KU LEUVEN
FACULTEIT PSYCHOLOGIE EN
PEDAGOGISCHE WETENSCHAPPEN

METHODOLOGIE VAN HET PEDAGOGISCH ONDERZOEK

# Being grateful for small mercies: The analysis of single-case data

**Isis Bulté**

Proefschrift aangeboden tot het verkrijgen van de graad van
Doctor in de Pedagogische Wetenschappen

Promotor: Prof. Dr. Patrick Onghena
Copromotor: Prof. Dr. Johan Vlaeyen

2013

ASSOCIATIE KU LEUVEN

**Isis Bulté, Being grateful for small mercies: The analysis of single-case data**

*Doctoral dissertation submitted to obtain the degree of Doctor in Educational Sciences, 2013*

Supervisor: Prof. Dr. Patrick Onghena. - Co-supervisor: Prof. Dr. Johan Vlaeyen

Despite the strong emphasis on large N designs in most statistical and methodological courses and handbooks, single-case designs are gaining popularity. They can provide a viable alternative or supplement to group studies to investigate causal research questions. This could be the case when it is unnecessary to collect data from a large group of subjects (e.g., generating pilot data) or when this is impossible (e.g., studying rare conditions). Although several data-analytic techniques have been suggested for the analysis of single-case experiments, visual analysis still remains the predominant method. An important reason is the absence of these techniques in popular statistical packages.

The focus of this doctoral research project is on providing user-friendly tools for analyzing single-case data. As a statistical technique, randomization tests were selected, because they are promising for analyzing the specific data resulting from single-case experiments. For the software development, the statistical programming language R was chosen. To enhance the user-friendliness, an implementation in the window-based graphical user interface, the R commander, was also created.

The developed software is presented in six separate manuscripts and contains functions for three important steps in analyzing single-case data: visual exploration, statistical significance testing, and effect size calculation. *Manuscript 1* presents the "SCVA" package, which helps researchers in making graphical representations of single-case data and to transform graphical displays back to raw data. *Manuscript 2* introduces the "SCRT" package for designing single-case phase and alternation experiments, as well as for conducting randomization tests on data resulting from these experiments. In *manuscript 3*, a systematic review on the use of multiple baseline designs is provided. Since it was found that most multiple baseline studies in the sample data relied solely on visual analysis of the data, in *manuscript 4* an extension of the "SCRT" package is presented for use with multiple baseline data. *Manuscript 5* explains the use of the "SCMA" package, with functions for probability combining and effect size calculation. In *manuscript 6*, all functionalities are combined in the "RcmdrPlugin.SCDA" package: a window-based graphical user interface, which overcomes the difficulties of the command line interface of R.

By making these data-analytic tools available and assemble them in one overarching user-friendly system, we wish to contribute to their familiarity with applied researchers, and hence their increased use in empirical studies. The scripts could also be used in stochastic simulation studies, and in this way add to the knowledge on the statistical properties of effect size measures and statistical techniques for single-case data.

**Isis Bulté, Wie het kleine niet eert: Het analyseren van single-case data**

*Proefschrift aangeboden tot het verkrijgen van de graad van Doctor in de Pedagogische Wetenschappen, 2013*

Promotor: Prof. Dr. Patrick Onghena - Co-promotor: Prof. Dr. Johan Vlaeyen

---

Ondanks de sterke nadruk op grootschalig onderzoek in de meeste statistische en methodologische handboeken en cursussen, winnen single-case designs aan populariteit. Ze kunnen een alternatief of nuttige aanvulling vormen op groepsvergelijkend onderzoek, wanneer men causale onderzoeksvragen wenst te beantwoorden. Situaties waarin geopteerd kan worden voor een single-case design zijn die waarin het niet nodig is om gegevens van een grote groep te verzamelen (bvb. het genereren van pilootgegevens) of situaties waarin dit niet mogelijk blijkt te zijn (bvb. het bestuderen van zeldzame aandoeningen). Hoewel in het verleden verschillende data-analytische technieken werden voorgesteld om single-case data te analyseren, blijft visuele analyse de meest gebruikte methode. Een belangrijke reden hiervoor is het ontbreken van technieken voor single-case data in de populaire statistische software pakketten.

De focus van dit doctoraatsproject ligt op het leveren van gebruiksvriendelijke instrumenten voor het analyseren van single-case data. Als statistische techniek werden randomiseringstoetsen gekozen, omdat deze veelbelovend zijn voor het analyseren van data met specifieke kenmerken eigen aan single-case experimenteel onderzoek. De software werd ontwikkeld in de statistische programmeertaal R. Om de gebruiksvriendelijkheid te vergroten, werd ook een implementatie in de R commander, een grafische gebruikersomgeving (GUI), voorzien.

De ontwikkelde software wordt voorgesteld aan de hand van zes afzonderlijke manuscripten. Ze omvat functies voor drie belangrijke stappen in het analyseren van single-case data: visuele exploratie, statistisch toetsen en het berekenen van effectgroottes. *Manuscript 1* presenteert het “SCVA” pakket dat onderzoekers helpt bij het maken van een grafische weergave van hun gegevens en bij het omzetten van grafieken naar ruwe data. *Manuscript 2* introduceert het “SCRT” pakket dat gebruikt kan worden voor het ontwerpen van single-case fase- en alternatiedesigns, en voor het uitvoeren van randomiseringstoetsen op de gegevens afkomstig uit deze experimenten. In *manuscript 3* wordt een systematische review over het gebruik van multiple baseline designs beschreven. Aangezien hieruit bleek dat in onderzoeken met deze designs bijna uitsluitend visuele data-analyse gebruikt werd, wordt in *manuscript 4* een uitbreiding van het “SCRT” pakket voor multiple baseline data voorgesteld. *Manuscript 5* beschrijft het gebruik van het “SCMA” pakket met functies voor het combineren van overschrijdingskansen en het berekenen van effectgroottes. En in *manuscript 6* worden alle mogelijkheden gecombineerd in het “RcmdrPlugin.SCDA” pakket: een GUI die de moeilijkheden van het werken in de R omgeving omzeilt.

Door deze instrumenten voor data-analyse ter beschikking te stellen en deze samen te voegen in een overkoepelend gebruiksvriendelijk systeem, willen we hun bekendheid bij toegepaste onderzoekers verhogen, en op die manier hun gebruik in empirisch onderzoek aanmoedigen. De R scripts kunnen ook gebruikt worden in stochastische simulatiestudies en zo bijdragen aan de kennis met betrekking tot de statistische eigenschappen van maten voor effectgrootte en statistische technieken voor single-case data.

# Dankwoord

Patrick en Johan, bedankt voor jullie deskundige begeleiding. Jullie waardevolle suggesties en opmerkingen hebben dit proefschrift mee vorm gegeven. Patrick, bedankt voor de grote vrijheid die je mij gaf en voor het blijvende vertrouwen in het slagen van dit project. Johan, bedankt voor je niet aflatende en uiterst aanstekelijke enthousiasme.

I thank the members of my supervisory committee Prof. Dr. Eva Ceulemans, Prof. Dr. Geert Crombez, Prof. Dr. Pol Ghesquière, Prof. Dr. Dirk Hermans, and Prof. Dr. David Howell for their valuable suggestions and feedback. I am honored that Prof. Dr. Eva Ceulemans, Prof. Dr. Dirk Hermans, and Prof. Dr. Rumen Manolov accepted to be jury members of my doctoral defense.

Veel dank ook aan de fantastische ex-collega's van het CMPO/CWP, die ondertussen vrienden geworden zijn. The Gang, bedankt voor de bijzonder aangename werksfeer, de fijne tijden en de steun in moeilijke momenten.

Vrienden en familie, bedankt voor jullie steun en vertrouwen, voor de ontspannende momenten buiten de werkuren, voor de nodige relativering, en voor zoveel meer.

Marjo, bedankt om er steeds weer te zijn voor mij…

# Table of contents


**GENERAL INTRODUCTION** **1**

**MANUSCRIPT 1- SINGLE-CASE VISUAL ANALYSIS** **17**

**MANUSCRIPT 2 – SINGLE-CASE RANDOMIZATION TESTS** **45**

**MANUSCRIPT 3 – MULTIPLE BASELINE DESIGNS: SYSTEMATIC REVIEW** **78**

**MANUSCRIPT 4 – MULTIPLE BASELINE RANDOMIZATION TESTS** **94**

**MANUSCRIPT 5 – SINGLE-CASE META-ANALYSIS** **116**

**MANUSCRIPT 6 – SINGLE-CASE DATA ANALYSIS** **134**

**GENERAL DISCUSSION** **166**

**REFERENCES** **176**

**APPENDIX** **207**

# General introduction

This doctoral dissertation focuses on the development of a user-friendly software package for analyzing single-case data. The presented R packages include functions for visual analysis, statistical significance testing and effect size calculation.

*First*, a general introduction of single-case experimental designs is provided. A short historical overview of the position of the individual in research is followed by a definition of single-case experimental designs. Also methodological challenges in collecting and analyzing single-case data are looked at.

*Second*, the development of the software is presented in six separate manuscripts. The first manuscript discusses the visual analysis of single-case data. In manuscript three, a systematic review of the use of multiple baseline designs is provided. Manuscripts two and four deal with the use of nonparametric randomization tests to statistically analyze single-case (multiple baseline) data. Manuscript five includes probability combining and the calculation of effect size measures on single-case data. And in manuscript six, all functionalities are combined in a window-based graphical user interface.

*Third*, the doctoral dissertation is completed with a general discussion in which the research project is evaluated. Possible extensions of the software package are suggested, and perspectives for future research are given.

# Single-case experimental designs

## The role of the individual: A historical perspective

At the end of the 19th and the beginning of the 20th century research was mainly focused on the intensive study of the individual. Researchers used only one or a few participants to draw conclusions, and some of these findings are still very relevant. Famous examples in psychology and education are Paul Broca, who identified the brain's speech centre by intensively studying one patient with a speech problem; Hermann Ebbinghaus, who discovered the forgetting curve by studying the efficiency of his own memory; Ivan Pavlov, who laid out the basic principles of classical conditioning based on the study of single dogs; J. B. Watson, who found evidence of classical conditioning in human beings with his case study on the acquisition of fear of pets in Little Albert; and of course Sigmund Freud, who published several case studies of which Anna O is probably the most well-known (see e.g., Barlow, Nock, & Hersen, 2009; Kazdin, 2011).

In the course of the 20th century this dominant position of small-scale individual studies shifted to a focus on large-scale group studies. A responsible underlying discovery was Quetelet's notion that several human traits are normally distributed. This led to the desire of estimating "l'homme moyen" or "the average man", for physical (e.g., height) as well as psychological (e.g., intelligence) characteristics. These attempts were facilitated by the development of descriptive statistics, for example by Galton and Pearson. Also

Fisher's work, with the invention of several sophisticated statistical procedures to make comparisons within and between groups (e.g., ANOVA), was, of course, of considerable interest in further expanding the study of individual differences. Furthermore, Fisher elaborated the possibility to make general statements about the distribution of characteristics in the population by making inferences from the studied sample. All these developments led to an emphasis on averages and intersubject variability and consequently observations of single organisms became rather irrelevant. By the 1950s, the group comparison approach had dethroned the study of the single case (see e.g., Barlow et al., 2009; Kazdin, 2011).

Because of several difficulties however, skepticism arose on the clinical usefulness of this early group comparison approach. Besides that it is not always feasible to find a large number of participants, for example with rare or unique conditions or because of economical considerations, it could also be considered unethical to withhold treatment from the control group. Most importantly, individual participants' results are lost in group averages, which makes them irrelevant for clinicians since this global result does not reflect which patient characteristics are related to which therapy response. Also the clinical course of a patient through treatment is often ignored, while this intraindividual variability is significant to practitioners. This all contributed to a renewed attention to the individual. The first response was a wave of correlational approaches (e.g., process research and naturalistic studies), based on which no causal relations could be inferred, which contributed to the so-called scientist-

practitioner split: the prevailing clinical research was considered unacceptable by most scientists, and the group comparison approach was seen as unsuited for studying applied problems. Luckily, the experimental single-case approach came as a compromise, balancing the focus on the individual and the experimental approach (see e.g., Barlow et al., 2009; Kazdin, 2011).

## Single-case experimental designs: a definition

Onghena and Edgington (2005) define a single-case experiment as "*a designed experiment in which one entity is observed repeatedly during a certain period under different levels ("treatments") of at least one independent variable*". Since it is an experiment, there has to be a deliberate manipulation of the independent variable (e.g., the introduction of a treatment), which allows making causal inferences. This clearly distinguishes single-case experiments from case studies and observational time series studies, which only permit to determine associations between variables. "Single-case" does not necessarily mean that only one person is involved; "one entity" could also refer to one family, one classroom, etc.… that is subjected to all experimental conditions. Important is that repeated measurements are taken, since no inferences can be made based on only one observation. Repeated observations also provide insight in what happened during the intervention and thus lead to more information than sheer pre-post treatment observations (e.g., cyclical patterns, improvement in the baseline phase, impact of major life events,…) (Barlow et al., 2009).

Just as in large-scale group studies, randomization is important in single-case experiments to be able to make valid causal inferences. Whereas in group studies randomization is usually accomplished by randomly assigning participants to experimental groups, in single-case experiments the measurement occasions can be randomly assigned to experimental conditions. This increases the *internal validity* of the study by excluding alternative explanations for the observed effects. For example, a lot of external (e.g., winning the lottery) and internal (e.g., spontaneous recovery) events could occur together with the introduction of the treatment, which could also explain the observed (lack of) behavioral change (Edgington, 1996; Kratochwill & Levin, 2010; Onghena & Edgington, 2005; Dugard, File, & Todman, 2012). Two types of single-case experimental designs in which this randomization can be easily incorporated are phase designs and alternation designs. In phase designs the measurement occasions are separated into blocks of experimental conditions, usually comparing baseline and intervention phases. Randomization can be accomplished by randomly determining the moment(s) of phase change. In alternation designs the experimental conditions are alternated along the measurement occasions. Randomization here occurs very similar as in group studies: for every measurement occasion it is determined at random which experimental condition will be present (cfr. assigning participants to experimental groups), so actually, the treatment order is determined randomly. An overview of more single-case experimental design types can be found, for example, in Barlow et al. (2009), Franklin, Allison, and Gorman (1997), Kazdin (2011), and Kratochwill and Levin (1992).

To demonstrate the *external validity* of an experiment, replications are required, since one cannot just generalize the observed results of a single case to other cases. These repetitions could be carried out one by one or at the same time. Familiar is the multiple baseline design: an extension of the basic single-case phase design, in which several of those phase designs are implemented simultaneously to different persons, behaviors, or settings. A characteristic feature is that the intervention is introduced in a staggered way to the different units (Bulté & Onghena, 2009; Ferron & Scott, 2005).

## Regained popularity

One of the instances in which these single-case experimental designs can be very useful, is in the gathering of pilot data in early research stages. An example of such a relatively new developing research area is the use of a graded exposure in vivo treatment for the reduction of pain-related fear. This treatment was developed based on the idea that chronic pain and phobia share important characteristics. The cognitive-behavioral model of fear of movement/(re)injury (Vlaeyen, Kole-Snijders, Boeren, & van Eek, 1995) predicts that when individuals with an acute pain episode catastrophically misinterpret their pain as a sign of serious tissue damage, they are likely to respond with pain-related fear. This fear causes them to engage in escape and avoidance behavior, which in turn leads to the maintenance of fear, a reduction of painful activities, and an augmentation of the disability. With a combination of education to illustrate this vicious circle, the determination of a personal graded hierarchy of avoided situations, and ultimately the actual exposure to

these feared activities, fear reduction can be achieved (Vlaeyen, Morley, Linton, Boersma, & De Jong, 2012). In order to investigate the applicability of this treatment to different pain problems, it was deemed useful to start collecting small-scale pilot data before engaging in large-scale RCT's. Also, the context of these therapies is very well-suited for the methodology of singe-case experiments, because of the need of an idiographic and customized approach in determining the treatment goals and developing the fear hierarchy. In addition, the treatment is set up as a behavioral experiment in which the patient performs activities to challenge the validity of his/her misinterpretations. Usually, -and also a common practice in cognitive behavior therapy- a diary is kept to be able to individually monitor the effectiveness of the treatment. This way, the need for repeated measurements is not considered as an additional burden, but rather as a logical part of the treatment. Mostly AB designs are used, in which diary observations are collected during pre-treatment (A) and treatment (B) phases. This basic design is sometimes extended to an ABCD design, with 6 and 12 month follow-up measurements. In an average therapy situation, different patients are seen consecutively. By combining several of those 'single-case experiments' using meta-analytical techniques, the external validity as well as the statistical conclusion validity of single-case designs could be improved (Onghena & Edgington, 2005).

Single-case experimental designs are clearly regaining popularity. This is visible in the appearance of special issues of applied journals dedicated to single-case research, for example in 'Remedial and

Special Education' (2013) and in the 'Journal of Applied Sport Psychology' (2013). Shadish and Sullivan (2011) reviewed all single-case designs published in 21 journals in the field of psychology and education in 2008. They found that in the category of intervention studies, those designs were more prevalent than group studies. 10.3% of the articles in the journals of the focus reported the use of single-case designs, of which 54.3% were multiple baseline designs, 8.2% phase designs and 8% alternation designs (the rest were combinations of design types). Also Hammond and Gast (2010) found that single-case designs are more represented in the literature than in the past decades, but group studies are still the majority.

## Challenges for analyzing single-case data

Not only the current spirit of the times, with still a preference for large scale group studies, but also several methodological challenges in collecting and analyzing single-case data lie at the basis of the lesser use of single-case designs. A lot has to do with the nature of the data: repeated observations taken from the same entity.

A first consequence is that reliable and stable methods of data collection should be used, which can be administered repeatedly. They must be sensitive to slight changes in behavior over time, as well as robust against reactivity; the effects of repetitive administration of the same test (e.g., learning effects). In single-case experiments collecting data often requires an active input of the subject and/or his environment, whether primary measures (e.g., frequency of occurrence of thumb sucking) or indirect observations (e.g., self-reports through questionnaires) are used. Disadvantages of

diary methods are that the data collection itself could act as an intervention that could alter behavior (e.g., when counting the number of cigarettes smoked daily), and that one should trust on introspection (e.g., risk of social desirability) (Barlow et al., 2009). Often when observational measurements are used, however, this is partly resolved by using multiple sources and calculating interrater reliability (Smith, 2012). Another problem is that the observational data collected often do not follow a normal distribution, which could cause a problem for many parametric statistical test. Almost all outcome measures in the Shadish and Sullivan (2011) study, for example, were count data.

A second and very important consequence of taking repeated measures over time is that serial dependence is present in the data. Serial dependence, or autocorrelation, reflects the interrelationships amongst subsequent data points, or, more specifically, the fact that the residuals are correlated (Barlow et al., 2009; Kazdin, 2011). There has been considerable debate on whether autocorrelation is present in single-case data. In his review, Huitema (1985) concluded that autocorrelation is rarely an issue in applied research, but this finding could have been biased because of the inclusion of studies with a limited number of observations. His result was contradicted by several other authors with estimates of 20 to 80% of the studies containing serial dependence (e.g., Busk & Marascuilo, 1988; Jones, Weinrott, & Vaught, 1978; Matyas & Greenwood, 1991; Matyas & Greenwood, 1997; Shadish & Sullivan, 2011; Sideris & Greenwood, 1997). The presence of autocorrelation has as a result that many

familiar statistical techniques, often developed for group comparison research, cannot readily be used. Conventional *t* and *F*-tests, for example, rely on many assumptions about the data of which the assumption of uncorrelated error terms cannot be violated without problems. Positive autocorrelation increases the risk of Type I errors, whereas negative autocorrelation more Type II errors tend to be made (Barlow et al., 2009; Kazdin, 2011). Also recent simulation studies have shown that autocorrelation has an impact: Manolov and Solanas (2008) demonstrated that effect sizes are linearly related to autocorrelation, and Smith, Borckhardt, and Nash (2012) found that an autocorrelation of .08 already negatively affects the ability to correctly infer significant level-changes using standardized mean difference methods. Serial dependence should thus be taken into account when statistically analysing the data. This could be done, for example, by removing the internal structure of the time series data with ARMA/ARIMA modelling before calculating effect size measures etc. These, however, require at least 50 observations per phase to be able to adequately identify the underlying models (Box, Jenkins, & Reinsel, 1994).

