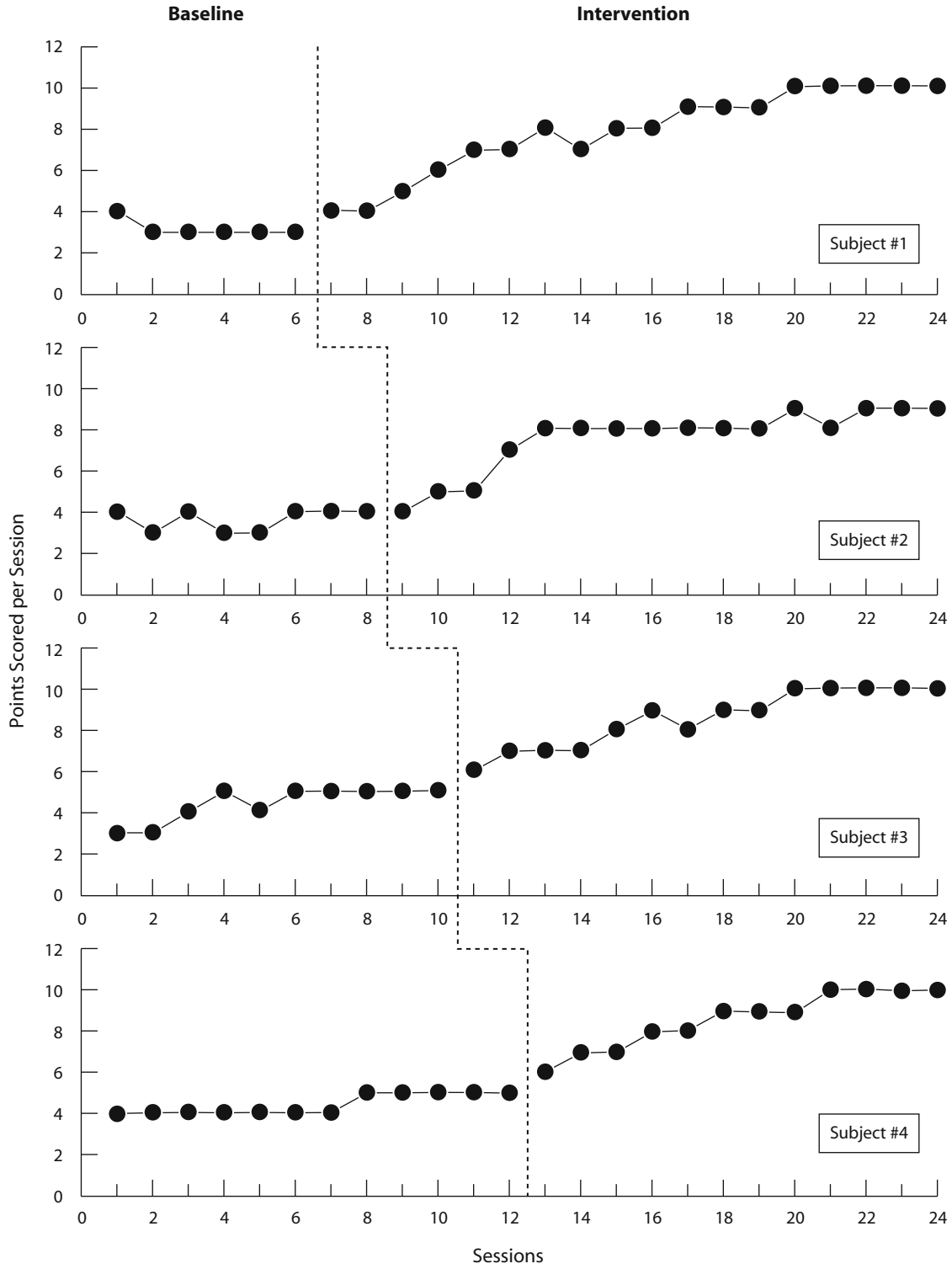

Figure 1. Results of the attentional shift training program. The numbers of points per session are shown for each of the 4 subjects. Each session consisted of 12 attempts (4 drills × 3 attempts), so the maximum possible score for each testing session was 12. From "The Effects of Attentional Shift Training on the Execution of Soccer Skills: A Preliminary Investigation," by S. G. Ziegler, 1994, *Journal of Applied Behavior Analysis*, *27*, p. 551. Copyright 1994 by the Society for the Experimental Analysis of Behavior. Reprinted with permission.