For single-case researchers in the social sciences, it is however not often feasible to collect these large numbers of repeated measurements. Especially when a substantial effort of the participant is required (e.g., when using self-report measures), this would be quite a burden. A careful balance must be made between having enough observations to obtain a representative sample and preventing study drop out. The more repeated measures requested,

the higher the risk of being confronted with missing data (e.g., because of forgetfulness, not having access to the data collection tool, intentional noncompliance,...) (Barlow et al., 2009; Smith, 2012). According to Shadish and Sullivan (2011), the number of observations in single-case studies ranges from 2 to 160, with a mean of 20, and with 90.6% having 49 for fewer measurements. This could not only cause problems for statistical analysis techniques such as time series analysis and multilevel modelling, but also for acquiring a representative baseline. The baseline in single-case experiments is characterized by the absence of the intervention and serves the same purpose as the no-treatment control group in group comparison studies. To serve as a good standard for comparisons when visually analyzing the data, baseline data should be relatively stable, without significant trends (particularly not in the hypothesized direction of the effect) and with minimal data overlap with the subsequent phase. Stability could sometimes be reached by extending the baseline, but this holds some practical and ethical limitations (e.g., patients get discharged, treatment is withheld,...) (e.g., Barlow et al., 2009; Franklin, Gorman, Beasley, & Allison, 1997; Parsonson & Baer, 1992). To prove that stability in the baseline pattern is reached, and thus that trends and extreme variability are absent, a minimum number of observations is required. Kratochwill et al. (2010), for example, recommend that a baseline phase includes at least five observations. Less observations also leaves only little data available to compare baseline and treatment data (for statistical inference or effect size calculation). In practice, this however does not always seem to be the case. Only 54.7% of the studies in the sample of

Shadish and Sullivan (2011) had at least five baseline observations. Smith (2012) found an average of 10.22 baseline observations, with 6 as modal number. Kazdin (2011) suggested using statistical analysis when baseline data contain trends or significant variability.

However, statistical analysis of single-case data still seems to be the exception rather than the rule. Less than 10% of the studies reviewed by Kratochwill and Brody (1978) used some kind of statistical analysis. This was confirmed by Busk and Marascuilo (1992). Brossart, Parker, Olsen, and Mahadevan (2006) estimated that this number had doubled by the year 2006. Smith (2012) reviewed all single-case experiments published in peer-reviewed journals between 2000 and 2010 and found an upward trend in the use of statistical analysis. However, visual analysis still clearly remains the predominant method of data analysis. Possible explanations are that statistical testing is still considered as unnecessary by some single-case researchers, partly because they are not familiar with their use, and that there exist no straightforward guidelines concerning which technique to use under which circumstances. For the analysis of single-case experiments, several data-analytic techniques have been suggested of which a nonexhausted overview is given in, for example, Barlow et al. (2009), Gorman and Allison (1997), Kazdin (2011), and Kratochwill and Levin (1992). The absence of these techniques in the popular statistical packages (e.g., SAS, SPSS) is without any doubt of considerable importance. This software includes various statistical techniques for evaluating between-group data, but, as indicated above, these tests cannot always simply be applied to single-case

data. For some of the proposed single-case techniques specific software is available, but no generic software package exists that has the scope of coverage and the ease of use of the more familiar statistical packages (Barlow et al., 2009; Kazdin, 2011).

# Doctoral research project

## Project focus and methodology

The focus of this doctoral research project is on providing user-friendly tools for analyzing single-case data. By making tools for data analysis available and assemble them in one overarching user-friendly system, we want to contribute to their familiarity with applied researchers and hence their use in empirical studies. The scripts could also be used in stochastic simulation studies, and in this way add to the knowledge on the statistical properties of effect size measures and statistical techniques for single-case data.

As statistical technique, randomization tests were selected. This choice was not arbitrarily; randomization tests hold many promises for analyzing the specific data resulting from single-case experiments: their results are not invalidated by the presence of autocorrelated residuals; they are not based on distributional assumptions (e.g., normality, homogeneity of variances); and they can be used with different types of data (e.g., counts, ranks,...) and data patterns (e.g., baseline). The only requirement is the incorporation of random assignment in the study design (see e.g., Dugard et al., 2012; Edgington, 1975, 1980b; Edgington & Onghena, 2007; Gorman & Allison, 1997; Onghena & Edgington, 2005).

For the development of the software package, the statistical programming language R was chosen (cran.r-project.org). Reasons for this selection are that it is open source and thus free of charge, that it runs on a variety of UNIX platforms, as well as on Windows and MacOS, that it has excellent graphical possibilities, and that it is a very powerful and flexible statistical environment (e.g., Crawley, 2005; Hornik, 2012). To enhance the user-friendliness of the software, an implementation in the window-based graphical user interface, the R commander (Fox, 2005), is anticipated. Unlike the command line interface of the basic R console, the R commander uses menus and dialog boxes, which are far more familiar to most since it looks very similar to, for example, SPSS. With the possibility of adding plug-in packages to the R commander (Fox, 2007), it can easily be extended to include more single-case data analytical techniques.

## Project outcomes: Development of software

The developed software contains functions for three important steps in analyzing single-case data: visual exploration, statistical testing, and effect size calculation. The separated packages were submitted to the Comprehensive R Archive Network (CRAN), to make them wider and more easily available for potential users. They were also introduced at international conferences and in several papers. The papers are presented in the next chapters of this dissertation.

Manuscript 1 presents the SCVA package (*Single-Case Visual Analysis*), which helps researchers in making graphical representations of single-case data and to transform graphical displays back to raw data (Bulté & Onghena, 2012). A basic plotting

function is provided, as well as some tools to facilitate the use of three interpretative principles for visually analyzing single-case data: plotting a measure of central location, displaying variability, and displaying trend. Also included is a function to extract raw data values from published graphs.

Manuscript 2 introduces the SCRT package (*Single-Case Randomization Tests)*, for 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 resulting from these experiments. The statistical technique included in this package is a nonparametric randomization test (Bulté & Onghena, 2008).

In manuscript 3, a systematic review on the use of multiple baseline designs is provided. More specifically the use of some design features (type of multiple baseline design, number of units, number of measurement times, and placement of the intervention point) and the techniques used to analyze the resulting data are considered. It was found that multiple baseline designs are still relatively popular in the applied behavioral sciences, but most of the studies in the review relied solely and visual analysis of the data.

Manuscript 4 therefore presents an extension of the SCRT package, in which all functions are expanded for use with data resulting from multiple baseline designs (Bulté & Onghena, 2009).

In manuscript 5, the use of the SCMA (*Single-Case Meta-Analysis)* package is presented (Bulté, Van den Noortgate, Heyvaert, &

Onghena, submitted). This package includes functions for probability combining, using two methods for the nonparametric combination of *p*-values: the multiplicative approach, using Pearson's formula, and the additive approach, using Edgington's method. Furthermore, several effect size measures can be calculated on single-case experimental data: the (pooled) standardized mean difference, the percentage of nonoverlapping data (PND), and the percentage of data points exceeding the median (PEM).

In the final paper, manuscript 6, all functionalities are combined in the RcmdrPlugin.SCDA package (*Single-Case Data Analysis*). This overcomes the difficulties of the command line interface of R, which is rather hard for practitioners who have never engaged in any basic programming. The SCDA package is a plug-in for the R commander: a window-based graphical user interface to R, with menus and dialog boxes, like, for example, SPSS (Bulté & Onghena, submitted).

# When the Truth Hits You Between the Eyes: A Software Tool for the Visual Analysis of Single-Case Experimental Data


## Abstract

Visual data analysis is an important first step when evaluating intervention effects. This also holds for analyzing data from single-case experiments. Because most software packages do not offer customized facilities for constructing single-case graphs and are not particularly suited to perform single-case visual data analyses, we created an R package to help researchers in making graphical representations of single-case data and to transform graphical displays back to raw data. In addition to a basic plotting function, we included some tools to facilitate the use of three interpretative principles for visually analyzing single-case data: plotting a measure of central location as a horizontal reference line; displaying variability with (trimmed) range bars, range lines, and trended ranges; and displaying trends with a vertical line graph, by fitting a robust linear trend, or by plotting running medians. Finally, we included a function to extract raw data values from published graphs.

# Introduction

Visual analysis is the primary method of evaluating data from single-case experiments (Busk & Marascuilo, 1992; Kahng et al., 2010; Kazdin, 2011; Parsonson & Baer, 1992). Although several statistical alternatives have been suggested over the years (for an overview, see, e.g., Campbell & Herzinger, 2010; Gorman & Allison, 1997; Houle, 2009), these alternatives are only rarely used (e.g., Brossart, Parker, Olson, & Lakshmi, 2006; Carter, 2009; Kahng et al., 2010). When looking at trends in single-case research, Kratochwill and Brody (1978), Busk and Marascuilo (1992), and Long and Hollin (1995) indicated that in less than 10% of their sampled studies statistical analyses were used. Parker et al. (2005) reanalyzed published single-case data and found over 65% of the studies in their sample relying on visual analysis alone. Effect sizes, confidence intervals, and tests of statistical significance were found in only 11% of the articles. A reason for this difference with group studies, in which statistical analyses are most common, lies in the two conflicting traditions from which single-case and group studies arose (Allison, Franklin, & Heshka, 1992). Scientific group study research relies on making inferences based on formal statistical procedures, largely based on the work of Ronald Fisher and Jerzy Neyman. Single-case designs and applied behavior analysis have their roots in the work of Murray Sidman and Burrhus Frederic Skinner, who rejected statistical analysis and relied solely on visual analysis of the graphed data (Allison et al., 1992). This dominance of visual analysis in published single-case research does not necessarily mean that it is generally

accepted to be the best data-analytic technique available. In past research, questions have been raised concerning the consistency (or reliability), the sensitivity, and the specificity of visual analysis.

## Consistency

Visual analysis is often criticized for having no empirically established formal decision guidelines, leaving room for subjectivity and inconsistency. This is demonstrated by the average poor interrater agreement that is found in several studies (e.g., Bobrovitz & Ottenbacher, 1998; DeProspero & Cohen, 1979; Fisch, 1998; Jones et al., 1978; Ottenbacher, 1990; Park, Marascuilo, & Gaylord-Ross, 1990; Ximenes, Manolov, Solanas, & Quera, 2009). This discouraging low consistency was not confirmed by all researchers, with, for example, Kahng et al. (2010) reporting a high level of agreement among different judges when replicating the 1979 DeProspero and Cohen study. Potential reasons for these contradicting results are procedural differences in the instructions and response measures used, and the fact that the earlier results could be outdated because of more and better training opportunities that exist for judges nowadays. Also, the alarming results found in earlier studies could be partly explained by the difference between real-life settings and the artificial settings used in the studies (Brossart, Parker, Olson, & Lakshmi, 2006; Parsonson & Baer, 1992).

Several variables were shown to increase interrater agreement: Training judges in using a standard trend estimation procedure, prior knowledge of the participant of whom the data are graphed, and the use of judgmental aids (Skiba, Deno, Marston, & Casey, 1989). The

area of expertise seems to have an influence, with statistically trained judges showing more agreement than single-case analysts (Harbst, Ottenbacher, & Harris, 1991). Some characteristics of the data also play a role: changes in mean shift and level across phases seem to yield higher agreement among judges than changes in variability and slope (DeProspero & Cohen, 1979; Gibson & Ottenbacher, 1988; Knapp, 1983), whereas data overlap and variability between phases seems to have little influence (Gibson & Ottenbacher, 1988). The effect of serial dependency is not clear. It is thought that autocorrelation negatively influences agreement among judges (Jones et al., 1978; Matyas & Greenwood, 1990), but some studies only found weak effects (Gibson & Ottenbacher, 1988).

## Sensitivity and Specificity

Proponents of visual analysis claim this method to be more conservative and less sensitive than statistical analysis (Kazdin, 2011). This insensitivity is seen as an advantage, because small treatment effects will be ignored and only very large effects, which are most likely also clinically (or practically) relevant, will be detected (e.g., Carter, 2009; Jones et al., 1978; Ottenbacher, 1990; Park et al., 1990; Parsonson & Baer, 1992). Overlooking consistent but relatively weak effects could however be problematic, for instance in developing new theories or technologies (Kazdin, 2011). Not all studies confirmed this claimed conservatism. Bobrovitz and Ottenbacher (1998) found a high agreement between the results of visual and statistical analyses, concluding that both methods are equally sensitive and specific. Fisch (1998) found very high Type I error rates, and Matyas and

Greenwood (1990) and Normand and Bailey (2006) found that participants were more likely to detect nonexistent effects (Type I errors) than they were to omit existing effects (Type II errors).

Since findings up till now are inconsistent and contradictory, concerns about the error rates in visual analysis remain, and questions about the quality of the validity studies have been raised. One issue involves the unrealistic conditions that are used, because validity studies differ in the extent to which they resemble natural conditions. In real-life situations, visual analysts have to make their judgment in a particular context and can judge the degree of effectiveness of a given intervention instead of being forced to make a dichotomous yes/no decision. Besides the available response options, also characteristics of the judges (e.g., trained/untrained) and the way of graphically presenting the data (e.g., addition of trend lines) could be confounding factors and more research is definitely needed (Brossart, Parker, Olson, & Lakshmi, 2006; Carter, 2009; Furlong & Wampold, 1982; Parsonson & Baer, 1992; Rojahn & Schulze, 1985; Wampold & Furlong, 1981). Allison et al. (1992) state that Type I error rates could even be larger in real-life applications because of the use of response-guided experimentation. If each data point is judged instantly and separately to determine how to continue the experiment, these multiple analyses increase the probability of making at least one Type I error. Allison et al. (1992) therefore correct the estimate of 10% for Type I error rates in visual analysis to 25%. Ferron and Jones (2006) presented a method to

control these Type I error rates, by using random assignment and a 'blind' data analyst.

A second issue concerns the validation criteria that are used in the studies. Some studies (e.g., Normand & Bailey, 2006) constructed data sets and graphs with known characteristics. However, many studies compare the results of visual analysis with those of statistical analysis, like the split-middle method (e.g., Ottenbacher, 1990; Richards, Taylor, & Ramasamy, 1997), time-series analysis (e.g., Jones et al., 1978), and randomization tests (e.g., Park et al., 1990) and different statistics tend to yield different conclusions. Since no consensus exists on which statistical test to use, without a power analysis none of them can be taken as the standard to evaluate the performance of visual judges (Brossart, Parker, Olson, & Lakshmi, 2006; Parsonson & Baer, 1992; Ximenes et al., 2009). For experiments with relatively large sample sizes, well-established procedures exist to determine the effects of the intervention. Provided that some assumptions about the population are met, parametric statistical tests (e.g., t-tests and ANOVAs) can be used validly. For small-n experiments the same procedures as for large-n designs do not necessarily apply because there are more doubts about the plausibility of the assumptions. The most important assumption that might be violated is that of serial independence (Kazdin, 2011). Autocorrelation might however also complicate the interpretations when visually analyzing single-case data, since it can be confounded with the existence of a treatment effect (e.g., Matyas & Greenwood, 1990; Rojahn & Schulze, 1985). Several statistical

techniques were proposed for analyzing single-case data (for an overview, see, e.g., Campbell & Herzinger, 2010; Gorman & Allison, 1997; Houle, 2009), but all were criticized for various reasons (e.g., they pose limitations on the researchers, they tend to ignore clinical significance,...), and there is still some debate on which of them is suitable. Several effect size indices based on visual analysis criteria have also been developed (e.g., Ma, 2006; Parker, Hagan-Burke, & Vannest, 2007), but here too no clear guidelines exist yet on which measure to use and how to interpret it.

## The Intraocular Trauma Test: Does the Result Hit You Between the Eyes?

Visual analysis is a necessary step when evaluating intervention effects. Even statisticians increasingly emphasize the importance of graphical data analysis. Wilkinson and the APA task force on Statistical Inference (1999) and Wilkinson (2005), for example, advise to first look at the data before computing any statistics. Most software packages, however, are not specialized in constructing single-case graphs. We hereby present a tool to aid researchers in making graphical representations of their single-case data, and transforming graphical displays back to raw data. To create the graphical functionalities, we used the programming environment R. This was chosen because, as an open source implementation of the S-PLUS language, it can be downloaded at no cost from the CRAN website (cran.r-project.org). R is a very powerful and flexible tool, which has very good graphical possibilities to obtain simple as well as complex visual displays of the data, and which is able to deal with

unusual data sets and problems (Kelley, 2007). The R scripts that we created are available on our website (ppw.kuleuven.be/english/research/mesrg).

## Visual Analysis: A Software Tool

Kazdin (2011) defines visual analysis as 'reaching a judgment about the reliability or consistency of intervention effects by visually examining the graphed data' (p. 286). Visual analysis depends on many characteristics of the data, but most importantly on the magnitude of changes across phases (differences in the mean level of performance and level shifts at the moment of phase change) and on the rate of those changes (changes in trend and the latency of changes). Generally speaking, three variables are taken into account: central location (and level), variability, and trend (Franklin, Gorman, Beasley, & Allison, 1997; Morley & Adams, 1991). Therefore we did not only create a basic plotting function to display data from single-case experiments, but we also incorporated some tools to facilitate the use of these three interpretative principles of visual analysis.

### From Data to Graph

Single-case data are usually displayed by plotting the measure of time on the abscissa and the dependent variable on the ordinate, as displayed in Figure 1. For single-case alternation designs (left panel), in which the basic strategy consists of the rapid alternation of two or more conditions within a single case, the data points of each condition are connected, resulting in multiple lines in the same area of the plot. In single-case phase designs (middle panel), comparisons

are made within a time series and the case's performance is evaluated over time across baseline (A) and intervention (B) phases. The graphical display connects the data points in each phase, yielding an interrupted line with a space between the phases. In this space, a vertical line is placed to mark the phase change. Some contrasting findings were published on the effects of these graphing conventions. Knapp (1983) found that judges were more conservative when the phases are not separated, whereas Carter (2009) found no substantive evidence that judgments were affected. If any effect, he found more Type I errors when the graphs are presented without phase change lines and when the data points across phase changes are connected. Because of the lack of a clear opinion and because these conventions are widely used, we decided to incorporate them in the graphical display of phase designs. Users who wish to change these settings can easily do so by slightly changing the R code. Data from multiple baseline AB designs (right panel), in which several AB phase designs are implemented simultaneously to different persons, behaviors, or settings, are displayed by plotting these different AB designs beneath each other, so that the staggered administration of the intervention becomes clear. As can be seen in Figure 1, standard labels are given to the conditions and phases (''A'' and ''B'') and to the abscissa (''Measurement Times'') and ordinate (''Scores''). These can be personalized within the code.

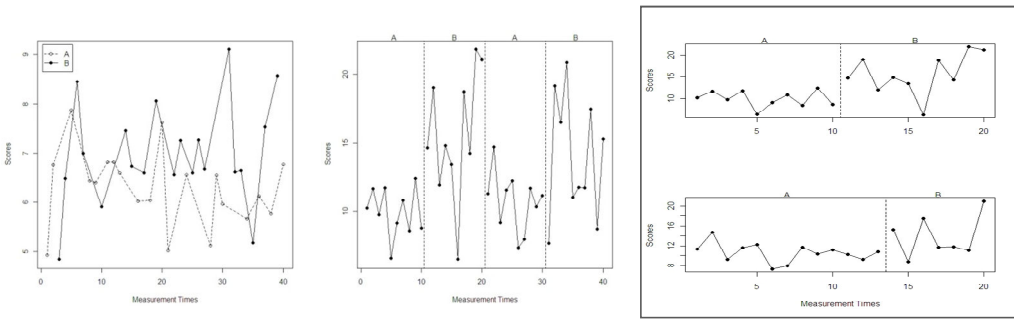



*Figure 1.* Graphical display of three hypothetical examples of single-case research design types. In the left panel an alternating treatments design with two conditions is shown (graph(design="ATD")). The middle panel displays an ABAB phase design (graph(design="ABAB")). And in the right panel an example of a multiple baseline design with two units is given (graph(design="MBD")).

These basic single-case graphs can be obtained by calling the function *graph(''design'')*, where with the design argument the user can specify which design was used in the experiment. Possibilities are a phase design *(''AB,'' ''ABA,'' ''ABAB'')*, an alternating treatments design *(''ATD''*), and a multiple baseline AB design *(''MBD''*). A more detailed explanation of these different design types can be found in Barlow et al. (2009). All the functions explained below will work for the above-mentioned design types. For illustration purposes we will focus on AB phase designs in what follows. The use of the general interpretative principles of central location, variability, and trend in the next paragraphs is largely based on Morley and Adams (1991) and Franklin, Gorman, Beasley, and Allison (1997).

## Central Location

Changes in level across phases appear to be the characteristic that judges use most often to come to a decision and it is the characteristic that is associated with the highest degree of interrater

agreement, rater certainty, and rater confidence (Bailey, 1984; Fisch, 1998; Furlong & Wampold, 1982; Gibson & Ottenbacher, 1988; Normand & Bailey, 2006; Parsonson & Baer, 1992; Wampold & Furlong, 1981). Central location can be incorporated in the single-case plot by superimposing a horizontal reference line on the raw time series. This will make treatment effects (demonstrated by differences in level) more visible, while also providing a basis for analyzing variability and trend. With the function *graph.CL(design,CL,tr)* the user can choose which measure of central tendency has to be plotted as a line parallel to the abscissa. An example of an AB phase design where the mean is plotted as a reference line for each phase is given in Figure 2. Note that there is no legend indicating which measure of central tendency is displayed. Users can mention this in the figure's caption. This recommendation also applies to most of the subsequent graphs.

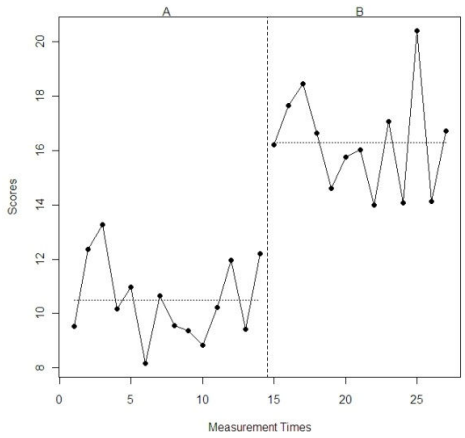



*Figure 2.* Hypothetical example of a possible display of central location for an AB phase design. The mean is plotted in each phase as a horizontal reference line superimposed on the raw time series data (*graph.CL(design="AB",CL="mean")).*

The mean (*CL=''mean''*) is not resistant to the influence of outliers, so sometimes it could be better to consider the median (*CL=''median''*). However, in small samples the median may also not be very representative, because it only takes into account the one or two central data points of the whole data set. A possible solution here is to use the broadened median (*CL=''bmed''*), which is calculated based on the three, four, five, or six middle values of the data set, depending on the total number of data points (Morley & Adams, 1991). This way this measure is sensitive to a larger proportion of the data than the median, while also being robust to the influence of outliers.

The median, however, has a larger standard error than the mean when the population is normally distributed (Wilcox, 2005). It is also strongly affected by the observations in the center of the distribution, which is not the case for the mean. Another way of dealing with the lack of robustness of the mean as a measure of central tendency is by discarding the observations in the tails of the distribution (the extreme values) and calculating the so-called ''trimmed mean'' on the remaining observations (*CL=''trimmean''*). The percentage of observations that has to be removed from each end of the distribution before computing the mean can be set with the *''tr''* argument, and can be any value from 0 (regular arithmetic mean) to 0.5. Usually 20% of the observations is trimmed (so *tr=0.2*).

While the trimmed mean removes a fixed proportion of the observations, whether they are outliers or not, Huber's M-estimators of location (Huber & Ronchetti, 2009) first evaluate each observation

to determine if it actually is an outlier compared to the rest of the data and then give less weight to those outlying values. For this evaluation a constant K needs to be specified that can have any value between 0 and ∞, and which actually comes down to balancing between robustness and efficiency: when K = 0, the M-estimator equals the median (maximally robust), while when K → ∞ it equals the arithmetic mean (maximally efficient) (Jeng, 2010). Therefore, M-estimators of location could be considered as a generic measure of central tendency of which the mean and the median are special cases. Usually a percentile of the standard normal distribution is chosen as the constant K. Wilcox (2005) suggests using K = 1.28, which corresponds to the 90th percentile of the standard normal distribution and covers 80% of the underlying distribution. When choosing a larger value for K, the coverage gets larger until the whole distribution is covered when K → ∞ (as with the mean). In other words, the larger K gets, the wider the range of plausible observations becomes, and the lesser observations will be classified as outliers. To display this generic measure of central tendency, the *CL* argument should be put to *''mest,''* and the desired value for the constant should be specified with the *tr* argument (e.g., *tr=1.28*). The function *mest(x,bend=1.28)* from Wilcox (2005) is used for the calculation of this measure. In his book, Wilcox (2005) provides R functions for several other robust estimators of location. These estimators could also be included by making slight adjustments to the R code.

## Variability

Relative variation within and between phases is often overlooked or ignored by visual analysts. This seems to be influenced by the area of expertise of the judges: Whereas statistically trained judges take into account the type of the effect and the amount of variability across phases, experts in visual analysis tend to look only at the magnitude of the effect (Furlong & Wampold, 1982; Wampold & Furlong, 1981). However, variability in the data could have an influence on the decisions made by the judges, with a higher degree of variability being related to higher Type I error rates (Matyas & Greenwood, 1990). By graphically displaying some estimates of variability, we want to separate variability from other aspects in the display, like level and trend with which it is often confounded. This way, we aim at making it easier to attend to variability when visually analyzing the data.

The function *graph.VAR(design,VAR,dataset,CL,tr)* was developed for this purpose. It includes two commonly used methods to demonstrate variability: range bars and range lines. Range bar graphs (*VAR=''RB''*) consist of a vertical line for each phase, created by connecting three points: An estimate of central tendency, the minimum, and the maximum, while the raw data points are usually not plotted. Depending on the desired measure of central tendency, the *CL* argument should be put at *''mean,'' ''median,'' ''bmed,'' ''trimmean,'' or ''mest''* (with the additional *tr* argument for the trimmed mean and the M-estimator). In the upper left panel of

Figure 3, a range bar graph with the mean as a measure of central tendency is displayed.

Range lines (*VAR=''RL''*) are superimposed on the raw data and therefore do include information about the timing of the data points. This way information about variability is displayed, while at the same time highlighting possible trends. A pair of lines is drawn parallel to the X-axis, and passing through the lowest and the highest values for each phase. An example is given in the upper middle panel of Figure 3.

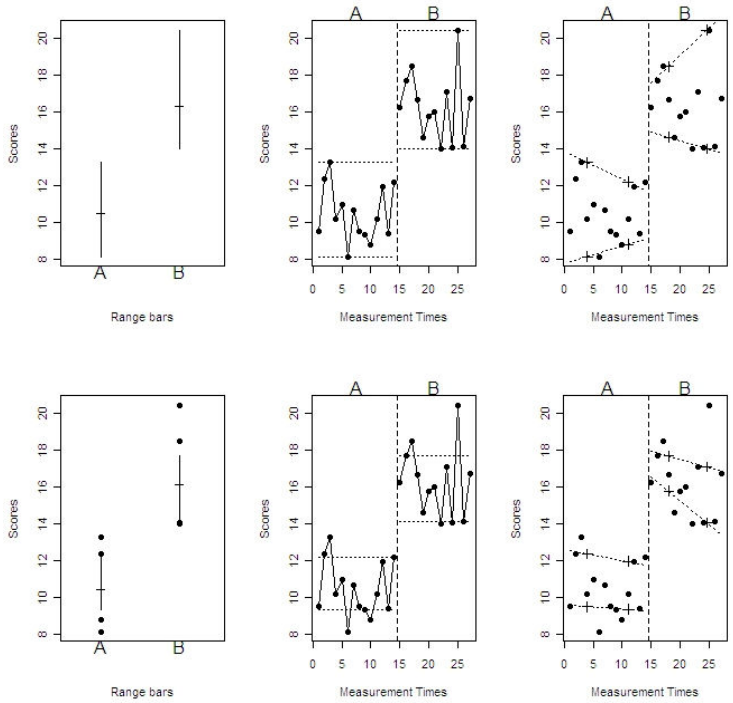



*Figure 3.* Hypothetical example of possible displays of variability for an AB phase design. The upper left panel shows range bars, with the mean as a measure of central tendency (*graph.VAR(design="AB",VAR="RB",CL="mean")),* and the upper middle panel displays range lines (*graph.VAR(design="AB",VAR="RL")).* In the upper right panel, a trended range is shown (*graph.VAR(design="AB",VAR="TR")).* In the lower panels, the trimmed versions of each variability display are plotted: trimmed range bars on the left, trimmed range lines in the middle, and trimmed trended ranges on the right. This is done by adding the *dataset="trimmed"* argument to the R command, e.g., for range lines: *graph.VAR(design="AB",VAR="RL", dataset="trimmed").*

One problem with these two methods of displaying variation is that the range is severely influenced by the presence of outliers. This may be overcome by using a trimmed range (*dataset=''trimmed''*), in which only a sample of the data is used. The original data set is trimmed with 10–20% by taking out the most extreme values (i.e., the lowest and highest data points). This is demonstrated for range bars (lower left panel) and range lines (lower middle panel) in Figure 3. Notice that for range bars, the omitted values are displayed as dots above and below the vertical line. Two other problems with range bars and range lines are that variability may be confounded with trend and that changes in variability within phases are not displayed. This could be solved by plotting a trended range (*VAR=''TR''*), as demonstrated in the upper right panel of Figure 3. For each phase two lines are drawn, displaying possible changes in variability. First the phase is divided into two halves along the X-axis. Then the middle time point is determined for each phase half, and the minimum and maximum values of the dependent variable within that half are plotted at that time point. Finally the minimum values of each half are connected, as well as the maximum values, which results in two lines. Trended ranges can also be applied to the data after the lowest and the highest values in each half of the data set have been trimmed, by setting the *dataset* argument to ''*trimmed*'' (lower right panel).

## Trend

Whereas level changes between phases are the characteristic associated with the highest degree of interrater agreement, changes

in trend are the characteristic that is most often associated with inconsistent interpretations (Bailey, 1984; Gibson & Ottenbacher, 1988; Parsonson & Baer, 1992). Judges seem to focus primarily on shifts in level, being unable to differentiate these from trends in the data, with similar difficulties in detecting trends for experienced and non-experienced judges (Fisch, 1998; Furlong & Wampold, 1982; Wampold & Furlong, 1981). After training judges in using a standard trend estimation procedure, interrater agreement as well as confidence in the judgments increased. However, training does not necessarily lead to better performance (Knapp, 1983; Richards et al., 1997). It can result in more conservative judgments or make analysts rely only on those criteria in which they have been trained. Besides training, the use of a judgmental aid may be important to visual inferences (Skiba et al., 1989).

Morley and Adams (1991) define trend as ‘‘a systematic shift in the value of the central location of the data set over time’’ (p. 100). A quick impression of the presence of such a trend can be obtained by drawing a vertical line graph, in which the deviations from each data point to a measure of central tendency are plotted against time. An example of such a graph is given in the upper left corner of Figure 4. If a positive trend would be present in the data, the vertical lines on the left side would be hanging below the central tendency line, and the vertical lines on the right side would be standing on it. With the function *graph.TREND(design,TREND,CL,tr)* we included some possibilities to display a potential trend in the data. Depending on the desired central tendency measure, a vertical line graph can be drawn

by setting the *TREND* argument to *‘‘VLP’’* and the *CL* argument to ‘‘*mean,’’ ‘‘median,’’ ‘‘bmed,’’ ‘‘trimmean,’’ or ‘‘mest’’* (and if necessary complemented with the *tr* argument).

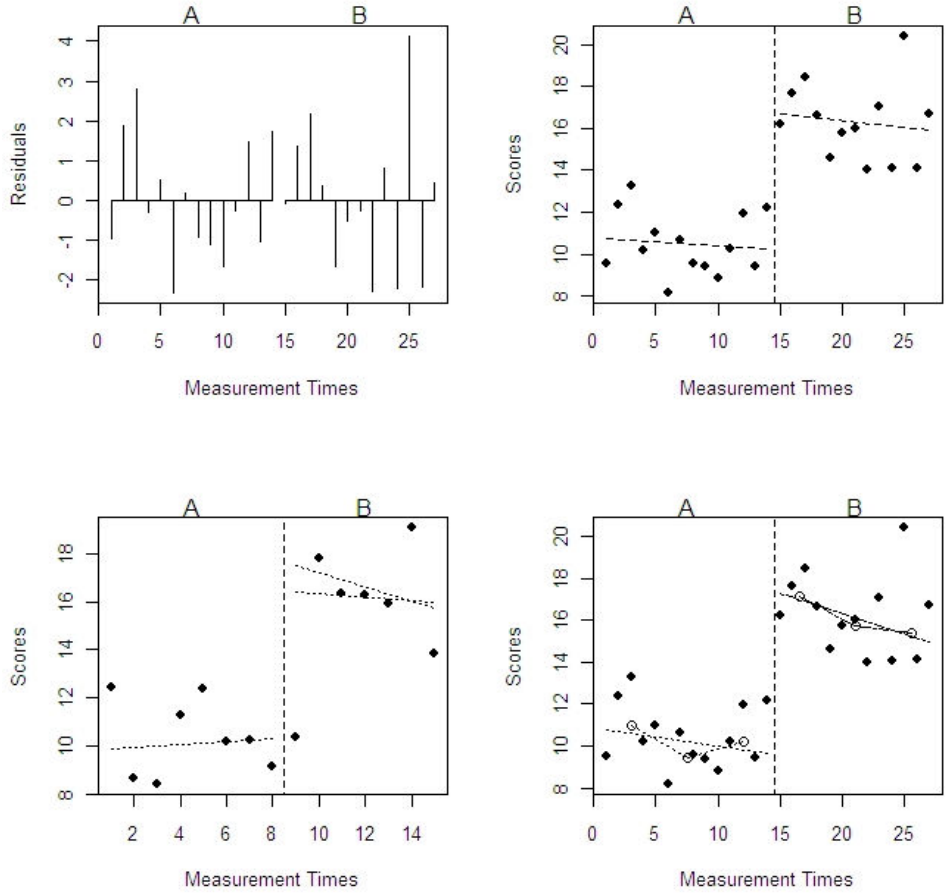



*Figure 4.* Hypothetical examples of linear trend displays for AB phase designs. In the upper left panel, a vertical line graph with deviations from the mean is displayed (*graph.TREND(design="AB",TREND="VLP",CL="mean")).* The upper right panel shows trend lines by least-squares regression (*graph.TREND(design="AB",TREND="LSR")).* Bottom left the split-middle method is demonstrated (*graph.TREND(design="AB", TREND="SM"))* and in the lower right corner we see the result of resistant trend line fitting (*graph.TREND(design="AB",TREND="RTL")).*

A trend in central location is usually displayed in single-case graphs by superimposing a linear function on the raw data, which shows if there is an increase or a decrease in the behavior over time. There is

evidence that the use of such regression or trend lines can increase interrater agreement, reliability, and decision accuracy (Bailey, 1984; Fisher, Kelley, & Lomas, 2003; Hojem & Ottenbacher, 1988; Parsonson & Baer, 1992; Rojahn & Schulze, 1985; Skiba et al., 1989). However, results are not all positive, because in some cases trend lines created dependencies, helped maintaining inconsistent judgments, and led to an overemphasis of trend and the neglect of other factors like level and variability (DeProspero & Cohen, 1979; Fisch, 1998; Harbst et al., 1991; Hojem & Ottenbacher, 1988; Skiba et al., 1989). Moreover, some researchers found no significant effect of using trend lines on the accuracy of decisions (Normand & Bailey, 2006). Fisch (1998) argues that these contradictory results could be partly explained by personal characteristics of the judges: experience seems to be more important than the use of visual guidelines. Trend lines are most easily drawn by 'eyeballing' a line that seems to bisect the distribution. Research has however shown that judges are not that good at deciding how and where exactly this line should be placed (Mosteller, Siegel, Trapido, & Youtz, 1981). Therefore we included some functions for fitting a robust linear trend through several methods.

The first possibility to draw a line that approximates the best linear fit of the data consists of standard linear regression. When setting the *TREND* argument to *''LSR'' (*least-squares regression), a regression line (Y = a + bX ) that minimizes the squared vertical distances between the line and the data points is calculated from the slope and the intercept. An example is shown in the upper right corner of

Figure 4. Just as with eyeballing, this method is however very much influenced by the presence of outliers. Several alternative, more robust, estimators have been proposed to replace ordinary least-squares regression. Wilcox (2005) gives a good overview, together with R functions to calculate those measures. The list of possibilities is very long (e.g., least median of squares, least trimmed squares, regression M-estimators with many variations, S-estimators), and we refer the interested reader to Wilcox (2005).

Another possibility to display trend is the split-middle method (*TREND=''SM''*), which is quite straightforward when phases have an even number of observations. The first step involves dividing the phase into two halves along the Xaxis. When there are, for example, ten observations, the first five observations are allocated to the left phase half, and the last five observations are allocated to the right phase half. In the second step two values are calculated for each phase half: The median of the dependent variable and the middle time value. This middle time value is the observation in the middle of the time series for an uneven number of observations (e.g., the 3 in a series of five observations), whereas for an even number of observations this is the average of the two middle time values (e.g., 2.5 in a series of four observations), which is an imaginary time point. The final step of the split-middle method connects the crossing of the median (Y-axis) and the middle time value (X-axis) of both phase halves. The same three steps are followed when phases have an unequal number of observations, with as a difference that in the first step of dividing the phase into two halves, one of the halves will have

an observation more than the other half. Therefore the division is conducted twice: once with the extra observation allocated to the left half and once with the extra observation allocated to the right half. Step two and step three are therefore also repeated, once for each way of allocating the data points, which results in two lines instead of one. The split-middle method is demonstrated in the lower left panel of Figure 4: in phase A the situation of an even number of observations is shown, while in phase B the two split-middle lines for a phase with an uneven number of observations are displayed.

Whereas the split-middle method is sufficient to display a linear trend for shorter time series (4–12 observations), the resistant trend line fitting method (*TREND=''RTL''*) (Tukey, 1977) is more often used with larger time series. Both methods are comparable in finding the medians of the dependent variable and the middle time values, but with resistant trend line fitting the data are divided in three sections along the X-axis instead of two. The slope of the resistant trend line is determined by the change in the medians between the two outer data sections, and the intercept is computed using all three data sections to make the line pass as closely as possible through the middle of the data. By plotting the two half-lines, connecting the coordinates (intersection of the mid-time point and the median of the dependent variable) of each of the three sections, it is possible to check whether a linear fit is adequate. In our example, in the lower right panel of Figure 4, the fit for the B phase is almost linear, because the line connecting the coordinates of the two outer

sections runs nearly through the midpoint of the middle section. This is not the case for the data in the A phase.

To further investigate the presence of a nonlinear trend in the data, we included the possibility of displaying running medians. With running medians, the time series is smoothed by dividing it into successive segments of a given size and calculating the median for each segment (Tukey, 1977). Three sizes of segments are easy to use with time-series data: running medians of three (*TREND=''RM3''*) can be considered when there are 6–10 observations (left panel of Figure 5), and for time series with more than 10 observations running medians of five (*TREND=''RM5''*) could be used (middle panel). These are calculated by finding the median of each batch of, respectively, 3 or 5 successive data points and plotting this value at the middle time point for that batch. This means that the first median value is plotted at the second time point (middle of 1-2-3) for RM3 and at the third time point (middle of 1-2-3-4-5) for RM5, reducing the time series by two (in the case of RM3) or four (in the case of RM5) data points. These two smoothers, however, often produce curves with high and low points. Another way of estimating the median of five successive values is by calculating running medians of batch size four, and then averaging each successive pair (*TREND=''RM42''*). As can be seen in the right panel of Figure 5 this results in a more smoothed curve. Of course many other smoothers exist, such as kernel and lowess smoothing. Wilcox (2005) provides R functions for making plots of these smoothed functions, as well as for more robust estimators.

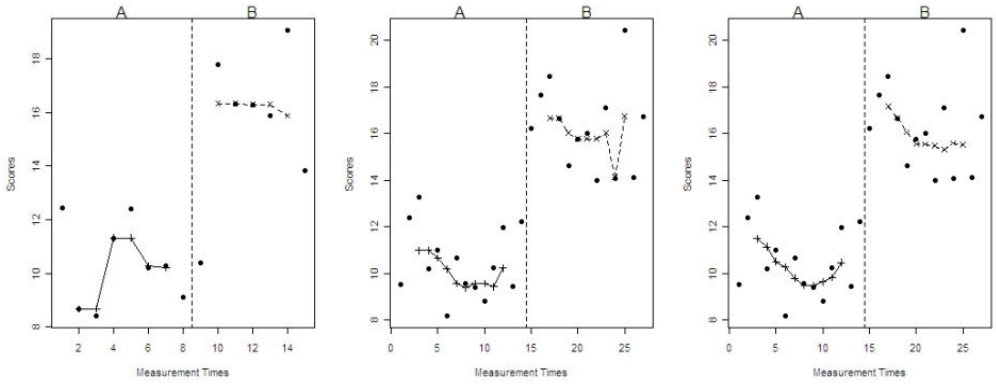



*Figure 5.* Examples of running medians superimposed on the hypothetical raw data. On the left, running medians of three are shown (*graph.TREND(design="AB", TREND="RM3*")). In the centre we displayed running medians of five (*graph.TREND (design="AB",TREND="RM5")),* and on the right side running medians of four averaged by pairs are plotted (*graph.TREND(design="AB",TREND="RM42")).*

## From Graph to Data

To construct these graphical representations, one usually needs the original data values. This is also the case when statistically reanalyzing published data, for example to integrate the results from a study in a meta-analysis. Unfortunately, raw data are often not available from published studies. Therefore, we included a function to extract the values from existing graphs (adaptation of Timothée, 2010): *graph.extract(MT,refX,refY,save).* Suppose that we want to do this for the data displayed in the left panel of Figure 5. We need to supply the number of observations (*MT=15*), two reference values for the X scale (e.g., *refX=c(2,14*)), and two reference values for the Y scale (e.g., *refY=c(10,16)).* When giving the command, a window pops up in which the user can indicate where the jpeg file with the graph can be found. Then the reference values should be located on the graph by clicking with the left mouse button, first the X values and then the Y values, and the same should be done sequentially for each

data point. As can be seen in the left panel of Figure 6, the marked reference points are indicated with a blue cross and the marked data points with a red circle. When finished, the calibrated true data points are displayed as output in the R console and plotted in a new graph. This is shown in the right panel of Figure 6. If one wants to save these data values to a file, the *save* argument should be put to *''yes,''* and another window will pop up where it is asked in which file the data should be saved. Here one can create a new txt file in the folder of choice. Notice that for this function the R package ReadImages should be installed and the graphs of which one wants to extract data should be in jpeg format.