ADVANTAGES AND LIMITATIONS

In several circumstances, single-case designs can provide a good alternative or supplement to group designs. Examples are the generation of pilot data in the early stages of larger group studies; research concerning rare types of experimental subjects; and, of course, when research funds are scarce and it is not possible to obtain enough subjects for large-scale group studies (Barlow & Hersen, 1984; Edgington & Onghena, 2007).

In a variety of research contexts, the multiple-baseline design satisfies critical empirical validity criteria, which include the internal validity, replication and generalization, and selectivity and discrimination (Koehler & Levin, 1998) better than other single-case designs do. By simultaneously monitoring the different units, a relative control over historical confounding variables can be obtained by which plausible rival hypotheses that could account for the observed effects can be ruled out (Barlow & Hersen, 1984; Kazdin, 1982; Onghena, 2005). When the observed effects are the result of time-related factors such as history or maturation, rather than experimental manipulation, these same factors would also be expected to influence the other units when they are still in the A phase (Harris & Jenson, 1985). This simultaneous monitoring also allows for a closer approximation of naturalistic conditions, where several target behaviors occur at the same time (Long & Hollin, 1995). Replication and generalization across units is established by applying the intervention to several units, which all receive a separate AB phase design. These replications are necessary for demonstrating the external validity of the experiment (Onghena & Edgington, 2005). Another advantage of multiple-baseline designs is that no reversal is required to demonstrate the effect of an intervention, so this technique can be especially useful when behavior appears to be irreversible, or when reversals are considered undesirable because of practical limitations or ethical concerns (Baer et al., 1968; Kazdin, 1982). Multiple-baseline designs are also well suited to clinical situations, because the intervention is first implemented in 1 patient, behavior, or setting, before it is extended. This is an advantage for the person who has to administer the treatment, who only needs to increase the scope after fully having mastered the initial application, as well as for the client, for whom a shaping program is followed in which more behaviors or situations are gradually incorporated into the program (Kazdin, 1982).

Of course, like all designs, multiple-baseline designs have limitations. Although they can provide rather strong inferences when all treatment effects occur at the point at which intervention is introduced for each unit, the conclusions are less clear when the effects are inconsistent across units (Ferron & Scott, 2005). This is the case when not all units change at the point at which the intervention is introduced, and it has some serious implications for internal validity when there are only 2 units, because extraneous effects cannot then be ruled out (Kazdin, 1982). Ambiguity can also arise when the units are interdependent, so that an intervention in one unit has an effect on the other unit(s) (Kazdin, 1973, 1982; Leitenberg, 1973). Both problems relate to the greatest threat to internal validity—namely, "history"—when events other than the treatment could have produced the observed effect (Christ, 2007; Shadish, Cook, & Campbell, 2002). Kazdin and Kopel (1975) made three recommendations to maximize the possibility of drawing valid inferences: (1) Researchers should select units that are as independent from each other as possible; (2) they should use 4 or more units rather than just a few; and (3) they should implement a reversal phase for 1 of the units.

Another potential problem is that the intervention is withheld temporarily from some of the units because of the staggered administration (Harris & Jenson, 1985). From this, ethical (withholding effective treatment) and methodological (e.g., boredom) difficulties may arise. These can, however, usually be avoided by including only very few data points in the baseline phases, by shortening the delay period between the implementation of treatments for successive units, or by applying the treatment to more than 1 unit at the same time (Kazdin, 1982). Finally, this design could be considered weaker than, for example, a withdrawal design, because potential treatment effects cannot be demonstrated directly but should instead be inferred from units not yet treated (Barlow & Hersen, 1984).

ANALYZING MULTIPLE-BASELINE DATA

Visual Analysis

For the analysis of multiple-baseline data, several techniques have been suggested. The oldest and still most popular one is visual inspection. In this nonstatistical method of data analysis, data are plotted on a graph, in which the *y*-axis represents the dependent variable and the *x*-axis represents units of time (Zhan & Ottenbacher, 2001). On the basis of these graphs, a judgment is reached about the reliability or consistency of intervention effects (Long & Hollin, 1995).

This method of data analysis undoubtedly has some advantages, such as the speed of making the graphs, yielding conclusions, and deriving hypotheses (Parsonson & Baer, 1992); in many cases, however, these advantages do not outweigh the difficulties. The main problems are the lack of concrete decision rules, the requirement of a particular pattern of the data (e.g., stable baselines without a trend in the direction of the expected change), and the overlooking of small but systematic effects (Kazdin, 1982). The accuracy and reliability of this method have been questioned because there has often been a lack of agreement among judges (e.g., DeProspero & Cohen, 1979). And especially when there is variability within phases, both Type II and Type I error rates are elevated to unacceptable levels (Matyas & Greenwood, 1990). Morley and Adams (1991) recommended complementing visual analysis with a statistical analysis of the data, whenever possible.

However, although several statistical tests have been suggested for use with multiple-baseline designs, they are still rarely used; whereas in group research statistical tests are commonly used to evaluate the effect of an intervention, in single-case (multiple-baseline) research, statistical tests are the exception rather than the rule.

F and *t* Tests

The most familiar group of statistical tests consists of the parametric *F* and *t* tests. However, these tests are controversial in the analysis of multiple-baseline data, because data from single-case research often violates some of the assumptions on which parametric tests depend (e.g., normality and homogeneity of variances). The assumption of serial independence is often especially problematic, because multiple-baseline data tend to have autocorrelated residuals that can seriously bias the results from *F* and *t* tests (see, e.g., Gorman & Allison, 1996; Hooton, 1991; Kazdin, 1982, 1984; Kinugasa et al., 2004; Ludbrook, 1994; Recchia & Rocchetti, 1982; Todman & Dugard, 2001). Before these tests are used on multiple-baseline data, the data should be demonstrably free from serial dependency. Otherwise, alternative statistical tests should be considered (Kazdin, 1984; Long & Hollin, 1995).

Time Series Analysis

One of the proposed alternatives to conventional *F* and *t* tests is time series analysis, which is suitable for the analysis of data when serial dependency is present and when the criteria for visual inspection (e.g., stable baselines) are not met (Kazdin, 1982, 1984). However, time series analysis requires many data points to determine the existence and the pattern of autocorrelation and to identify the model correctly (Box, Jenkins, & Reinsel, 1994; Crosbie, 1993; Tryon, 1982). This could cause problems for multiple-baseline designs, in which the phases are usually rather short, so that the intervention will not be withheld for a very long time. Another difficulty of time series analysis is the complexity of the mathematical theories on which it is based (Gorman & Allison, 1996; Hartmann et al., 1980).

Split-Middle Technique

In addition, several nonparametric statistical tests have been proposed for use with multiple-baseline designs. One approach is the combined use of the split-middle technique and a binomial test. This method can reveal the nature of the trend in the data by plotting linear trend lines (celeration lines) that best fit the data, then applying a binomial test to see whether the number of data points in the intervention phase falls above (or below) the projected line of the baseline (Kazdin, 1982; Kinugasa et al., 2004). Requirements for this test are that several observations are needed in the different phases and should be made at equally spaced intervals in each phase (Kazdin, 1982). The split-middle technique is easy to compute and can be used with a small number of data points, an advantage for use with multiple-baseline designs (Zhan & Ottenbacher, 2001); however, it is not suited when data are autocorrelated (Kinugasa et al., 2004).

Randomization Tests

Other nonparametric tests that have been suggested for the analysis of multiple-baseline data, and that we believe provide a strong alternative, are randomization tests. Their most important advantages are that they are free from the assumption of random sampling (Edgington, 1973; Edgington & Bland, 1993), and they are not based on distributional assumptions or assumptions about the homogeneity of variances. Also, the presence of serial dependence or trends in the data will not invalidate the result of a randomization test (Arndt et al., 1996; Hooton, 1991; Ludbrook, 1994; Recchia & Rocchetti, 1982; Wilson, 2007). The basic approach of randomization tests is very straightforward. The essence is that some aspect of the experimental design needs to be randomized. The randomization test is then based on permutations that mirror the random assignment used in the experiment (Ferron, Foster-Johnson, & Kromrey, 2003). The null hypothesis states that there is no effect of the intervention (Edgington & Onghena, 2007). This is tested by locating the observed value of the test statistic in the randomization distribution (an equivalent of the sampling distribution in parametric statistical testing): The randomization test's *p* value is equal to the proportion of test statistics that exceed or equal the observed test statistic. The null hypothesis is rejected when this value is less than or equal to the predetermined significance level α (Murray, Varnell, & Blitstein, 2004; Potvin & Roff, 1993; Strauss, 1982). A more comprehensive step-by-step explanation of the randomization test procedure can be found in Bulté and Onghena (2008).