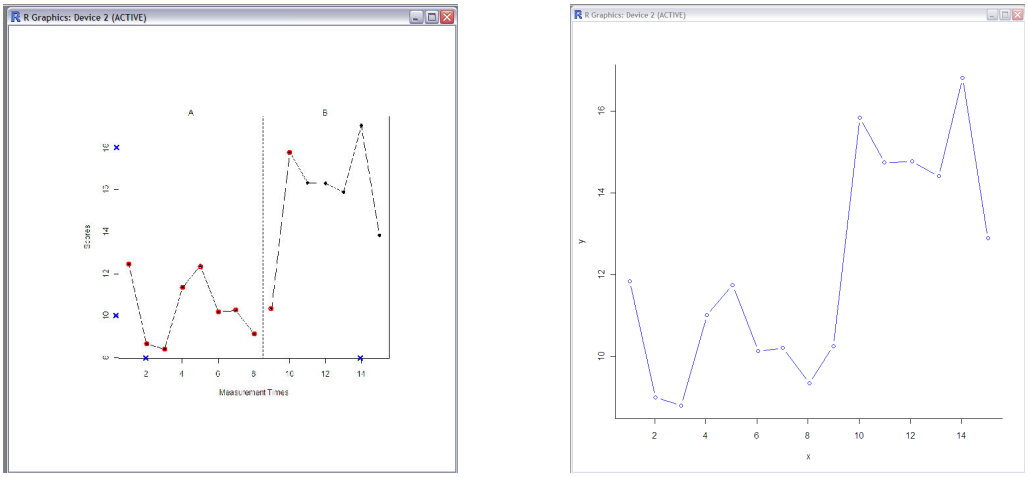



*Figure 6.* Print screen of the '*graph.extract*' function. On the left is shown how the reference values and data points are located on the graph. In the right panel the output is demonstrated: the calibrated true data points are returned in the R console and a new graph is created as a quick visual check.

An overview of all graphical functions is given in Table 1. Table 2 summarizes the possible arguments.

*Table 1.* Overview of Graphical Functions

| Function name | Description |
|---|---|
| graph(design) | Makes a graphical representation of the single-case data |
| graph.CL(design,CL,tr) | Plots a measure of central tendency as a horizontal reference line superimposed on the raw time series data |
| graph.VAR(design,VAR,dataset,CL,tr) | Displays information about variability in the data |
| graph.TREND(design,TREND,CL,tr) | Visualizes systematic shifts in central location of a data set over time |
| graph.extract(MT,refX,refY,save) | Extracts data values from published graphs |

*Table 2.* Overview of Graphical Arguments

| Argument | Values |
|---|---|
| design | Indicates what type of single-case design has been used: a phase design ("AB", "ABA", "ABAB"), an alternating treatments design ("ATD"), or a multiple-baseline AB design (“MBD”) |
| CL | Indicates what measure of central tendency should be plotted: "mean", "median", "bmed" (broadened median), “trimmean” (trimmed mean), or “mest” (M-estimator) |
| tr | For a trimmed mean: indicates which proportion (0-0.5) of the observations has to be removed from each side before calculating the mean. A common choice when calculating a trimmed mean is trim=0.2.<br><br>For a M-estimator of location: indicates the constant against which the observations are evaluated to determine if they are extreme. This constant is usually chosen at 1.28 (the .9 quantile of the standard normal distribution) (Wilcox, 2005). |
| VAR | Indicates what kind of variability information should be plotted: range lines ("RL"), range bars ("RB), or a trended range ("TR") |
| dataset | Indicates whether the whole dataset should be used ("regular" = default) or the 10-20% extreme values from each phase should be removed ("trimmed") |
| TREND | Indicates what trend visualization should be plotted: vertical line plot ("VLP"); trend lines by means of least squares regression ("LSR"), split-middle method ("SM"), resistant trend line fitting ("RTL"), or running medians ("RM3","RM5", "RM42") |
| MT | Number of observations (measurement times) |
| refX | Reference values for the X scale: c(x1,x2) |
| refY | Reference values for the Y scale: c(y1,y2) |
| save | Indicates whether the data values should be saved to a file (save="yes") or just displayed as output in the R console (save="no" = default) |

# Discussion

Although there have been serious concerns regarding the consistency, sensitivity, and specificity of visual analysis of single-case data, this kind of analysis remains a necessary step in any thoughtful data-analytic procedure. In some instances, it could even be considered a sufficient step. When a visualization is that strong that the conclusions ‘‘hit you between the eyes’’ (what is sometimes called the ‘‘Intraocular Trauma Test’’), statistical inference is of little relevance and *p*-values do not really have a surplus value. When the effect of a behavior modification intervention is very clear, it is unnecessary to use complicated statistical modeling techniques such as time-series analysis to state the obvious. Graphs can also be sufficient when the purpose is to give feedback to individual patients. Graphical displays of single-case data are very well-suited for online monitoring of the patient’s progress, by altering them every time new data become available, and for communicating results in an easy and understandable way. In other instances, of course, when the effect is not that plain to see or when the results have to be combined in a meta-analysis, statistical data analysis might be a useful supplementary technique. Even strong proponents of visual analysis acknowledge that statistical results can be valuable in the absence of a stable baseline, when results must be shared unambiguously with other professionals, and when testing new treatments of which the effects cannot be predicted (Parker et al., 2005).

This possible need for complementary statistical analyses can be easily accomplished by working in R. We did not only choose this programming language because of its graphical possibilities, but also because it is a statistical environment, which facilitates the combined use of visual and statistical analyses. Moreover, as indicated before, the flexibility of R enables users to change parts of the code according to their demands (e.g., labels of the axes). This requires some experience with the program, but this is easily acquired. We acknowledge that there might be a threshold to start working with R, because of the lower user-friendliness of its command line interface. However, to use the functions described in this paper in their standard versions, one does not need to know much about R. After downloading and installing R (Hornik, 2010, gives a detailed explanation), the functions can be read into R by choosing ‘‘File’’ and selecting ‘‘Source R Code’’ from the top menu. For easy access and use, we suggest saving the files containing the R functions on the local disk. The text files (.txt) containing the data are most easily made in a text editor (e.g., EditPad or NotePad) or in Excel (saved as ‘‘text (tab delimited)’’). For phase designs and alternating treatments designs the file should consist of two columns: one containing the condition labels (‘‘A’’ and ‘‘B’’) and the second with the observations. It is important not to label the rows or columns. For multiple baseline AB designs the data file should consist of these two columns for each unit (so twice as many columns as there are units). In the future, more and more forms of graphical user interfaces for R will become available, what will enhance the user-friendliness (but maybe also decrease the flexibility).

# An R package for single-case randomization tests


## Abstract

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.

# Introduction

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 treatments. 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.

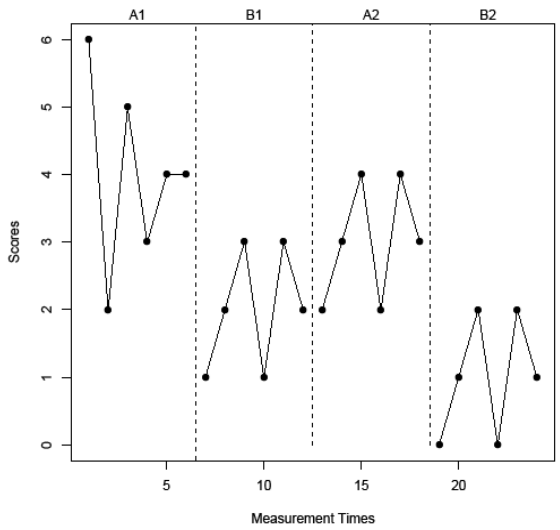



*Figure 1.* Hypothetical example of an ABAB design (Onghena, 1992).

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.

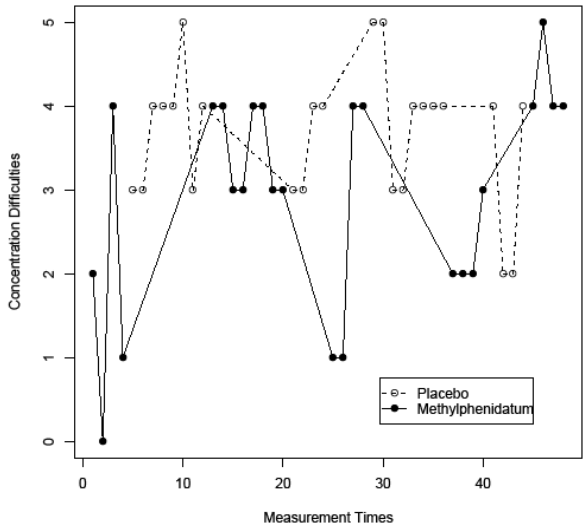



*Figure 2.* Alternation design with a treatment condition and a placebo condition to test the impact of methylphenidatum on the concentration difficulties of an adolescent (Baplu, 2005).

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 AAABBB AAAAAB

ABBBBB AAAABB BBBBBB

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 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 |
| BAABB | 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 | ABBABA |
| ABAABB | BABBAA | ABABAB | BABABA | BAABAB |
| ABABBA | BABAAB | ABBAAB | BAABBA | |

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 obtained 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, 1980b; 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.

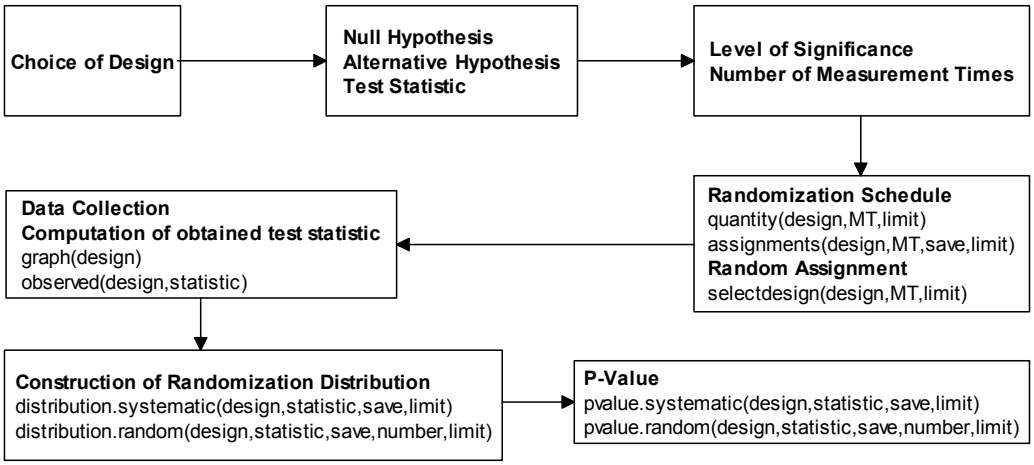



*Figure 3.* Step by step procedure for conducting a single-case randomization test, with R functions.

## 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 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:

$$T = \left(\overline{A_1} + \overline{A_2}\right) - \left(\overline{B_1} + \overline{B_2}\right)$$

## Step 3: Determination of the level of significance and the number of measurement times

Before gathering the data, the level of significance, $\alpha$, 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, $\alpha$, 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 $\alpha$ 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 possible 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 selectde*sign(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.

Table 1. *Hypothetical Data from Onghena (1992), Collected in an ABAB Design with 24 Measurement Times (MT)*

| Condition | A1 | A1 | A1 | A1 | A1 | A1 | B1 | B1 | B1 | B1 | B1 | B1 | A2 | A2 | A2 | A2 | A2 | A2 | B2 | B2 | B2 | B2 | B2 | B2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 |

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: $T = \left(\overline{A_1} + \overline{A_2}\right) - \left(\overline{B_1} + \overline{B_2}\right) = \left(4 + 3\right) - \left(2 + 1\right) = 4$

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 consider 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.

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 |

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 $\alpha$, 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 1013 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="AB",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.

# Multiple Baseline Designs in the Journal of Applied Behavior Analysis: A Systematic Methodological Review at their 40th Anniversary.


## Abstract

Since their first appearance in the applied literature 40 years ago, multiple baseline designs gained fame and still are very popular. In this review, we look at customs and shortcomings in the application of those designs in the *Journal of Applied Behavior Analysis,* over the past 15 years. More specific, we are interested in some design features, like the type of multiple baseline designs used, the number of units and the number of measurement times per unit, and the way the placement of the intervention points is decided. Our interest lies also in the techniques that are used to analyze data from multiple baseline designs in general, and more in particular in the application of randomization tests.

# Introduction

Multiple baseline designs can be considered as one of the most important contributions of the applied behavior sciences to general research methodology. The rationale and the term as such were first presented in the applied literature by Baer, Wolf, and Risley (1968) in their innovative article in the first issue of the *Journal of Applied Behavior Analysis* (*JABA*). Since then, multiple baseline designs gained fame and they are still very popular in recent applied behavioral research (Ferron & Scott, 2005; Long & Hollin, 1995).

With multiple baseline designs reaching the mature age of 40, we considered it appropriate to take a closer look at the recent practices and prospects of this versatile research methodology. In particular, we wanted to summarize the customs and call attention to some specific shortcomings and opportunities for optimization in application and reporting. We were especially interested in the techniques used to analyze data collected through multiple baseline designs in general and the use of randomization tests in particular.

## The Multiple Baseline Design Rationale

The popularity of multiple baseline designs is probably due to their flexibility and simplicity (Barlow & Hersen, 1984; Ferron & Scott, 2005; Kazdin, 1982). Multiple baseline designs are well suited to flexibly investigate a wide variety of problems and populations in a large range of (applied) research domains. Furthermore, they can be considered as simple extensions of the most basic type of single-case research designs, the AB phase design, in which all baseline

measurements (the A phase) precede all treatment measurements (the B phase). The researcher starts by taking repeated measurements of the behavior during the baseline phase, then introduces an intervention and continues to record the target behavior during the treatment phase (Barlow & Hersen, 1984; Kazdin, 1982; Onghena, 2005). Multiple baseline designs extend this basic AB design by subjecting several persons, behaviors, or settings under study simultaneously to a separate AB design (Barlow & Hersen, 1984; Kazdin, 1982; Onghena, 2005). For convenience, these separate persons, behaviors, or settings will be called 'units' throughout the remainder of this article.

The most distinguishing feature of the multiple baseline design is that the intervention is introduced in a staggered way to these different units, by extending each baseline phase a little further than the previous one (Barlow & Hersen, 1984; Kazdin, 1982, 2003; Kinugasa, Cerin, & Hooper, 2004). This way, the different units do not only serve as replications, but also as controls to determine what changes one can expect without the introduction of the intervention to that specific unit. When each unit changes if and only if the intervention is introduced, the effects can be attributed with more confidence to the intervention instead of to extraneous effects (Baer et al., 1968; Barlow & Hersen, 1984; Kazdin, 1982; Kinugasa et al., 2004).

## Design Considerations

For this review, we were interested in some design features based on which multiple baseline designs can be classified. A first distinguishing feature is the *type* of multiple baseline design: across

subjects, across behaviors, or across settings. In a multiple baseline across subjects design, an intervention is applied sequentially to several participants (Barlow & Hersen, 1984; Kazdin, 1982; Kinugasa et al., 2004). This variation of the design is most appropriate when the particular behavior one wants to change is constant among different persons (Kazdin, 1982). In a multiple baseline across behaviors or across settings design, the intervention is applied sequentially to respectively different behaviors within the same subject or to several independent situations, settings or time periods in a given subject (Barlow & Hersen, 1984). The multiple baseline across behaviors design is well-suited when one wants to alter several behaviors in one subject, while the variation across settings can be used when a particular behavior needs to be altered in several situations (Kazdin, 1982). Also more complex designs can be thought of by extending this basic structure to form mixed designs, for instance, combining the across subjects and across behaviors variants, by observing two behaviors of each of four participants (Ferron & Scott, 2005).

Multiple baseline designs can vary not only with respect to the types of units that are involved, but also in the procedure used to determine the *intervention points* throughout the different units. Traditionally, a response-guided assignment procedure is used, in which the intervention points are chosen based on the emerging data patterns, for instance when a researcher waits with intervening until baseline data are stable (Ferron & Jones, 2006; Kazdin, 1982). An alternative procedure consists in the use of a systematic assignment

schedule, where the points to intervene are decided a priori, in order to obtain, for example, an even staggering across time (Ferron & Scott, 2005). And finally, a random assignment approach can be used, in which the intervention points are determined at random (Edgington, 1975, 1996; Onghena, 1992).

A third and final way to classify multiple baseline designs is by the *number* of units (i.e., behaviors, persons, or settings) and the number of observations per unit. Other things being equal, the inclusion of more than two units is preferable, because this provides a greater number of replications of the effect and thus allows for more confidence that the observed effect was indeed the result of the intervention (Kazdin, 1982). The same is true for the number of observations: more measurement occasions yield stronger statements about the effect of the treatment (Ferron & Scott, 2005). Barlow and Hersen (1984) recommend a minimum of three to four units, and Kazdin and Kopel (1975) suggest the use of four or more units.

## Analyzing Multiple Baseline Data

For the review, we also wanted to look at which analyses are most frequently used with multiple baseline data. Several techniques have been suggested, of which the oldest one is visual inspection. Here, researchers let the data speak for themselves and look at a graph to determine whether the intervention had any effects (Long & Hollin, 1995). Although this way of data-analysis without doubts has several advantages (for a scholarly review see e.g., Parsonson & Baer, 1992), these do in many instances not outweigh the difficulties, like the lack

of concrete decision rules, the high risk of Type I errors, and the requirement of a specific pattern in the data (e.g., stable baselines) (Kazdin, 1982; Matyas & Greenwood, 1990). Also, there is often a lack of agreement among judges with respect to the conclusions drawn from visual analysis, so the reliability and accuracy of this method have been questioned (see e.g., DeProspero & Cohen, 1979). Therefore, Morley and Adams (1991) recommended to subject data to both visual and statistical analysis, whenever this is possible.

But which statistical test should be used when analyzing multiple baseline data? Regretfully, the well-known parametric tests (e.g., *t* tests and ANOVAs) are often inappropriate for the analysis of multiple baseline data, because they require assumptions about the data that are infrequently met in practice (e.g., normality, homogeneity of variances, and independence of errors). Most problematic is the assumption of serial independence because multiple baseline data tend to have autocorrelated residuals. It is well-known that the presence of autocorrelation can seriously bias the results from *t* and *F* tests, for instance, by inflating Type I errors (see e.g., Gorman & Allison, 1997; Hooton, 1991; Kazdin, 1982, 1984; Kinugasa et al., 2004; Ludbrook, 1994; Recchia & Rocchetti, 1982; Todman & Dugard, 2001).

A promising suggested alternative for the analysis of multiple baseline data are nonparametric randomization tests. Their most important advantages include that they 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); that they are free from the assumption of random sampling (Edgington, 1973; Edgington & Bland, 1993); and that their application is straightforward and easy to communicate to applied researchers and practitioners (Bulté & Onghena, 2008; Edgington & Onghena, 2007; Onghena & Edgington, 2005). Several randomization test strategies have been suggested for use with multiple baseline data (see e.g., Koehler & Levin, 1998; Marascuilo & Busk, 1988; Revusky, 1967; Wampold & Worsham, 1986).

## Objectives of the Review

Long and Hollin (1995) already examined the articles published in *JABA* and in *Behavioral Psychotherapy* (now *Behavioral and Cognitive Psychotherapy*) over the period from 1987 to 1991 to uncover the trends in single case research. They reported that 59% of all articles in *JABA* in this time frame were devoted to single case studies, among which 52% were multiple baseline designs (plus 6.5% 'combination designs', that were mostly multiple baseline designs with alternating treatments or reversal). Long and Hollin (1995) also mention that in 92% of the cases, these studies relied solely on visual analysis of the data (in contrast to the single-case studies in *Behavioral Psychiatry*, of which 19% used statistical analyses to complement the visual inspection).

To find out if anything changed over the past 15 years, we took a closer look at the articles published in *JABA* from 1993 to 2007. We wanted to see if multiple baseline designs still are as popular in this journal as 20 years ago and if the methods of data analysis have

changed over the decennia. The purpose was to get a clear picture of the use and method of reporting of multiple baseline designs and to get an idea of the populations and topics studied with multiple baseline designs.

## Method

We used the search engine ISI Web of Knowledge (apps.isiknowledge.com/) to find all articles, published in *JABA* between 1993 and 2007, that used a multiple baseline design. As key words we entered 'multiple baseline design' (topic), 'Journal of Applied Behavior Analysis' (publication name) and '1993-2007' (year published). Only studies that actually employed a multiple baseline design for the collection of empirical data were retained.

## Results

With the aforementioned search key words, we received 95 hits. Of those we had to eliminate two because they were not really applications of multiple baseline designs. One concerned instructions on the creation of graphs for single subject designs (Carr & Burkholder, 1998), and the other employed three alternating treatments designs (one for each participant) (Paramore & Higbee, 2005). An alphabetical list of the remaining 93 articles is provided in Appendix A. These articles account for 9.5 % of all articles published during this period in *JABA.* Figure 1 shows the fluctuations in number of articles over the years.

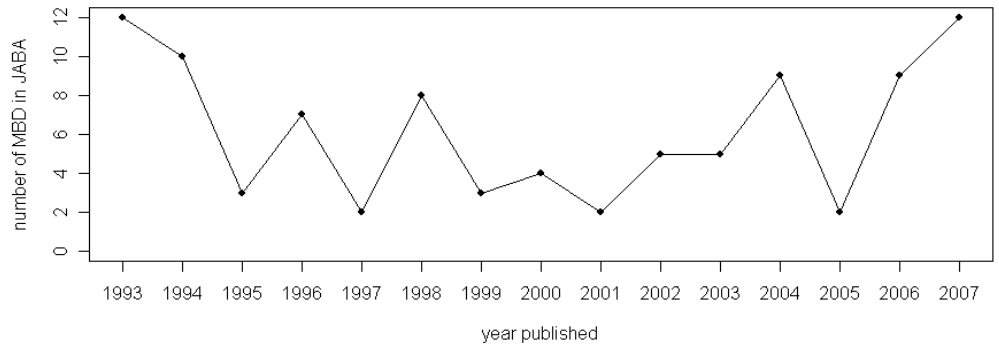



*Figure 1*. The number of articles published in the Journal of Applied Behavioral Analysis over the past 15 years in which a multiple baseline design was used.

The articles included a variety of populations and problems studied. In most studies, the population of interest comprised of children, ranging from 'normal children and infants' (20.43%) to children with developmental disorders (44.09%) (e.g., PDD, ADHD, mental retardation) and medical problems (6.45%) (e.g., cystic fibrosis, agenesis of corpus callosum). The remainder of the articles concerned 'normal adults' (18.28%) and adult inpatients (10.75%) (institutionalized adults with dementia or mental retardation). A large range of problems were investigated, going from social problems and problem behavior, over academic and sports performance, to staff/teacher/parent training programs and safety behavior. In Appendix B, a detailed overview is provided.

When we took a closer look at the design features, it was apparent that the multiple baseline design across subjects was the most popular variant (65.6%). Sixteen studies made use of a multiple baseline across behaviors design (17.2%) and in 14 articles (15.1%) a

multiple baseline across settings design was used. We also found two implementations of a mixed design (2.1%), in which a combination of a multiple baseline design across subjects and across behaviors was applied. Regarding the placement of the intervention points, response-guided and systematic assignment procedures were roughly equally present, with a minor advantage in favor of the response-guided designs (53.76%). Of the response-guided studies, 39 contained an explicit statement (e.g., treatment was implemented after stable performance in baseline was demonstrated), while the other 11 appeared to be response-guided because of an irregular spacing of the interventions. Eleven articles made an explicit statement about the systematic assignment of intervention points, and 32 had an equal spacing of the intervention points what is indicative for a systematic assignment. Only one study explicitly mentioned the random assignment of the order in which the units would be subjected to the intervention. Concerning the number of units per study and the number of observations per unit, we calculated the median number for each of the articles. Of these median numbers, details are provided in Table 1.

*Table 1*. Descriptive Statistics Including Tukey's Five Number Summary (Minimum, Lower-hinge, Median, Upper-hinge, Maximum) and the Mean, for the Number of Units per Study and the Number of Observations per Unit.

| | Minimum | Lower-hinge | Median | Upper-hinge | Maximum | Mean |
|---|---|---|---|---|---|---|
| Units | 2 | 2 | 3 | 3 | 8 | 3.09 |
| Observations | 8.5 | 23 | 35 | 52.5 | 200 | 42.94 |

Finally, we looked at the way in which the data from the multiple baseline designs were analyzed. We found that in all of the articles, the authors presented graphical displays and visually inspected the data. In 75.3% of the studies these plots were accompanied by some descriptive statistics, usually means and ranges. In only 5 out of the 93 articles the visual analysis was complemented with statistical analyses on the time-series data (in four articles, *t* tests were used, and in one article the authors calculated *t* tests on Pearson correlations). In Appendices B and C, master tables are provided containing all the information used in this section.

## Discussion

Although the prevalence of 93 articles is lower than the percentages found by Long and Hollin (1995), we can conclude that that multiple baseline designs still are relatively popular in the applied behavior sciences over the last 15 years. In this time period we noticed some fluctuations in the frequencies. Whereas in 1993 multiple baseline designs were used in 18.75% of all articles in *JABA*, after 1994 this number dropped to only three instances in 1995. During the late nineties of last century and the first years of this century, the prevalence of multiple baseline designs stayed rather low with some exception peaks in 1996, 1998 and 2004. From 2006 onwards, an upward trend appears to be visible. In both 2006 and 2007, multiple baseline designs again account for more than 15% of the articles in the *Journal of Applied Behavior Analysis.*

When we took a closer look at some design features, it became apparent that the multiple baseline across subjects design was applied most frequently. Besides that this is the most well-known variant, this observation could also be explained by the fact that the majority of the studies were interested in the effects of some behavioral intervention (e.g., extinction) on one kind of problem behavior (e.g., self-injurious behavior) in several subjects (e.g., adults with mental retardation).

Concerning the placement of the intervention points, we found that response-guided and systematic assignment procedures were about equally present. This observation, however, needs to be interpreted with caution, because it could be an artifact of the way of coding. In most studies, it was not explicitly mentioned how the placement of the intervention points was decided. Often we had to derive this from the graphical displays of the data. When the intervention points appeared to be spaced equally, we assumed that a systematic procedure was used, because it would be rather unusual to obtain such an equally staggered pattern with response-guided assignment. This could however also be possible. On the other hand, when the spacing of the intervention points appeared to be unequal, we assumed that this was due to a response-guided assignment procedure. However, the researchers could have decided beforehand to obtain such an irregular spacing, or the placement of the intervention points could be due to other factors, like sickness,… and not so much based on an emerging pattern in the data, like for instance stable baselines.

We found a median number of three units per study, which seems large enough, since the use of a minimum of two units is in many situations quite adequate and sufficient (Kazdin, 1982). This median of three units concurs with the recommendation of Barlow and Hersen (1984) to use minimum three to four units. Kazdin and Kopel (1975) however suggested at least four units. This is because the use of more units can demonstrate effects in a clearer and more convincing way. First of all, more units provide more replications, based on which stronger conclusions about treatment effects can be made (Ferron & Scott, 2005). Second, this has to do with ambiguity in the interpretation of results from multiple baseline designs. Although multiple baseline designs can provide rather strong inferences when all treatment effects occur at the time that the intervention is introduced for each unit, the conclusions are less clear when the effects are inconsistent across the different subjects, behaviors, or settings (Ferron & Scott, 2005). This is the case when not all units change at the point at which the intervention is introduced and it can have some serious implications for the internal validity when there are only two units, because then extraneous effects cannot 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 are related 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). Besides using four or more units, Kazdin and Kopel (1975) made two other recommendations to maximize the possibility

to draw valid inferences: researchers should select units that are as independent from each other as possible and implement a reversal phase for one of the units. These suggestions could attenuate the ambiguity.

Regarding the number of observations per unit, we found a median number of 35. This median of 35 measurement occasions seems reasonably high enough. However, the minimum number of 8.5 appears rather worrisome, since more observations can yield stronger conclusions about treatment effects. A mitigating factor is the fact that several studies included in the review used a multiple probe technique in which observations are done on an intermittent instead of a continuous basis (Horner & Baer, 1978), which without doubts has brought down the minimum.

Finally, concurring with Long and Hollin (1995), we found that most of the studies relied solely on visual analysis of the data. Even worse than the fact that statistical analyses were almost always lacking, is that all studies that did use statistics, employed *t* tests, which are controversial in the analysis of multiple baseline data. As indicated before, most problematic is the assumption of serial independence. When one wants to use these tests, the safest approach is to perform a preceding analysis of serial dependency. Unless computation of autocorrelations shows that the data are free from such serial dependency, alternative tests should be considered (Kazdin, 1984; Long & Hollin, 1995). We believe that randomization tests provide a strong alternative and may be preferred for analyzing data from multiple baseline designs. Here, the key feature is the randomization

of an aspect of the design: the experimenter has to randomly determine the moments of phase change. This random determination of the start of the different intervention points may seem incompatible with response-guided experimentation, where the transition from one phase to the next is based on the emerging pattern in the data. However, this concern can be partially addressed by, for example, starting the random assignment after all baselines have stabilized (Edgington, 1975, 1980a; Ferron & Sentovich, 2002; Koehler & Levin, 1998). Once the randomization schedule is decided, data can be collected according to the randomly chosen design, and the observed test statistic can be calculated from the obtained raw data (Adams & Anthony, 1996). By locating the observed test statistic in the randomization distribution (an equivalent of the sampling distribution in parametric statistical testing), the statistical significance of the outcome can be determined (Murray et al., 2004). The proportion of test statistics in the randomization distribution that exceeds or equals the observed test statistic gives the randomization test's *p* value. When this value is less than or equal to the predetermined significance level $\alpha$, the null hypothesis is rejected (Potvin & Roff, 1993).

We believe that the fact that we found no articles in which randomization tests were used is mainly due to the fact that researchers are not acquainted with these analyses and that most widely used statistical software programs do not include the possibility of conducting randomization tests for multiple baseline designs. To overcome this problem, we created an R package to

perform randomization tests on single-case data, including data from multiple baseline designs. R is an open source implementation of the S-PLUS language, which can be downloaded at no cost from the CRAN website (Comprehensive R Archive Network; cran.r-project.org) and which runs on a variety of UNIX platforms, as well as on Windows and MacOS (Hornik, 2008). Besides that it is for free, R has the advantages of being a very powerful and extremely flexible statistical modeling tool, which is able to deal with unusual data sets and problems (Crawley, 2005; Kelley, 2007); and of having very flexible graphical possibilities, by which simple as well as complex displays of the data can be obtained (Dalgaard, 2002). Our SCRT-R package, to conduct the randomization tests, can be obtained at no cost from ppw.kuleuven.be/cmes/SCRT-R.html.

To conclude, we would like to encourage researchers to state more clearly how the study was designed, especially how the intervention start points for each of the units in the multiple baseline design were decided. This will make it a lot easier to interpret the results of the study, and will make it more feasible for other researchers to try to replicate the study as accurately as possible. We would also like to encourage applied researchers to complement their visual analysis of the data with a statistical one. We believe that in many instances randomization tests are well-suited for this purpose.

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


## Abstract

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.

# Introduction

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 et al., 1968; Barlow & Hersen, 1984; Hayes, 1981; Kazdin, 1982; Kinugasa et al., 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).

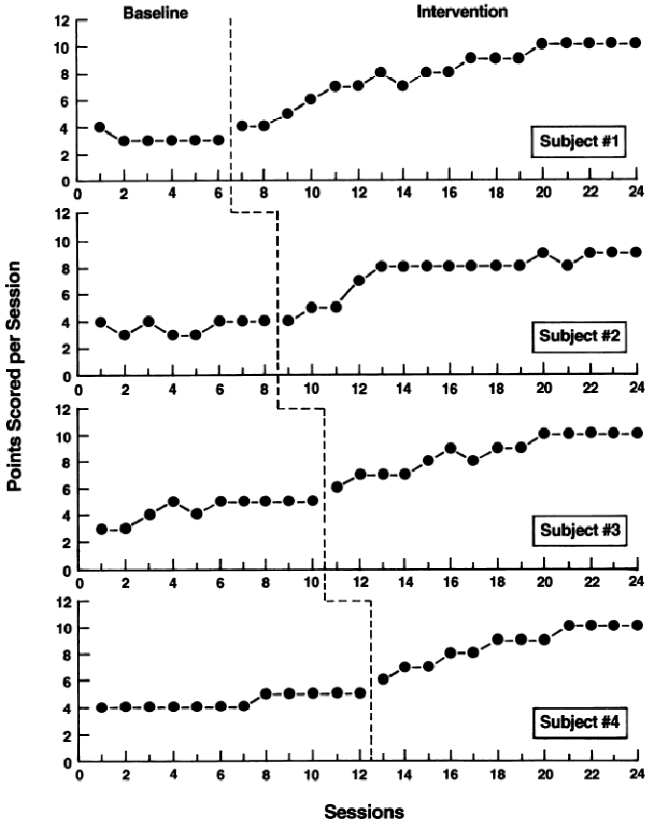



*Figure 1.* Results of the attentional shift training program. The number of points per session are shown for each of the four subjects. Each session consisted of 12 attempts (4 drills x 3 attempts), so the maximum possible score for each testing session was 12.

## 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 et al., 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, 1997; 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 et al., 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, 1997; 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 et al., 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 $\alpha$ (Murray et al., 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, 1980a; 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 *i*th 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.

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., Edit-Pad 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 the possible intervention start points for the units in our example.

*Table 1.* Data Obtained in the Experiment of Ziegler (1994)

| 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 |

*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 |

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="BA",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. Appendix D 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 observed 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.

# Meta-analysis of single-case studies: The SCMA package


## Abstract

Combining data from individual cases in different studies or in the same study can be a meaningful source of information for social and behavioral sciences. While group studies provide information on the general effect in the population and single-case studies provide information on the effect for a specific individual, a meta-analytical approach of single-case studies can combine the strengths of both kinds of studies. However, one of the obstacles for the researcher interested in this approach is the lack of user-friendly and comprehensive software to carry out these analyses. Therefore we developed a package for this kind of meta-analysis of single-case studies, for probability combining as well as for calculating some effect size measures on this data, using the statistical programming language R.

# Introduction

A meta-analysis is a quantitative integration of two or more studies, in which statistical methods are used to combine and compare the results of the primary studies and to come to more general or nuanced conclusions (Hedges, 1992; Hedges & Olkin, 1985; Mullen & Rosenthal, 1985; Shadish et al., 2002). By integrating the results of several studies, statistical power tends to be larger, more accurate effect size estimations are possible, and more convincing generalizations can be made. Also, moderator variables can be identified to explain the differences in study outcomes. Meta-analyses can help bridging the scientist-practitioner gap, by preventing delays in the introduction of effective treatments and leading to the timely identification of adverse effects (Zwahlen, Renehan, & Egger, 2008). Finally, meta-analyses are generally considered to be more informative, more accurate, and more objective than narrative methods and have less chance of giving distorted results (Adams et al., 1997).

Researchers have attempted to statistically summarize the results of several studies since the beginning of the 20th century. However, it is only since the article of Glass (1976) that the term 'meta-analysis' was introduced and that this idea was explicitly elaborated. Glass's contribution increased the popularity of this method spectacularly in disciplines such as social and behavioral sciences. Furthermore, this popularity led to the development and refinement of different techniques and to an explosion in the number of review articles (Hedges, 1992).

In sharp contrast to the many meta-analyses in which group comparison studies are combined and compared, only very few meta-analyses can be found in which the results of single-case designs are integrated (Shadish & Rindskopf, 2007). However, single-case designs make a unique and substantial contribution to clinical as well as to more fundamental process and outcome research (Kratochwill & Levin, 1992; Onghena & Edgington, 2005) and the poor involvement of single-case studies in meta-analyses holds the risk that the conclusions drawn from meta-analyses are incomplete, and in the worst case even distorted (Faith, Allison, & Gorman, 1997; Shadish & Rindskopf, 2007; Van den Noortgate & Onghena, 2003a). An important reason for omitting results from single-case studies in quantitative research syntheses is that the developers of meta-analytical techniques focussed mainly on the aggregation of research results from group studies. Different methods for applying meta-analytical techniques to the combination of single-case data were suggested, but no generally accepted methodology has yet been developed.

In this article we will first give an overview of several suggested methods for meta-analyzing single-case data. More in particular we will focus on some methods for combining *p*-values and combining effect sizes. For a part of these methods R packages are already available (e.g., 'metafor' and 'nlme'). We will conclude this article with the introduction of an R package for probability combining; an approach for which no other free software exists yet. The package also contains the possibility to calculate some measures of effect size for single-case data.

# Single-case research synthesis

The suggested methods for meta-analysis can roughly be divided into two classes. In the first, a general *p*-value is calculated by statistically combining the *p*-values of a number of studies, to determine whether there is statistically significant evidence for an effect. In the second class of meta-analytical techniques, the observed size of the treatment effect is calculated for each case, and these effect sizes are combined to come to one general effect size estimate, or are analyzed using adapted ANOVA or regression models to explore the moderating effect of case and/or study characteristics.

## Methods based on p-values

Although the most widely used method of data analysis for single-case studies is still visual inspection (Brossart, Parker, Olsen, & Mahadevan, 2006), several statistical tests have been suggested for analyzing single-case data. Examples are parametric *t* and *F* tests, time series analysis, nonparametric rank tests, randomization tests, and hierarchical linear models. The result of a statistical test is a (one-tailed) *p*-value, which indicates the probability that the obtained result, or a more extreme one, would have been observed if the null hypothesis were true.

Several meta-analytic procedures allow for the combination of *p*-values of a number of independent cases. These procedures result in a statement about a general (or omnibus) significant effect. The earliest methods classified results as 'significant' or 'nonsignificant' and used the binomial distribution to determine the expected

number of significant results given a true null hypothesis (Jones & Fiske, 1953; Rosenthal, 1978). Later, more powerful methods were developed in which probability values from independent significance tests are combined to statistically integrate research. Important assumptions are that the combined results are statistically independent and that the studies test essentially the same hypothesis. A global null hypothesis, which states that all separate null hypotheses are true, and a global alternative hypothesis, which states that at least one of the sub null-hypotheses is not true, are formulated. A functional combination of the separate *p*-values is used as a test statistic to determine whether a general significant result is obtained. The general combined *p*-value is calculated as the probability of getting a test statistic as small as or smaller than the observed test statistic (Edgington & Onghena, 2007).

The Fisher method (see e.g., Jones & Fiske, 1953; Pesarin & Salmaso, 2010; Rosenthal, 1978) relies on the addition of the natural logarithms of the reciprocals of *n* independent *p*-values. The negative natural logarithm of a uniformly distributed stochastic variable follows an exponential distribution. When stochastic variables that are exponentially distributed are scaled by a factor of two, this results in a value that follows a chi-square distribution with two degrees of freedom (Johnson, Kotz, & Balakrishnan, 1995). The sum of *n* independent chi-square values, each with two degrees of freedom, follows a chi-square distribution with 2*n* degrees of freedom: $\chi^2 = -2\sum_{i=1}^{n} \log_e p_i$. This sum is used as a test statistic to

determine the significance of the overall combination of *p*-values. Edgington (1972a) refers to the Fisher method as the 'multiplicative method', because, although the arithmetic operation performed is addition, logarithms are added and the sum of the logarithms of *p*-values equals the logarithm of the product of these *p*-values. He suggests using this method in the case of heterogeneous *p*-values. A possible disadvantage of the multiplicative method is that it can yield results that are inconsistent with the sign test (Rosenthal, 1978).

Edgington (1972a) proposed an alternative, which he labeled the 'additive method', and which is based on the addition of the *p*-values of two or more independent studies. The question asked to determine the general significance of the outcome is: what is the probability of obtaining such a small (or an even smaller) sum of *p*-values under the null hypothesis? This probability is given by $\frac{S^n}{n!} - \binom{n}{1}\frac{(S-1)^n}{n!} + \binom{n}{2}\frac{(S-2)^n}{n!} - \binom{n}{3}\frac{(S-3)^n}{n!} + \ldots$, where S is the sum of the n *p*-values. The plus and minus signs alternate and additional terms are added as long as the number in the numerator that is subtracted from *S* is smaller than *S*. These correction terms are needed to take into account the impossibility of having individual *p*-values that exceed the value of 1. Edgington (1972a) argues that this method has the advantage of being more powerful for homogeneous *p*-values (i.e., all *p*-values indicate the same result) than the multiplicative method. Another advantage is that when combining several *p*-values of .50, the combined *p*-value will always equal .50, no matter how many *p*-values are combined. This is not the case for

the multiplicative method, for which the combined *p*-value tends to become smaller and smaller as the number of *p*-values to combine gets larger. Hedges and Olkin (1985), however, state that the additive method can also result in wrong conclusions, since large *p*-values can compensate small ones. Additionally, Pesarin and Salmaso (2010) remark that this method is not consistent, more specifically that there is no guarantee that the null hypothesis will be rejected when it is false, if an infinite number of *p*'s is considered.