By applying the randomization schedules to the different units, the necessary random assignment can be incorporated easily in multiple-baseline designs (Onghena, 1992). Because in multiple-baseline AB designs the order of the phases cannot be altered (all A observations always precede all B observations), the randomization cannot be applied to the treatment order. The only aspect of the design that can be manipulated, and consequently randomized, is the timing of the first intervention point (i.e., the start of the B phase) for each of the units. Even with response-guided experimentation, in which the emerging data pattern is taken into account, this should not be a problem; one could, for example, start the random assignment from the moment at which all baselines have been stabilized (Edgington, 1975, 1980; Ferron & Sentovich, 2002; Koehler & Levin, 1998). For multiple-baseline designs, several randomization test strategies have been suggested over the years.

In 1986, Wampold and Worsham presented a randomization test in which the subjects (or behaviors or situations) are assigned randomly to the different units, so that the order in which they are subjected to the treatment is determined at random but the intervention points are fixed for each separate unit. This way, the intervention points are placed according to the demands of the researcher and a staggered introduction of treatment can be obtained. For a design with *N* units, there are *N*! different ways in which the subjects, behaviors, or settings can be assigned to the different AB designs (Edgington, 1992).

In the strategy proposed by Marascuilo and Busk (1988), the start of the intervention phase is determined randomly for each unit on the basis of the rationale of Edgington (1975). This leads to k^N possible assignments (for a multiple-baseline design with *N* units and *k* possible start points for the intervention phase). The amount of control over the staggering of the interventions is less than with

the Wampold–Worsham (1986) approach, but the number of possible assignments will be larger, yielding a smaller possible *p* value (Ferron & Sentovich, 2002).

Koehler and Levin (1998, 2000) tried to combine the best of both worlds in their randomization test by randomly assigning the subjects (behaviors, or settings) to the different AB designs, as well as randomly assigning the start points of the intervention. This strategy results in

$$N!\prod_{i=1}^{N} k_i$$

possible assignments (for a design with N units and k_i possible start points for the ith unit, provided that there is no overlap between the possible start points of the different units).

Ferron and Sentovich (2002) found that the power of these suggested randomization tests is similar and therefore concluded that, because of the similarity in power, researchers should be able to focus on design considerations (and thus choose the test that best matches the design used in the study) when choosing among the alternative randomization strategies. Because the procedure suggested by Koehler and Levin (1998, 2000), by obtaining a systematic staggering of the interventions, best retains the integrity of the multiple-baseline design, we prefer their method. In the following, we will provide an R package for analyzing multiple-baseline data with the Koehler–Levin (1998) randomization test.

AN R PACKAGE FOR ANALYZING MULTIPLE-BASELINE DATA WITH RANDOMIZATION TESTS

Most of the widespread statistical software packages, like SPSS or SAS, do not include randomization tests for multiple-baseline data. Koehler and Levin (2000) devised their own program, RegRand, to calculate *p* values according to their regulated randomization procedure. This software program, however, is Macintosh-based and has no IBM PC-compatible version (Koehler & Levin, 2000). We already created an R package, SCRT-R, to perform randomization tests on data from single-case phase and alternation designs (Bulté & Onghena, 2008). Here, we present an extension for multiple-baseline AB data. R runs on a variety of UNIX platforms, as well as on Windows and MacOS (Hornik, 2008). As an open-source implementation of the S-PLUS language, it can be downloaded at no cost from the Comprehensive R Archive Network Web site (CRAN; cran.r-project.org). R is extremely flexible and can be used for statistical modeling as well as for graphical applications (Crawley, 2005; Dalgaard, 2002; Kelley, 2007).

Before being able to use the R functions, the R package needs to be installed. The installation process is very straightforward. Hornik (2008) gives a detailed description of how to download and install R for Windows, Macintosh, or UNIX. The R functions explained below can be found on ppw.kuleuven.be/cmes/SCRT-R.html. For easy access and use, we suggest saving the files containing the functions on the local disk; afterward, they can be read into R if you click on "File," choose "Source R Code," and select the appropriate file. To demonstrate the R functions, the example of Ziegler (1994) will be used again. Although in her original study, no random assignment procedure was used to determine the start points of the intervention phase for the different subjects, for illustrative purposes we will assume that such a random selection process did take place. Her data, extracted from the visual display in Figure 1, are provided in Table 1.

Table 1
Data Obtained in Ziegler's (1994) Experiment

Subject 1		Subject 2		Subject 3		Subject 4	
Phase	Score	Phase	Score	Phase	Score	Phase	Score
"A"	4	"A"	4	"A"	3	"A"	4
"A"	3	"A"	3	"A"	3	"A"	4
"A"	3	"A"	4	"A"	4	"A"	4
"A"	3	"A"	3	"A"	5	"A"	4
"A"	3	"A"	3	"A"	4	"A"	4
"A"	3	"A"	4	"A"	5	"A"	4
"B"	4	"A"	4	"A"	5	"A"	4
"B"	4	"A"	4	"A"	5	"A"	5
"B"	5	"B"	4	"A"	5	"A"	5
"B"	6	"B"	5	"A"	5	"A"	5
"B"	7	"B"	5	"B"	6	"A"	5
"B"	7	"B"	7	"B"	7	"A"	5
"B"	8	"B"	8	"B"	7	"B"	6
"B"	7	"B"	8	"B"	7	"B"	7
"B"	8	"B"	8	"B"	8	"B"	7
"B"	8	"B"	8	"B"	9	"B"	8
"B"	9	"B"	8	"B"	8	"B"	8
"B"	9	"B"	8	"B"	9	"B"	9
"B"	9	"B"	8	"B"	9	"B"	9
"B"	10	"B"	9	"B"	10	"B"	9
"B"	10	"B"	8	"B"	10	"B"	10
"B"	10	"B"	9	"B"	10	"B"	10
"B"	10	"B"	9	"B"	10	"B"	10
"B"	10	"B"	9	"B"	10	"B"	10

To guarantee that the R functions work properly, we suggest that researchers follow a few guidelines when creating the text (.txt) file containing the data. This data frame can be made most easily in a text editor (e.g., EditPad or Notepad) or in Excel, with the file saved as "text (tab delimited)." It should consist of two columns for each unit (if made in a text editor, separated by a tab): the first column with condition labels ("A" and "B"), and the second column with the obtained scores. This way, each row represents one measurement occasion. By contrast to Table 1, it is important not to label the columns or the rows. A plot similar to that in Figure 1 can be obtained by typing the command `graph(design="MBD")` into the R console, after which R will open a pop-up window to ask in what file the data to be graphed can be found.

When conducting a randomization test, the intervention start points should be chosen randomly, given the restrictions imposed by the researcher. For the R functions in this article, an additional text file should be created with the possible start points for each unit. Herein, each row should contain all possibilities for one unit, separated by a tab, so that there are as many lines in the file as there are units in the experiment. Each line (including the last one) should be closed by a return and, again, the rows and columns should not be labeled. The numbers given in Table 2 are

Table 2
Hypothetical Possible Start Points Associated With the Units in the Example

Unit	Possible Start Points
1	3 4 5 6 7
2	5 6 7 8 9
3	7 8 9 10 11
4	9 10 11 12 13

the possible intervention start points for the units in our example.

Of course, when one wants to assure a staggered administration of the intervention, no overlap between the possibilities of the different units can exist. With response-guided designs, one could for example start counting the measurement times, beginning with the first observation after stable baseline data have been obtained for all units. This way, when the randomly chosen start point of the intervention for the first unit is 5, the treatment should start on the 5th day after baselines have stabilized.

After the experiment has been conducted, the data can be analyzed statistically. With randomization tests, as indicated before, the *p* value is calculated by locating the observed test statistic within the randomization distribution. So, first the observed test statistic has to be calculated from the observed raw data. Then the randomization distribution is derived by calculating the test statistic for every possible permutation of the data. (Note that with this particular randomization test, the randomization distribution is not constructed by shuffling all observations, because that would arrange the data in an unrealistic order. With multiple-baseline AB designs, all A measurements precede all B measurements, so the only thing that can be shuffled is the start of the B phase.) Finally, the randomization test's *p* value can be calculated as the proportion of test statistics in the randomization distribution that exceeds or equals the observed test statistic. However, with multiple-baseline designs it is often not feasible to compute this *p* value by hand; because of the large data sets and a lot of possible permutations, it can become too cumbersome even for computers. To form an idea of the computational time needed to calculate the *p* value, the function `quantity(design="MBD")` can be used. This returns the number of possible assignments, given the possible start points for each unit, as if there were no overlap, according to the formula of Koehler and Levin (1998):