Because of the computational burden with the additive method when there are a lot of *p*-values to be combined, Edgington (1972b) developed the 'normal curve method' for situations with four or more probability values. Based on the assumption of approximate normality of the theoretical distribution of the means, this method uses the mean of the added probability values. First, an ordinary mean of the *n p*-values is calculated. Under the null hypothesis, each *p*-value has a uniform distribution [0,1], this is a distribution with mean .50 and standard deviation $\sqrt{\frac{1}{12}}$. Next, the *z*-score of the obtained mean is calculated by the formula $\frac{\overline{p} - .50}{\sqrt{\frac{1}{12}} / \sqrt{n}}$. Then the probability of the mean of the added probability values is determined by the use of normal curve tables.

Both the multiplicative and the additive method assume that the individual *p*-values come from a continuous distribution. This assumption of continuity, however, does not hold for all statistical

tests. *P*-values from nonparametric tests, for example, cannot take any value between 0 and 1; for small data sets, only a few *p*-values might be theoretically possible. If this discreteness of the distribution of *p*-values is ignored, the combined probability is overestimated. Edgington and Haller (1984) describe methods to combine discrete *p*-values (potentially combined with continuous *p*-values), both for the multiplicative and the additive case.

Other suggested methods include Stouffer's method, in which the *Z*'s associated with the *p*-values are added and divided by the square root of the number of studies, after which the resulting *Z*-value is converted to an overall *p*; Mosteller and Bush's modification, that computes a *t*-test on the obtained *Z* value; and Winer's suggestion of adding *t* values and dividing the sum by the square root of the degrees of freedom (see e.g., Rosenthal, 1978, for an overview and comparison of some methods). Strube and Miller (1986) found comparable power rates for all those combination procedures. Pesarin and Salmaso (2010) describe even more combination methods (e.g., Liptak, Tippett, Lancaster combining functions). Because each of those can give slightly different overall *p*-values, he suggests iterating the combination procedure by applying more than one combining function to the same partial tests and then combining the resulting second order *p*-values into a third order of combination by means of one of the combining functions. All these methods are not especially designed for combining *p*-values from single-case studies, but could certainly be used for this purpose.

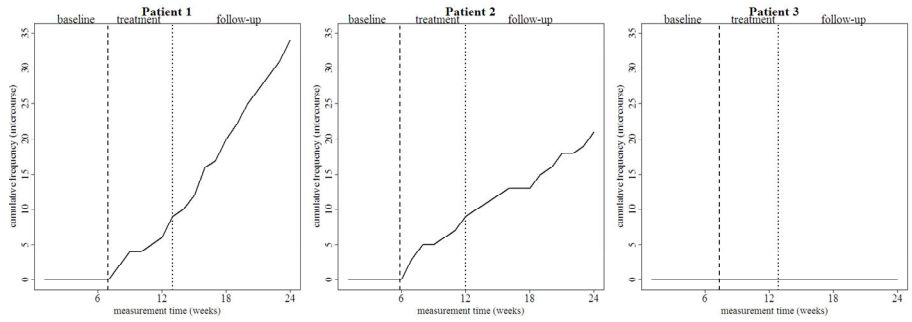
As an illustration, consider the results of ter Kuile et al. (2009). They investigated the effectiveness of therapist-aided exposure for women with lifelong vaginismus, using a replicated single-case AB phase design. Figure 1 shows the cumulative frequency of successful sexual intercourse attempts per week for patients 1, 2, and 3. The single-case data for each patient were analyzed using randomization tests, with a restriction of minimum 42 and maximum 84 baseline observations. As can be seen in Figure 1, both combined *p*-values are significant at a .05 level while none of the ten individual *p*-values is. This demonstrates one of the advantages of this meta-analytical approach: the statistical power increases considerably (see e.g., Strube, 1985).

## Methods based on effect sizes

The APA Task Force on Statistical Inference recommended to always provide some effect size measure when reporting a *p*-value (Wilkinson and the Task Force on Statistical Inference, 1999). These effect sizes can help evaluating the stability of the results, by comparing them to previously reported ones, and help to inform power analyses and meta-analyses. For single-case data, several effect size measures have been suggested.

Various versions of a standardized mean difference (SMD; left panel of Figure 2) were proposed (e.g., Busk & Serlin, 1992; Gingerich, 1984). These versions are based on the SMD that is commonly used in group research. Whereas in large-*n* studies this involves taking the standardized difference between the means of the treatment and the control group (e.g., Cohen's *d*, Hedges's *g*, or Glass's Δ), for single-

case data the mean of the baseline condition is subtracted from the mean of the treatment condition and this difference is divided by the baseline (or pooled) standard deviation (Beretvas & Chung, 2008).




| Individual p-values | | | | |
|---|---|---|---|---|
| Patient 1: p = 0.1163 | Patient 2: p = 0.0930 | Patient 3: p = 1.0000 | Patient 4: p = 0.0698 | Patient 5: p = 0.1628 |
| Patient 6: p = 0.4651 | Patient 7: p = 0.0698 | Patient 8: p = 0.0698 | Patient 9: p = 0.4651 | Patient 10: p = 0.4651 |
| **Combined p-values** | | | | |
| Multiplicative: combined p = 0.0317 | | Additive: combined p = 0.0126 | | |

*Figure 1.* The one-sided p-values of 10 cases are combined using the multiplicative (Fisher) and the additive (Edgington) procedure.

In addition, several nonparametric measures based on overlap between the data in the baseline and the treatment phases have been suggested. Scruggs, Mastropieri, and Casto (1987a) introduced the 'percentage of nonoverlapping data' (PND; middle panel of Figure 2), which is defined as the proportion of data points in the treatment phase that exceeds the most extreme value in the baseline phase (i.e., the proportion lower than the lowest baseline point for interventions designed to decrease unwanted behavior, and the proportion higher than the highest baseline point for interventions intended to increase desired behavior). Extensive debating between

Scruggs and colleagues and opponents of their method has been going on in scientific literature (see, e.g., Allison & Gorman, 1993; Scruggs, Mastropieri, & Casto, 1987b; White, 1987), causing Scruggs and Mastropieri (1994) to acknowledge that their procedure suffers some shortcomings. Despite clear caveats, such as extreme sensitivity to outliers and sample size, PND seems to be the most commonly used metric to describe treatment effects in single-case studies (Beretvas & Chung, 2008; Maggin, O'Keeffe, & Johnson, 2011). Ma (2006) introduced the 'percentage of data points exceeding the median' (PEM; right panel of Figure 2). For this metric, the overlapping data points are defined as the data points in the treatment phase that are located above the extended middle (median) line of the previous baseline phase (or below this line, if the undesired behavior is expected to decrease).

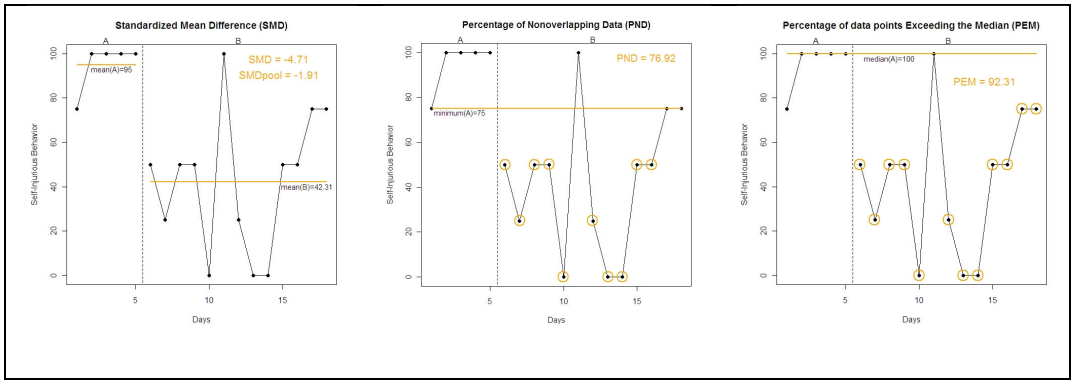



*Figure 2.* As an illustration, the Standardized Mean Difference (left panel), the Percentage of Nonoverlapping Data (middle panel), and the Percentage of Data points Exceeding the Median (right panel) are calculated for fictitious data.

Other suggested effect sizes are based on the use of generalized least-squares regression approaches. Gorsuch (1983) was one of the first to suggest the use of such methods for estimating effect sizes from single-case studies, and several similar suggestions followed (e.g., Allison & Gorman, 1993; Center, Skiba, & Casey, 1985-1986;

White, Rusch, Kazdin, & Hartmann, 1989). In these models the scores are regressed on the phase, the trend and the trend*phase interaction, and regression coefficients are calculated based on the ordinary least squares criterion. Van den Noortgate and Onghena (2003b) complemented these methods by suggesting to use (standardized) coefficients as effect size measures, instead of trying to summarize different coefficients into one measure. They do not calculate the effect size based on the *F* ratio, but instead use the regression coefficients themselves. By dividing the regression coefficients by the root-mean squared error (and the corresponding (co)variance matrix by the mean squared error), they can be standardized and as such made comparable over different studies. In the absence of a linear trend, this effect size measure simplifies to the standardized mean difference.

The effect size measures listed above provide an overview of the most frequently used methods for describing single-case data. More recently, other effect size measures were developed. An overview and comparison can be found in, e.g., Brossart, Parker, Olsen, & Mahadevan (2006); Manolov and Solanas (2008); Manolov, Solanas, and Leiva (2010); Manolov, Solanas, Sierra, and Evans (2011); Parker and Brossart (2003); Parker and Hagan-Burke (2007); Parker, Vannest, and Davis (2011); and Wolery, Busick, Reichow, and Barton (2010).

Not only *p*-values, but also effect sizes can be combined. The simplest method is just taking the (weighted) average of all effect sizes to obtain one overall measure that reflects the general effect of the

intervention over the different studies. This could for example be done with the effect size estimates based on data (non)overlap. The most recent technique entails the use of hierarchical linear models, which have a greater capacity than ordinary regression techniques in modeling correlated residual-error structures and allow more extensive modeling of within-case and between-case variation (Shadish, Rindskopf, & Hedges, 2008). Van den Noortgate and Onghena (2003b) suggest using such models for combining effect sizes (standardized coefficients) of single-case data. For a two-level model, with measurements nested in cases, this results in $d_j = \gamma_0 + u_j + e_j$ , with $d_j$ being the observed standardized mean difference of case *j*, $\gamma_0$ the overall standardized mean difference, $u_j$ a deviation typical of case *j*, and $e_j$ the sampling error for the observed standardized mean difference from case *j*. When in some studies more than one case is included, this model could be extended to a three-level model with measurements grouped in cases, and cases grouped in studies. Furthermore, study or case characteristics can be included as covariates to explain possible variation. Even more possibilities exist in the model to synthesize the results of several single-case studies by combining the raw data of the studies (Van den Noortgate & Onghena, 2003a).

## SCMA: A software package

One of the obstacles for the researcher interested in meta-analyzing single-case data is the lack of user-friendly and comprehensive software to carry out these computations and analyses. Although over the past years some R packages for meta-analysis were developed (e.g., the Metafor package of Viechtbauer, 2010), they do not entail the possibility to combine *p*-values or to calculate single-case effect size measures. To fill this gap, we developed the R package SCMA. The package is available on CRAN (cran.freestatistics.org/index.html) and contains two functions: *ES* and *combine*.

The *combine* function contains two methods for the nonparametric combination of *p*-values: the multiplicative approach, using Pearson's formula, is evoked by the command *combine(method="x")* and the additive approach by the command *combine(method="+")*. After giving one of those commands, a pop-up window asks in which file the *p*-values to be combined can be found. Instead of selecting the file by navigating through pop-up menus, also the argument *pvalues=read.table("…")* can be used, where the three dots should be replaced by the path and file name. Make sure to use double backslash (\\) instead of single backslash (\) and to enclose the path and file name in double quotes (" "), e.g., *combine(method="+", pvalues=read.table("C:\\Documents and Settings\\pvalues.txt"))*. The *p*-values should be stored in a .txt file with one column (all *p*-values below one another), without column or row labels.

With the *ES(design,ES,data)* function, several effect size measures can be calculated by selecting the appropriate *ES* argument. We included the *standardized mean difference* which uses the standard deviation of the A phase as denominator (*ES="SMD")*, the *pooled standardized mean difference* which uses the square root of the pooled variance as denominator (*ES="SMDpool")*, and two nonparametric effect size measures based on overlap between data in the baseline and the treatment phases: the *percentage of nonoverlapping data* (*ES="PND+"* or *ES="PND-"*, depending on whether an increase or a decrease of behavior is expected following the intervention) and the *percentage of data points exceeding the median* (*ES="PEM+"* or *ES="PEM-"*). All these measures can be calculated for single-case alternation designs (*design="CRD"* for a completely randomized design; *design="ATD"* for an alternating treatments design; *design="RBD"* for a randomized block design), single-case phase designs (*design="AB"; design="ABA";* or *design="ABAB"*, depending on the number of phases), and multiple-baseline designs (*design="MBD"*). A more detailed explanation of these design types can be found in Barlow et al. (2009). The .txt files containing the data can easily be made in a text editor or in Excel (saved as "text (tab delimited)"). For phase designs and alternation designs, the file should include the data for one case and consists of as many rows as there are observations and of two columns: one containing the condition labels ("A" and "B" for studies with two conditions/phases and "A1", "A2", "B1", and "B2" when there are more phases) and one containing the observations. The rows and columns should not be labeled. For multiple baseline AB designs, with

multiple subjects, settings or outcomes as units, the data file should consist of these two columns for each unit (so twice as many columns as there are units). The *data* argument operates the same way as the *pvalues* argument in the above described *combine* function. The resulting effect size measures can be combined using other existing R commands and packages.

Please note that another obstacle when statistically re-analyzing published data is that researchers do not always have access to the raw data. If however graphs were presented in the original studies, the *graph.extract* function can be used, with which data values can be extracted from existing graphs (SCVA R-package: Bulté & Onghena, 2012).

## Discussion

The advantages of meta-analyses have been questioned since the beginning. Because of the controversial results from the first appearing meta-analyses on psychotherapeutic outcomes, and the possible subjectivity in the selection, classification, and coding of the primary studies, questions were raised on the claimed objectivity and superiority of this method. Other main objections include the fact that the findings could be distorted by 'publication bias', the inclusion of studies of varying methodological quality ('garbage-in, garbage-out'), and the heterogeneity of the integrated studies (e.g., Eysenck, 1978, 1994; Shadish et al., 2002; Wilson & Rachman, 1983; Zwahlen et al., 2008). Like with all scientific methods, criticisms on meta-analyses will continue to appear, but this also helps improving the

methodology. The publication of potential problems leads to suggested solutions. Even though meta-analysis without doubt has limitations, many of the criticisms are not specific and would apply to all review methods, with the alternatives (e.g., narrative reviews) often being even worse (Shadish et al., 2002).

Because no consensus has yet been reached concerning the methods of calculating effect sizes and performing meta-analyses with single-case studies, there is no consistency in the published studies. This is problematic for the comparability of the results across different studies and the combining of effect sizes over studies. The above mentioned methods for combining *p*-values, as well as the effect size measures all have their own merits and limitations. More thorough research concerning the conditions under which each particular method could be preferred is therefore definitely needed. Based on the importance of having an empirically guided selection of metrics, future research should contrast the suggested measures of effect size to study the conditions under which each method is optimal. Additionally, methods for nonparametrically combining *p*-values of *p*-values should be examined in more detail. Loughin (2004) studied six methods for combining independent *p*-values and found that most of them have their own strengths, depending on the characteristics of the data. Some are better when only a few studies provide evidence against the global null hypothesis, while others are better when the evidence is more homogeneous, i.e. all studies are against or are in favour of the null hypothesis. As mentioned earlier, the additive and the multiplicative approach for the combination of *p*-values are

considered as complementary, with the former method being more powerful for homogeneous *p*-values, and the latter method being more powerful for heterogeneous *p*-values (Edgington, 1972a; Edgington & Haller, 1984). This statement was however empirically not strongly founded, but merely illustrated with a few numerical examples. This hypothesis of differences in sensitivity should be tested by means of a stochastic simulation study, to investigate which is most suited under which circumstances. Finally, the statistical power and robustness of those nonparametric combination approaches should be compared with those of the parametric multilevel approach.

# The Single-Case Data Analysis package: Analyzing single-case experiments with the R software


## Abstract

This paper describes RcmdrPlugin.SCDA, a plug-in package that integrates three R packages in the R commander interface: SCVA (for Single-Case Visual Analysis), SCRT (for Single-Case Randomization Tests), and SCMA (for Single-Case Meta-Analysis). This way the plug-in package covers three important steps in the analysis of single-case data.

# Introduction

To investigate research questions in educational, behavioral, and medical science, single-case experiments are very well suited. We want to bring these experiments to the attention of (applied) researchers and suggest a software package to analyze data resulting from single-case experiments.

Single-case designs are increasingly popular in educational, behavioral, and medical research (Hammond & Gast, 2010; Matson, Turygin, Beighley, & Matson, 2012; Shadish & Sullivan, 2011; Swaminathan & Rogers, 2007). Bliss, Skinner, Hatau, and Carroll (2008), for example, classified all articles published in four school psychology journals (*School Psychology Quarterly, School Psychology Review, Journal of School Psychology,* and *Psychology in the Schools*) between 2000 and 2005 and found that, with the exception of 2004, single-subject designs were more prevalent than group designs. In this time frame, 55% of the causal-experimental studies in these journals utilized single-subject designs. This may seem odd given the strong emphasis on large *N* designs in most statistical and methodological courses and handbooks for the educational, behavioral, and medical sciences.

A possible explanation for the findings of Bliss et al. (2008) lies in the fact that single-case designs can provide a viable alternative or supplement to group designs to answer causal questions. This can be said about educational, behavioral, and medical research in general, but the published applications of single-case designs are certainly not

that predominantly present in all subareas. In many subareas of educational, behavioral, and medical research single-case designs however there is a huge potential for single-case designs to complement the standard group designs because of the necessity of pilot data in early stages of larger group studies, because of the relevance of research concerning rare types of participants (e.g., patients with a very specific neuropsychological disorder due to brain injury), or the examination of idiographic questions like 'does this intervention (e.g., restructuring) work for this particular organization?'; and, of course, when research funds are scarce and it is not possible to obtain enough participants for large-scale group studies (Barlow et al., 2009; Edgington & Onghena, 2007; Franklin, Allison, & Gorman, 1997; Kazdin, 2011). Because of the close link of single-case evaluations to individual care, single-case designs are also ideally suited to bridge the scientist-practitioner gap (Barlow, Hayes, & Nelson, 1984; Bliss et al., 2008).

Unfortunately, most of the commonly used statistical software packages, like SPSS and SAS, do not present readily available procedures or options for designing single-case experiments and analyzing single-case data. To fill this gap, we present an R package for designing single-case experiments and analyzing single-case data.

R was chosen as the computational environment, because it is open source software, running on a variety of UNIX platforms, as well as on Windows and MacOS (Hornik, 2012). R has excellent graphical possibilities, but is also a very powerful and flexible statistical environment, which facilitates the combination of visual and

statistical data analysis (Kelley, 2007). However, because of the use of a standard command line interface, R is not very user-friendly. Especially for practitioners who never engaged in any basic programming, the threshold to start working with R can be too high.

Fortunately, Fox (2005) created a window-based graphical user interface (GUI) to R, called the "R Commander" with menus and dialog boxes (very similar to, e.g., SPSS), which are far more familiar to most people (see Figure 1). Working with the R commander is hands-down: by selecting the menus, submenus open which lead to dialog boxes. Each dialog box contains a 'help' button, which refers the user to a help page with more information. By making selections in the dialog boxes, R commands are generated and executed.

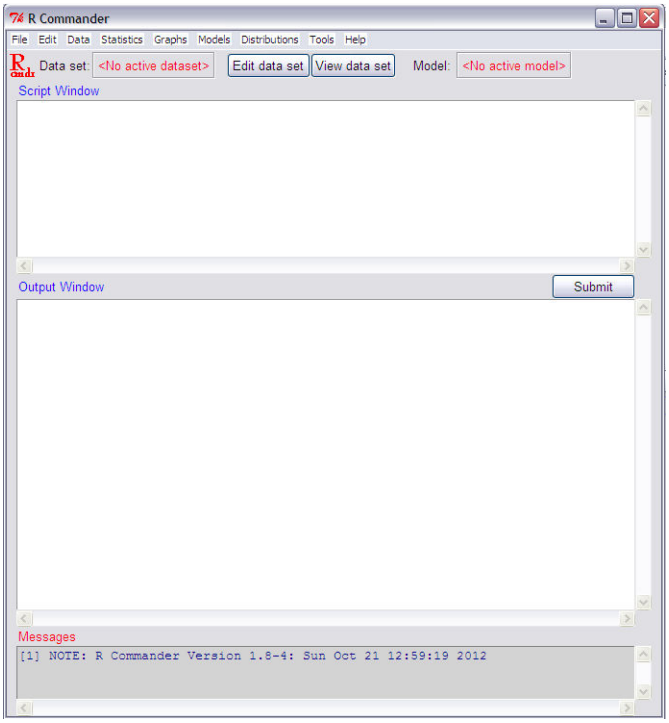



*Figure 1*. The R commander interface when starting up.