$$N! \prod_{i=1}^{N} k_i$$

(with N = number of units and k_i = number of possibilities for unit i). After typing this command, R will ask in which file the possible start points can be found. In our example, it resulted in 15,000 possible permutations. Since it would take a lot of computer time to calculate the "exact" *p* value that would use all of the test statistics, in this case it would be better to use the "Monte Carlo" version of the randomization test, which uses only a simulated distribution, to reduce the required calculation capacity (Besag & Diggle, 1977; Recchia & Rocchetti, 1982). The accompanying function is `pvalue.random`. Besides the already known `design` argument, it has three additional arguments: With `statistic`, the user can define the test statistic that should be used, by choosing `"A-B"`, `"B-A"`, or `"|A-B|"`. These are multivariate test statistics that stand for the (absolute value of the) mean difference between the condition means. If needed, other test statistics, such as differences in slopes or intercepts, can be adopted easily by means of small adjustments to the R script. Actually, any test statistic sensitive to the predicted treatment effect could be used with randomization tests. In our example, we expect a difference in level, which can be reflected by a difference between means. Because we expect the scores in the B phases to be larger than the scores in the A phases, we will use the directional test statistic `"B-A"`. With the `save` argument, one can indicate whether the randomization distribution should be saved into a file (`save="yes"`). And finally, the `number` argument serves to specify the required number of randomizations (e.g., 1,000). So, for our example, `pvalue.random(design="MBD",statistic="B-A",save="yes",number=1000)` should be typed into the R console; then, several pop-up windows will open. In the first, the location of the data should be specified, and in the second, the location of the possible start points. When `save` is set to `"yes"`, a third window will appear in which the user can indicate where the randomization distribution should be saved. This can be an existing file chosen by name from the list, or a new file that can be created by submitting the file name with a .txt extension. In this latter case, R will ask for confirmation ("The file does not exist yet. Create the file?"). When `save="no"`, this last window will not be shown. The resulting *p* value for our example is .006. This means that the mean difference between the phase means is statistically significant at a 5% level. The null hypothesis of no treatment effect can thus be rejected, as was already suggested by a visual inspection of the data. Note that, because a random sample of the test statistics is used with this Monte Carlo randomization test, this *p* value could be slightly different each time the function is used. If the number of possible assignments is smaller, the function `pvalue.systematic` can be used, with the same arguments, except here `number` is not needed. The Appendix gives an overview of all R functions explained above, with some additional ones.

When designing an experiment with a multiple-baseline design, the functions `assignments` and `selectdesign` can be convenient. With the first one, all possible permutations, given the potential start points for each unit provided by the user, can be displayed as output in the R console or saved into a file. Observe that R returns the possible combinations of start points for the units: For each unit, a start point is randomly chosen, and the combinations are then shuffled to randomly assign the different subjects, behaviors, or settings to the units. The second function, `selectdesign`, randomly selects one of these data arrangements on the basis of which data can be collected in the experiment. If knowing the ob-

served value of the test statistic is of interest, `observed` can be used. Finally, to display the randomization distribution (and if necessary save it to a file), the function `distribution.systematic` or `distribution.random`, depending on whether the exhaustive or the nonexhaustive variant is preferred, can be applied.

DISCUSSION

Although multiple-baseline designs are often used by applied researchers, to analyze the resulting data statistical techniques—and randomization tests in particular—are rarely applied. We believe that this is mainly because researchers are not acquainted with these analyses, and most widely used statistical software programs do not include the possibility of conducting randomization tests for multiple-baseline designs. By providing the rationale as well as an extension to our SCRT-R software package, we tried to fill this gap with this article. We hope this will encourage experimenters to supplement their visual analysis of the data with a statistical one, especially in instances where there is a lot of variability within the phases and where treatment effects cannot be easily detected with the naked eye.

Where the power of randomization tests with single-case phase AB designs is rather small, the combination of several of those AB experiments into a multiple-baseline design increases the power (Onghena & Edgington, 2005). Ferron and Sentovich (2002) showed that with 4 units and 20 measurement occasions, a large treatment effect ($d \geq 1.5$) can already be detected (power $> .80$). In our example, Ziegler (1994) used 4 units with 24 observations per subject. If we calculate the p value for each AB design separately, we find a p value of .33 for each of the units, which is not smaller than any commonly used significance level. For the randomization test on all units simultaneously, on the other hand, the p value equaled .006.

We stress again that in the original study, Ziegler (1994) did not use a random assignment procedure to decide the start points of the intervention phase; her example was used for illustrative purposes only. Although randomization tests are sometimes used when no random assignment has taken place, this is a practice we do not want to encourage, because it endangers the validity of those statistical tests. Without random assignment, the p value resulting from a randomization test has no probabilistic basis, and can only be tentatively interpreted as a descriptive ratio comparing an observed test statistic with reference test statistics that could have resulted from virtual, but plausible, alternative randomizations using the same kind of design (Edgington & Onghena, 2007; Winch & Campbell, 1969).

In this article, our intention is not to underestimate the importance of conducting a visual analysis of the data; and, although we focused on calculating the randomization test's p value, neither do we want to encourage researchers to see this as the only or ultimate technique for analyzing data from single-case studies. Not only should p values be interpreted with caution; because statistical significance does not always entail clinical significance, and the significance level of .05 is nothing but an arbitrarily chosen measure (Turk, 2000; Wilson, 2007), it also is only a part of the whole story. Attention should be paid to the magnitude or importance of the effect, not only to whether or not an observed effect is statistically significant (Robinson & Levin, 1997). Visual analysis, statistical significance, and effect size measures should be combined to obtain a comprehensive view of the results of the study. For the future, it would be interesting to investigate which measure of effect size is most suited for multiple-baseline data (d, percentage of nonoverlapping data, or others) and to develop tools in R to calculate these measures.

AUTHOR NOTE

The authors thank Susan G. Ziegler for giving permission to use her data and figure, and David C. Howell and an anonymous reviewer for their useful comments on an earlier version of this article. Correspondence concerning this article should be addressed to I. Bulté, Katholieke Universiteit Leuven, Centre for Methodology of Educational Research, Andreas Vesaliusstraat 2 bus 3762, B-3000 Leuven, Belgium (e-mail: isis.bulte@ped.kuleuven.be).

REFERENCES

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.

BAER, D. M., WOLF, M. M., & RISLEY, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, **1**, 91-97.

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 C*, **26**, 327-333.

BOX, G. E. P., JENKINS, G. M., & REINSEL, G. C. (1994). *Time series analysis: Forecasting and control* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

BULTÉ, I., & ONGHENA, P. (2008). An R package for single-case randomization tests. *Behavior Research Methods*, **40**, 467-478.

CHRIST, T. J. (2007). Experimental control and threats to internal validity of concurrent and nonconcurrent multiple-baseline designs. *Psychology in the Schools*, **44**, 451-459.

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.

DEPROSPERO, A., & COHEN, S. (1979). Inconsistent visual analysis of intrasubject data. *Journal of Applied Behavior Analysis*, **12**, 573-579.

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). Overcoming obstacles to single-subject experimentation. *Journal of Educational Statistics*, **5**, 261-267.

EDGINGTON, E. S. (1992). Nonparametric tests for single-case experiments. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 133-157). Hillsdale, NJ: Erlbaum.

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., & JONES, P. K. (2006). Tests for the visual analysis of response-guided multiple-baseline data. *Journal of Experimental Education*, **75**, 66-81.

FERRON, J., & SCOTT, H. (2005). Multiple baseline designs. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 3, pp. 1306-1309). New York: Wiley.

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

GORMAN, B. S., & ALLISON, D. B. (1996). 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. 119-158). Mahwah, NJ: Erlbaum.

HARRIS, F. N., & JENSON, W. R. (1985). Comparisons of multiple-baseline across persons designs and AB designs with replications: Issues and confusions. *Behavioral Assessment*, **7**, 121-127.

HARTMANN, D. P., GOTTMAN, J. M., JONES, R. R., GARDNER, W., KAZDIN, A. E., & VAUGHT, R. S. (1980). Interrupted time-series analysis and its application to behavioral data. *Journal of Applied Behavior Analysis*, **13**, 543-559.

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. (2008). *The R FAQ: Frequently asked questions on R.* Retrieved July 18, 2008, from cran.r-project.org/doc/FAQ/.

KAZDIN, A. E. (1973). Methodological and assessment considerations in evaluating reinforcement programs in applied settings. *Journal of Applied Behavior Analysis*, **6**, 517-531.

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., & KOPEL, S. A. (1975). On resolving ambiguities of the multiple-baseline design: Problems and recommendations. *Behavior Therapy*, **6**, 601-608.

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

KINUGASA, T., CERIN, E., & HOOPER, S. (2004). Single-subject research designs and data analyses for assessing elite athletes' conditioning. *Sports Medicine*, **34**, 1035-1050.

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.

LEITENBERG, H. (1973). The use of single-case methodology in psychotherapy research. *Journal of Abnormal Psychology*, **82**, 87-101.