An advantage of the R commander is that these R commands are not just executed 'behind the scene'. They also appear in the *script window,* so R users can always see the code and adapt it if necessary. This visibility of commands can also be a useful first step in learning R, by gently getting acquainted with the R language. In this script window, also other commands can be typed, or previous commands can be rerun. In the *output window*, the given command appears together with the output, and in the *messages window*, error messages, warnings, and notes appear (Fox, 2005).

An additional advantage of the R commander is that users can add their own menus and dialog boxes. This extensibility became even more practical with the possibility of writing plug-in packages (Fox, 2007). Besides the already mentioned advantages of R and the R Commander, a huge advantage is that all functionalities that are already available in the standard R Commander and in other plug-in packages can be used. Instead of developing a stand-alone GUI, we therefore chose to create a plug-in for the R Commander. The SCDA (Single Case Data Analysis) plug-in we created is a GUI for our three R packages: SCVA (Single Case Visual Analysis; Bulté & Onghena, 2012), SCRT (Single Case Randomization Tests; Bulté & Onghena, 2008, 2009) and SCMA (Single Case Meta-Analysis; Bulté et al., submitted).

## The RcmdrPlugin.SCDA package: an illustration

When planning an experiment, first the study design should be carefully chosen based on, among other things, the research question. Then data can be collected according to the selected

design. Single-case data analysis, just like any other analysis of empirical data, best starts with a visual exploration of the data. Unless the visualization is that convincing that the effect of the intervention is very obvious (e.g., a dramatic shift in level without any variability or trend), statistical data analysis might be a useful supplementary technique. And often it is not only useful to know whether an intervention had a statistically significant effect, but also what the size of the effect was. The SCDA plug-in for the R commander provides R functions for each of those steps of research design and data analysis. It adds one menu item to the R commander ("SCDA"), which contains three submenus (see Figure 2). Each of the submenus leads to several menu items.

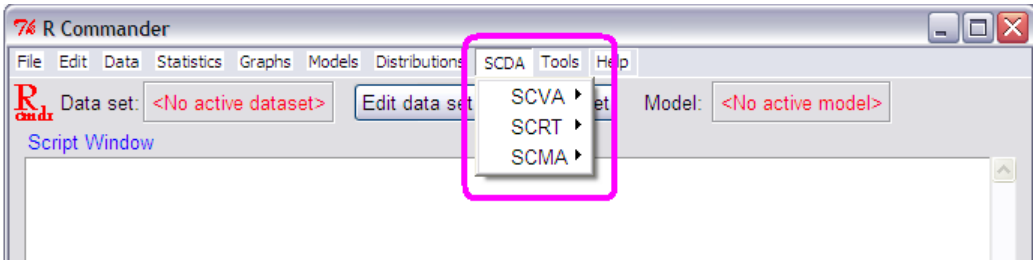



*Figure 2.* The R commander with the SCDA plugin.

**SCVA (Single-Case Visual Analysis)**

- Graphical display
- Plot measure of central tendency
- Plot estimate of variability
- Plot estimate of trend

**SCRT (Single-Case Randomization Tests)**

- Design your experiment
  -Number of possible assignments
  -Display all possible assignments
  -Choose 1 possible assignment

- Analyze your data
  -Observed test statistic
  -Randomization distribution
  -P-value

**SCMA (Single-Case Meta-Analysis)**

- Calculate effect size
- Combine *p*-values

In what follows we will illustrate the functionality of the package with an example from ter Kuile et al. (2009). A more detailed manual can be found in the manual.

Ter Kuile et al. (2009) investigated the effectiveness of therapist-aided exposure for lifelong vaginismus. Lifelong (or primary) vaginismus occurs when a woman has never been able to have sexual penetrative intercourse during her whole life. From a cognitive-behavioral perspective, fear and avoidance behavior are connected to vaginismus. Therefore it was hypothesized that exposure to the feared stimuli (i.e., penetration) would increase the ability to have sexual intercourse. The exposure therapy was aided by a trained female therapist and consisted of vaginal penetration exercises (performed by the woman herself) at the hospital, together with several specific homework assignments in which the partner was involved.

## Research design and data collection

Ter Kuile et al. (2009) used a replicated single-case AB-phase design to test whether the therapy would lead to an increase in sexual intercourse. In an AB-phase design all baseline measurements (A) precede all treatment measurements (B). The idea behind this design

is to be able to attribute a change in intercourse behavior after the exposure onset to the therapy. To control for time-related confounding variables, randomization was incorporated in the study design: the random aspect was the start of the exposure therapy. This random determination of therapy onset is of course not unlimited. To decide whether the therapy had any effect on the ability to have intercourse, baseline as well as treatment observations are needed. Therefore, in our illustration a constraint of a minimum of one week of diary recordings in each phase is guarded. With a total of 24 weeks of data recordings, this leads to 23 possible start points for the therapy (after week 1, after week 2, … , after week 23) (Figure 3: SCRT -> Design your experiment -> Number of possible assignments).

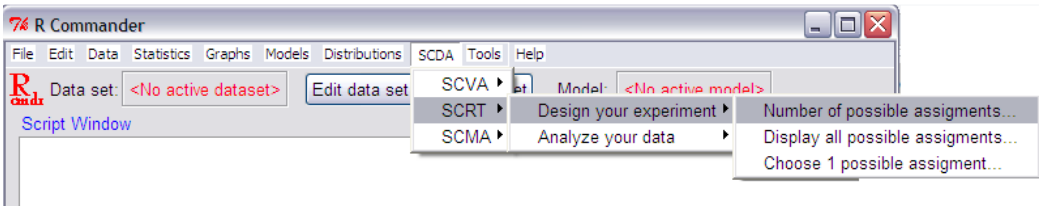



*Figure 3.* The RcmdrPlugin.SCDA menu, with the SCRT (Single-Case Randomization Tests) 'design your experiment' submenu.

By replicating this experiment over several participants, the strength of the findings is increased and more general statements can be made. Ten patients in sexology clinics, who suffered from primary vaginismus, participated in the study. They kept a daily diary in which they noted (amongst other things) whether they were able to have sexual intercourse with their partner that day. To analyze these data, they should be put into the R commander. This can be accomplished

by reading in a created a text file with the observations or by entering the observations directly as an active data set (see manual).

We will first focus on the diary data of the first patient (P1). She started recording her sexual intercourse attempts from the moment she was referred to the outpatient sexology clinic. The start of the exposure therapy was randomly determined after seven weeks of 'baseline' and she continued filling in the diary for seventeen more weeks.

## SCVA (Single-Case Visual Analysis)

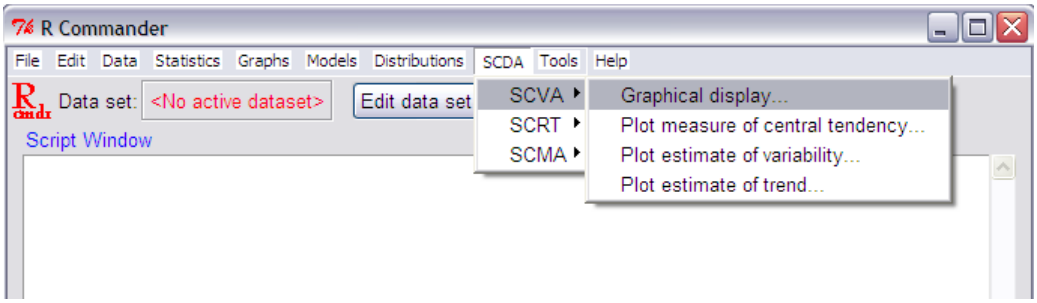



*Figure 4.* The RcmdrPlugin.SCDA menu, with the SCVA (Single-Case Visual Analysis) submenu.

A first impression can be obtained by making a graphical representation of her data (SCDA -> SCVA -> Graphical display; see Figure 4). The left panel of Figure 5 shows the average number of successful intercourse attempts per week for P1. The dotted vertical line indicates the start of the intervention phase. The introduction of the exposure therapy clearly made a difference, since before therapy this woman had never experienced sexual intercourse with penetration. The shift in mean level between the phases is illustrated

in the right panel of Figure 5, where the phase means are plotted as a horizontal reference line on the raw data.

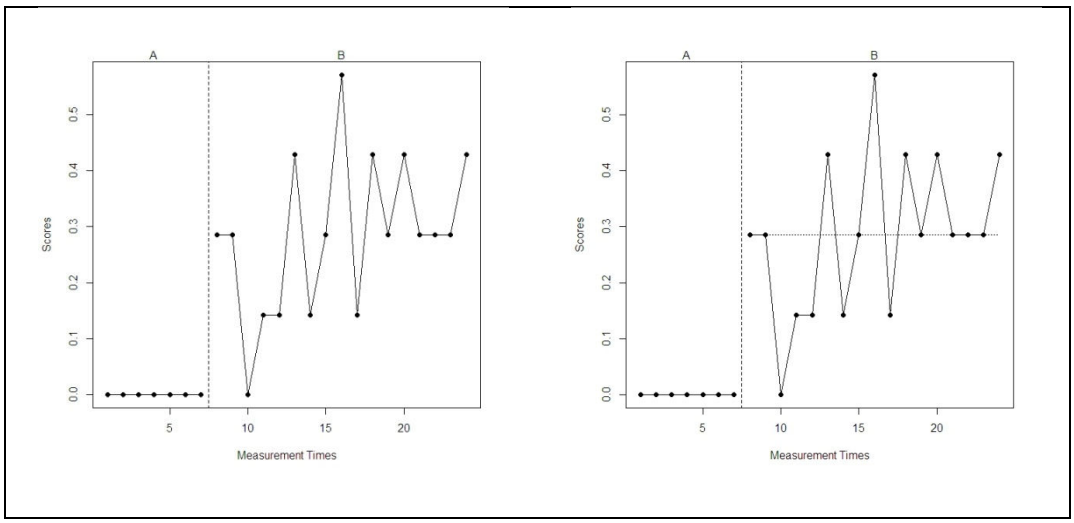



*Figure 5.* Graphical representation of the diary data of P1, averaged per week (SCDA -> SCVA -> Graphical display; see Figure 4). On the right, the shift in mean level is visualized (SCDA -> SCVA -> Plot measure of central tendency; see Figure 4).

The range bars in the upper left panel of Figure 6 illustrate the lack of variability in the baseline phase. In the treatment phase variation is shown by the vertical line connecting the minimum and the maximum value, with a small horizontal bar marking the phase mean. The trended range lines depicted in the upper right panel of Figure 6 show that the variability decreases over time. There is also an upward trend in level noticeable from the linear function in the lower left panel of Figure 6. That this trend is not entirely linear can be seen from the nonlinear smoothed curve produced by calculating running medians of batch size four and averaging each successive pair.

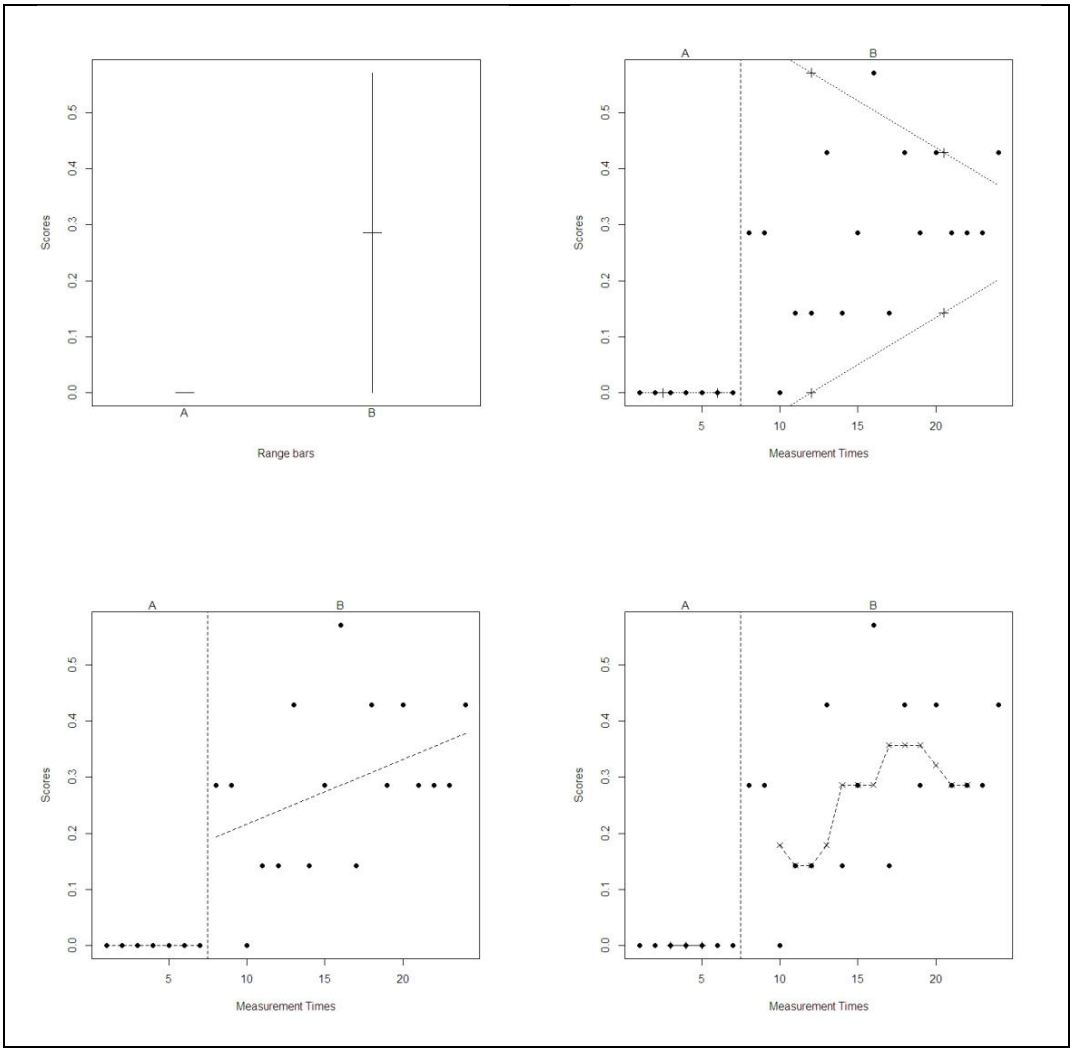



*Figure 6.* The upper panels illustrate the variation in the data (SCDA -> SCVA -> Plot estimate of variability; see Figure 4): range bars on the left and trended range lines on the right. Possible trends in the data are made visible in the lower panels (SCDA -> SCVA -> Plot estimate of trend; see Figure 4). On the left a linear function is drawn on the raw data by means of ordinary least-squares regression and on the right a nonlinear trend is visualized by displaying running medians of batch size four averaged by pairs.

## SCRT (Single-Case Randomization Tests)

Visual analysis is an important first step when evaluating intervention effects. In addition, several statistical tests might be conducted to evaluate the statistical significance of the intervention effects. The SCDA GUI includes functions to conduct randomization tests:

permutation tests based on random assignment, to test a null hypothesis about treatment effects in a randomized experiment (Onghena & Edgington, 2005). The alternative hypothesis in our illustration was that the exposure therapy will lead to an increase in successful sexual intercourse attempts. In other words, the mean of the treatment observations is expected to be higher than the mean of the baseline observations. Therefore, the difference between those means was chosen as test statistic.

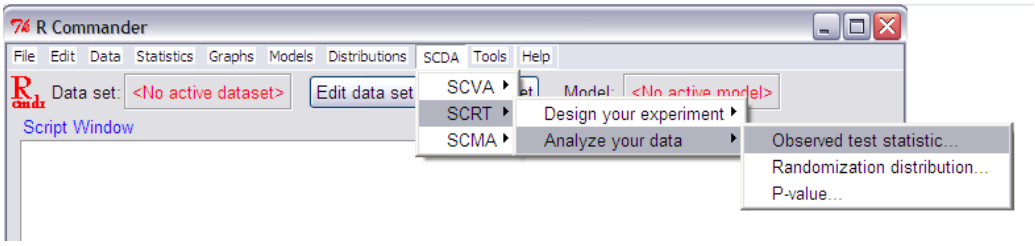



*Figure 7.* The RcmdrPlugin.SCDA menu, with the SCRT (Single-Case Randomization Tests) 'analyze your data' submenu.

Visual analysis of the data already indicated that the exposure therapy had a positive effect on the intercourse frequency of P1. This effect is also statistically significant, shown by the randomization test's *p*-value of .04 (Figure 7: SCRT -> Analyze your data -> P-value). More information on randomization tests can be found, for example, in Bulté and Onghena (2008).

## SCMA (Single-Case Meta-Analysis)

The combination of visual analysis and statistical significance testing, however, does not tell the whole story. It is often also useful to know what the size of the effect is. How large an effect is, is expressed by means of effect size measures (Figure 8: SCDA -> SCMA -> Calculate effect size). The pooled standardized mean difference, which uses

the pooled standard deviation of both phases, for P1 equals 2.35. The percentage of data in the treatment phase that is higher than the median of the baseline phase (=”PEM”) is 94%.

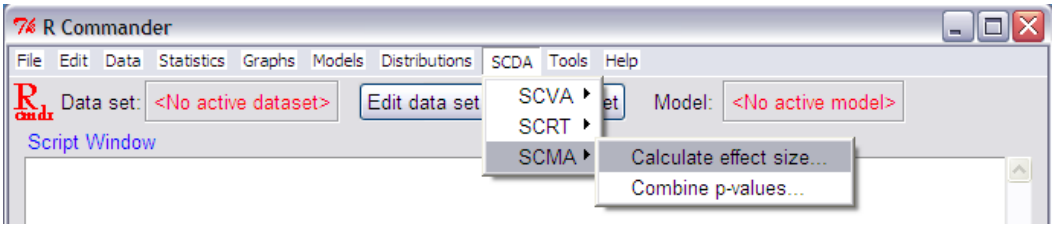



*Figure 8.* The RcmdrPlugin.SCDA menu, with the SCMA (Single-Case Meta-Analysis) submenus.

These effect sizes are not only used to show the magnitude of an effect, but they also help comparing results over different studies and they can be combined in meta-analyses to come to one general estimate of the effect size. Another meta-analytical procedure is statistically combining the *p*-values of several sequentially replicated single-case experiments, to determine whether a general significant result is obtained (Figure 8: SCDA -> SCMA -> Combine P-values). Ten women with primary vaginismus participated in the study of ter Kuile et al. (2009). When combining their results with the additive method for combining independent *p*-values, a general significant result of $p = .0016$ was obtained.

## Discussion

With this paper we wanted to briefly illustrate the use of the SCDA plugin package for the R commander. A more elaborated overview of all functionalities and an explanation of the different parts in the dialogue windows is given in the manual.

The presented software package contains one technique for statistically analyzing the data resulting from single-case experiments. We believe that these randomization tests have several advantages over some other techniques suggested in the literature. Most importantly, the presence of serial dependency in the data will not invalidate the result of a randomization test. This is of considerable interest in the context of single-case experiments, since these data tend to have autocorrelated residuals, which can seriously bias the results of for example *t*-tests, by inflating the Type I error rate. Another advantage is that they are free from the assumption of random sampling. This random sampling assumption, on which the probability tables of parametric tests are built, is usually an unrealistic ideal and most experiments do not use randomly sampled subjects. Also, unlike parametric *t*-test, randomization tests are not based on assumptions about the homogeneity of variances. But they do have one requirement: the incorporation of random assignment not only enhances the internal validity of the study, as indicated before, but it also justifies the application of a randomization test that is based on the random assignment used in the experimental design of the study. This way it is possible to infer causal relations between the treatment and the observed changes in behavior. Since randomization tests acquire validity by mirroring the random assignment schedule used in the study, their extreme versatility permits researchers to make valid tests by devising a test that is suitable for the particular design used. They can be used with all sorts of data (continuous, discrete, ranks,...) and all kinds of data patterns (trends, outliers, skewed distributions, zero variance in baseline,...)

(see e.g., Edgington, 1973, 1980b; Edgington & Onghena, 2007; Gorman & Allison, 1997; Kazdin, 2011; Ludbrook, 1994; Onghena & Edgington, 2005; Dugard et al., 2012).

An alternative to these nonparametric tests is time series analysis, which investigates whether there is a significant change in level and/or slope between the phases. This technique is also suitable for the analysis of data when serial dependency is present. It, however, requires many data points to determine the existence and the pattern of autocorrelation and to indentify the model correctly, which could cause problems for small-*n* experiments where the phases are usually rather short. Another difficulty is the complexity of the mathematical theories on which it is based (Box et al., 1994; Gorman & Allison, 1997; Kazdin, 2011). Another alternative is hierarchical linear modeling (HLM). Van den Noortgate and Onghena (2003a, 2003b) suggested using such a model for calculating effect sizes and for combining effect sizes of single-case data or combining the raw data of several studies. By modeling the hierarchical structure, the possible dependence of the scores is taken into account. Also study or case characteristics can be included as covariates to explain possible variation. For time series analysis already a plug-in for the R commander exists (RcmdrPlugin.epack). As far as we know, this is not the case for hierarchical linear modeling, but R packages do exist (e.g., 'nlme').