LONG, C. G., & HOLLIN, C. R. (1995). Single case design: A critique of methodology and analysis of recent trends. *Clinical Psychology & Psychotherapy*, **2**, 177-191.

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

MARASCUILO, L. A., & BUSK, P. L. (1988). Combining statistics for multiple-baseline AB and replicated ABAB designs across subjects. *Behavioral Assessment*, **10**, 1-28.

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.

MORLEY, S., & ADAMS, M. (1991). Graphical analysis of single-case time series data. *British Journal of Clinical Psychology*, **30**, 97-115.

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.

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). New York: Wiley.

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

PARSONSON, B. S., & BAER, D. M. (1992). The visual analysis of data, and current research into the stimuli controlling it. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 15-40). Hillsdale, NJ: Erlbaum.

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

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.

SHADISH, W. R., COOK, T. D., & CAMPBELL, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

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

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

TRYON, W. W. (1982). A simplified time-series analysis for evaluating treatment interventions. *Journal of Applied Behavior Analysis*, **15**, 423-429.

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

WAMPOLD, B. E., & WORSHAM, N. L. (1986). Randomization tests for multiple-baseline designs. *Behavioral Assessment*, **8**, 135-143.

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

WINCH, R. F., & CAMPBELL, D. T. (1969). Proof? No. Evidence? Yes. The significance of tests of significance. *American Sociologist*, **4**, 140-143.

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

ZIEGLER, S. G. (1994). The effects of attentional shift training on the execution of soccer skills: A preliminary investigation. *Journal of Applied Behavior Analysis*, **27**, 545-552.

APPENDIX
Overview of the Different Functions, Their Specific Arguments, What They Do, and the Purpose of the Pop-Up Windows

Function	Arguments	Description	Windows
quantity	design: "MBD" (multiple-baseline design)	Calculates how many possible data arrangements there are for a specific design	1. Start points
assignments	design: "MBD" (multiple-baseline design) save: "yes" (save the possible assignments to a file) or "no" (display the possible assignments as output in the R console)	Generates all the possible data arrangements for a specific design	1. Start points 2. Save
selectdesign	design: "MBD" (multiple-baseline design)	Randomly selects one data arrangement among all theoretically possible permutations	1.Start points
graph	design: "MBD" (multiple-baseline design)	Makes a graphical representation of the data	1. Data
observed	design: "MBD" (multiple-baseline design) statistic: "A−B" (mean phase A minus mean phase B), "B−A" (mean phase B minus mean phase A), or "|A−B|" (absolute value of the difference between the phase means)	Calculates the observed test statistic from the obtained raw data	1. Data
distribution.systematic	design: "MBD" (multiple-baseline design) statistic: "A−B" (mean phase A minus mean phase B), "B−A" (mean phase B minus mean phase A), or "|A−B|" (absolute value of the difference between the phase means) save: "yes" (save the distribution to a file), or "no" (see the distribution as output in the R console)	Constructs the systematic randomization distribution under the null hypothesis	1. Data 2. Start points 3. Save
distribution.random	design: "MBD" (multiple-baseline design) statistic: "A−B" (mean phase A minus mean phase B), "B−A" (mean phase B minus mean phase A), or "|A−B|" (absolute value of the difference between the phase means) save: "yes" (save the distribution to a file), or "no" (see the distribution as output in the R console) number: how many randomizations are required	Constructs the random randomization distribution under the null hypothesis, where all the test statistics are calculated	1. Data 2. Start points 3. Save
pvalue.systematic	design: "MBD" (multiple-baseline design) statistic: "A−B" (mean phase A minus mean phase B), "B−A" (mean phase B minus mean phase A), or "|A−B|" (absolute value of the difference between the phase means) save: "yes" (save the distribution to a file), or "no" (see the distribution as output in the R console)	The statistical significance of the outcome is obtained by locating the observed test statistic in the randomization distribution	1. Data 2. Start points 3. Save
pvalue.random	design: "MBD" (multiple-baseline design) statistic: "A−B" (mean phase A minus mean phase B), "B−A" (mean phase B minus mean phase A), or "|A−B|" (absolute value of the difference between the phase means) save: "yes" (save the distribution to a file), or "no" (see the distribution as output in the R console) number: how many randomizations are required	The statistical significance of the outcome is obtained by locating the observed test statistic in the randomization distribution	1. Data 2. Start points 3. Save

(Manuscript received August 6, 2008;
revision accepted for publication October 1, 2008.)