As a useful extension of the RcmdrPlugin.SCDA package, these techniques could be included. Not only the HLM suggestion of Van den Noortgate and Onghena (2003a, 2003b), but also other methods

for combining effect sizes could be adopted in the package. The simplest method is probably just taking the (weighted) average of all effect sizes to obtain one overall measure that reflects the general effect of the intervention over the different studies. This could for example be done with the effect size estimates already included in the package. Additionally, more effect size measures could be added, such as the percentage of all non-overlapping data (PAND; Parker, Hagan-Burke, & Vannest, 2007). Other possibilities for combining effect sizes include the three approaches presented by Busk and Serlin (1992), which differ in the assumptions made about the distribution and the variability of the data. Also the inclusion of more methods for combining *p*-values could be interesting: Stouffer's method, in which the *Z*s associated with the *p*-values are added and divided by the square root of the number of studies, after which the resulting *Z*-value is converted to an overall *p*; Mosteller and Bush's modification, that computes a *t*-test on the obtained *Z* value; Winer's suggestion of adding *t*-values and dividing the sum by the square root of the degrees of freedom (Rosenthal, 1978); and the iterative procedure for combining *p*-values, by applying more than one combining function to the same partial tests and then combining the resulting second order *p*-values into a third order of combination by means of a combining function, until the final *p*-value becomes reasonably invariant (Pesarin, 2001; Pesarin & Salmaso, 2010).

# Manual: Getting started with the RcmdrPlugin.SCDA

## Installation

To make the SCDA GUI work, R should be downloaded and installed. This can be done at no cost from the CRAN (Comprehensive R Archive Network; cran.r-project.org) website. Hornik (2012) gives a detailed explanation of how to do this for Windows, Macintosh and UNIX (the R package presented here is, however, only tested on Windows). Once R is running, the GUI can be installed from within the R console (Packages -> Install packages). After choosing a CRAN mirror nearby, the 'RcmdrPlugin.SCDA' package should be selected from the list. Note that the downloading process can take a while because several supporting packages are downloaded automatically as well. Then the RcmdrPlugin.SCDA package can be loaded into R by selecting 'Packages -> Load package' (or by typing the command *library('RcmdrPlugin.SCDA')* into the R console). Only this last step needs to be repeated when using the GUI. Note that under Windows, the R commander functions best with the single-document interface (SDI: Edit -> GUI preferences -> Single or multiple windows: SDI). By loading the RcmdrPlugin.SCDA package, the R commander opens with 'SCDA' as an additional menu button (Figure A1).

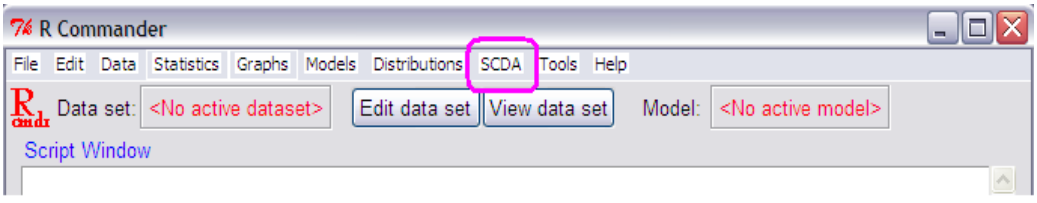



*Figure A1.* By loading the RcmdrPlugin.SCDA package, the "SCDA" menu button is added to the R commander interface.

## Data input

Most of the functions in the SCDA GUI need data input, which can be taken from a .txt file that has been created in advance in a text editor (e.g., EditPad or NotePad) or in Excel (save the file as 'Text(Tab delimited)'). Text files containing the raw data should consist of two tab-separated columns for single-case phase and alternation designs: the first with the condition labels ("A" and "B" when there are two conditions/phases, and "A1", "B1", "A2" and "B2" for three or four phases) and the second with the observations. For multiple-baseline designs they should consist of these two columns for each unit. This way, each row represents one measurement occasion. Text files containing *p*-values for the 'combine' function should consist of one column of *p*-values. In text files containing the possible start points for multiple baseline designs, each row should contain all possibilities for one unit, separated by a tab. It is important not to label the rows or columns.

Another way of data input is by using the active data set of the R commander. Creating an active data set has the advantage that also other functions than the SCDA-functions (e.g., built-in statistical functions of the R commander or from other plug-ins) can be applied to the data. This way one can fully benefit from the functionality of the R commander and its plug-ins. There are several ways to construct this active data set: a text file created as described above can be read into the R commander via the Data menu (Figure A2: Data -> Import data -> from text file, clipboard, or URL -> uncheck the box 'Variable names in file' if necessary -> OK -> select the text file

within the 'Open file' dialog box), or data can be entered directly via 'Data -> New data set'. The active data set can be consulted at any time by clicking the 'View data set' button.

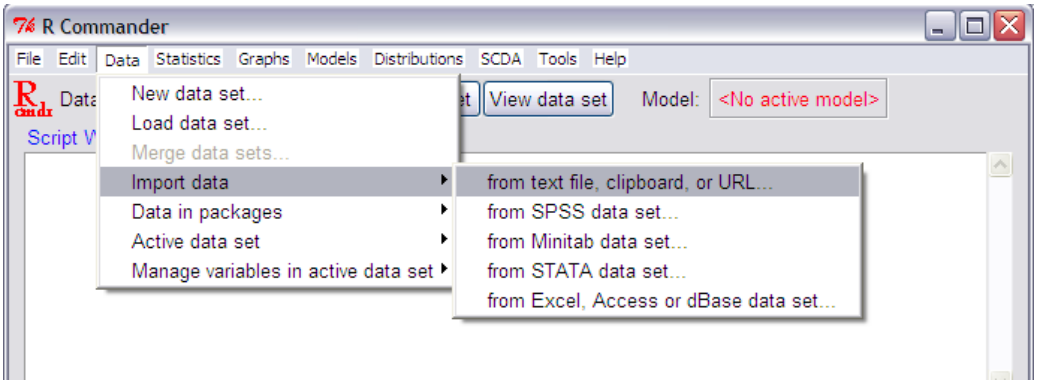



*Figure A2.* Constructing the active data set by reading a text file into the R commander.

## Functions

The SCDA menu contains three submenus with several menu items. Each menu item opens a dialogue box.

### SCVA (Single-Case Visual Analysis)

- **Graphical display**: the observed single-case data are plotted.

  As can be seen in Figure A3, two selections should be made in the dialogue box.

  In section A, the design type used has to be indicated. The options are:

  -*Phase designs:* comparisons are made within a time series and the subject's performance is evaluated over time across baseline (A) and intervention (B) phases.

  -> *AB Phase Design*: all baseline measurements precede all treatment measurements.

-> *ABA Phase Design*: withdrawal or reversal design in which the treatment is administered between two baseline phases.

-> *ABAB Phase Design:* an extra intervention phase is added.

-*Alternation design:* the basic strategy is the rapid alternation of two or more conditions within a single subject.

-> *Completely Randomized Design:* the random assignment procedure mirrors the one used when randomly assigning subject to different groups for an independent samples *t* test (no restrictions).

-> *Alternating Treatments Design:* the temporal clustering of treatments is prevented by ensuring that the randomization does not permit more than a specified number of successive time blocks with the same condition.

-> *Randomized Block Design*: adjacent treatment times are grouped together in blocks and the conditions are assigned randomly within each block.

-*Multiple Baseline Design*: This is an extension of the basic AB phase designs, in which several of those AB designs are implemented simultaneously to different persons, behaviors, or settings. A characteristic feature is that the intervention is introduced in a staggered way to the different units.

In section B, the data file in which the data can be found should be selected. This could either be the active data set or

a previously created text file. In the latter case, the file can be chosen with the 'Select File' button.

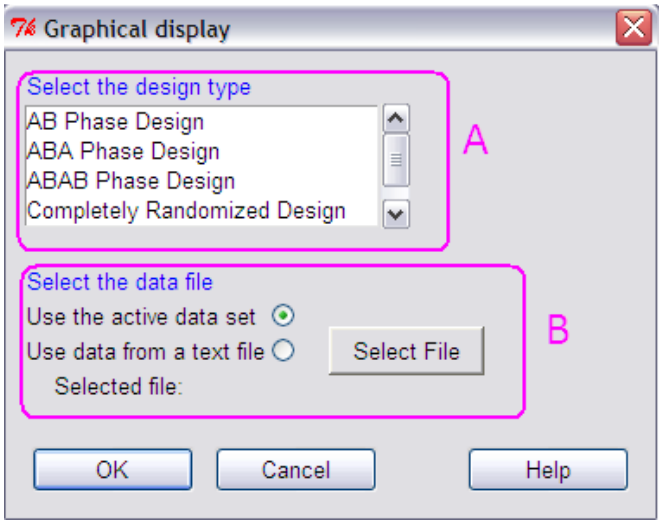



*Figure A3.* The dialogue box for the 'graphical display' menu.

- **Plot measure of central tendency:** a measure of central tendency is plotted as a horizontal reference line superimposed on the raw time series data.

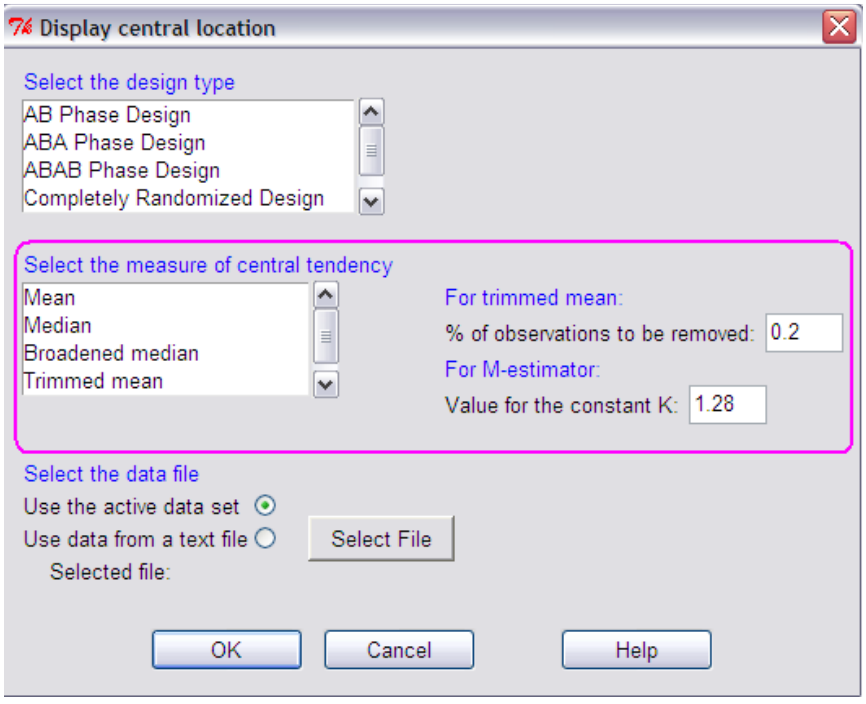



*Figure A4.* The dialogue box for the 'Plot measure of central tendency' menu.

In the marked section of the dialogue box (see Figure A4), the measure of central tendency should be selected. There are four built-in possibilities:

-*Mean:* the arithmetic mean (the average) of the data. This measure is rather sensitive to outliers.

-*Median:* the middle value of the data (for an even number of data points, the median equals the average of the two central data points). This measure is more robust than the mean, but it only takes into account the central data points, while disregarding other numerical information.

-*Broadened median:* calculated based on the three, four, five or six middle values of the data set (depending on the total number of data points). This measure is more robust than the mean, and sensitive to a larger proportion of the data than the median.

-*Trimmed mean*: calculates the mean after discarding the extreme observations. The percentage of observations that has to be removed from each end of the distribution can be indicated in the upper right box (Figure A4). This can be any value from 0 (= regular arithmetic mean) to 0.5. The default value is 20 percent. The trimmed mean is more robust than the mean and less affected by observations in the centre of the distribution than the (broadened) median.

-*M-estimator:* Huber's M-estimator of location first evaluates each observation to determine if it is actually an outlier compared to the rest of the data and then gives less weight

to those outlying values. For this evaluation a constant K needs to be specified that can have any value between 0 and ∞. Usually a percentile of the standard normal distribution is chosen. Wilcox (2005) suggests using K=1.28, which corresponds to the 90th percentile of the standard normal distribution and covers 80 percent of the underlying distribution. By determining the sensitivity of the estimator, one can balance between robustness and efficiency (see e.g., Huber & Ronchetti, 2009).

- **Plot estimate of variability:** information about variability in the data is displayed by one of three methods.

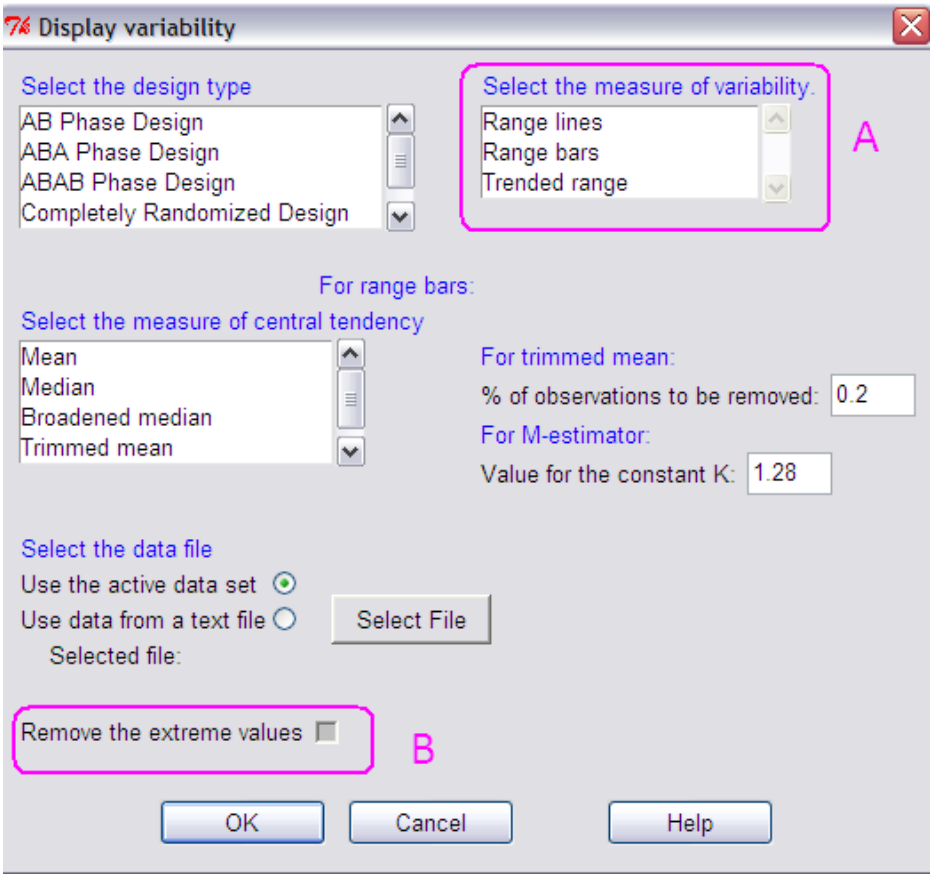



*Figure A5.* The dialogue box for the 'Plot estimate of variability' menu.

In box A (Figure A5) the measure of variability should be selected. The three options are:

-*Range lines* are a pair of lines parallel to the X-axis, passing through the lowest and highest values for each phase, and superimposed on the raw data.

-*Range bar graphs* consist of a vertical line for each phase, connecting the minimum and the maximum value, with a small horizontal bar crossing this line to display a measure of central location ((trimmed) mean, (broadened) median, or M-estimator). This estimate of central tendency should be selected in the middle part of the dialogue box (Figure A5).

-*Trended ranges* display changes in variability within phases by two lines, one connecting the minimum values of the phase halves and one connecting the maxima.

For all these methods the influence of outliers may be lessened by using a trimmed range, in which only a sample of the data set is used. This can be selected in box B (Figure 5): default the whole dataset is used, but by checking the box the 10 to 20 percent extreme values are removed from each phase.

- **Plot estimate of trend:** This function visualizes systematic shifts in central location over time using several methods.

  The method of trend visualization should be selected (Figure A6):

  -A *vertical line plot* draws the deviations from each data point to a measure of central tendency against time. The measure

of central tendency should be selected in the middle part of the dialogue box (Figure A6).

-*Trend lines* superimpose a linear function on the raw data, which shows if there is an increase or a decrease in the observed behavior over time.

-> *least squares regression*: minimizes the squared vertical distances between the regression line and the data points.

-> *split-middle method*: connects the crossings of the median dependent variable value and the middle time value of both phase halves

-> *resistant trend line fitting*: comparable to the split-middle method, but here the phases are divided into three sections instead of two. This makes this method more suited for longer time series.

-The presence of a nonlinear trend can be displayed with *running medians*, with which the time series is smoothed by dividing it into successive segments of a given size and calculating the median for each segment (Tukey, 1977). Three sizes of segments are easy to use with time series data: *batch size 3, batch size 5, and batch size 4 averaged by pairs*.

## SCRT (Single-Case Randomization Tests)

- **Design your experiment**

  **-Number of possible assignments:** the number of assignment possibilities for the specified design is calculated.

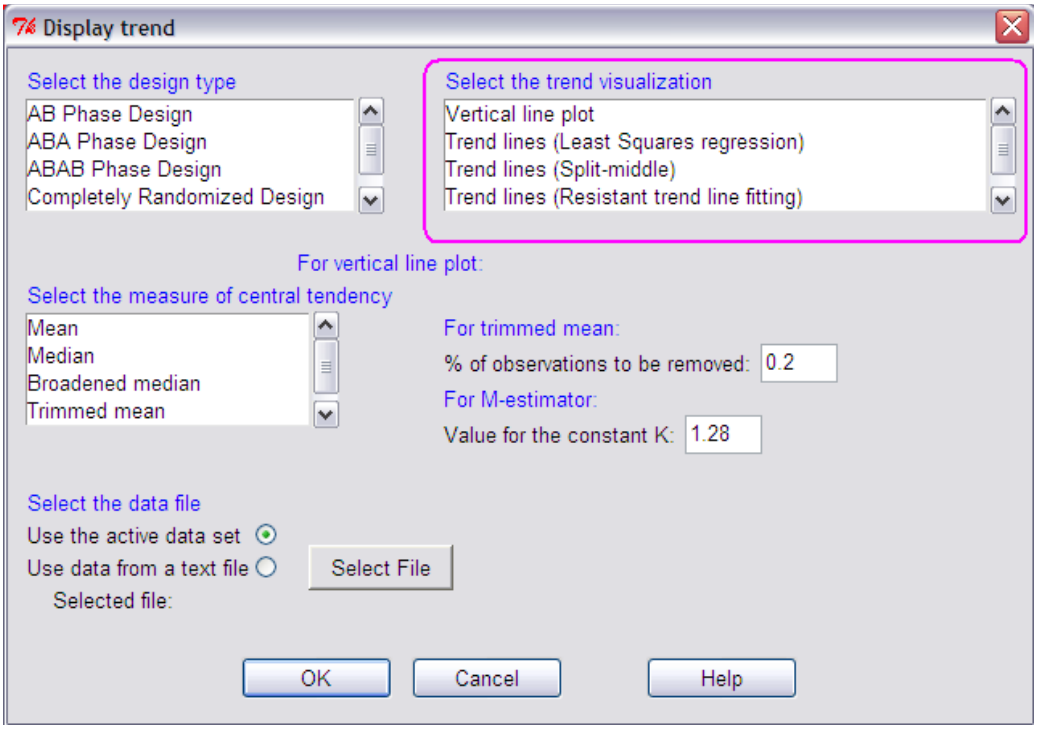



*Figure A6.* The dialogue box for the 'Plot estimate of trend' menu.

In box A (Figure A7) the total number of observations in the experiments should be indicated. This is not necessary when the design used is a multiple baseline design. In box B a constraint on the randomization schedule should be provided. In phase designs the moment of phase change is randomly determined, so the restriction is placed on the minimum number of observations per phase. In alternation designs the observations are randomly assigns to different conditions, so in other words the treatment order is randomly determined. To avoid too long sequences of the same condition, in alternating treatment designs there is a restriction on the maximum number of consecutive administrations of the same condition. For multiple baseline designs, there is one extra selection to be made: in box C the

location of the file with the possible start points should be selected (see 'Data input').

**-Display all possible assignments:** all assignment possibilities for the specified design are enumerated. In the dialogue box (Figure A8) can be chosen if the possibilities should only be displayed in the output window of the R commander, or also should be saved to a file. The location to save the assignment possibilities has to be indicated with the 'Select location' button.

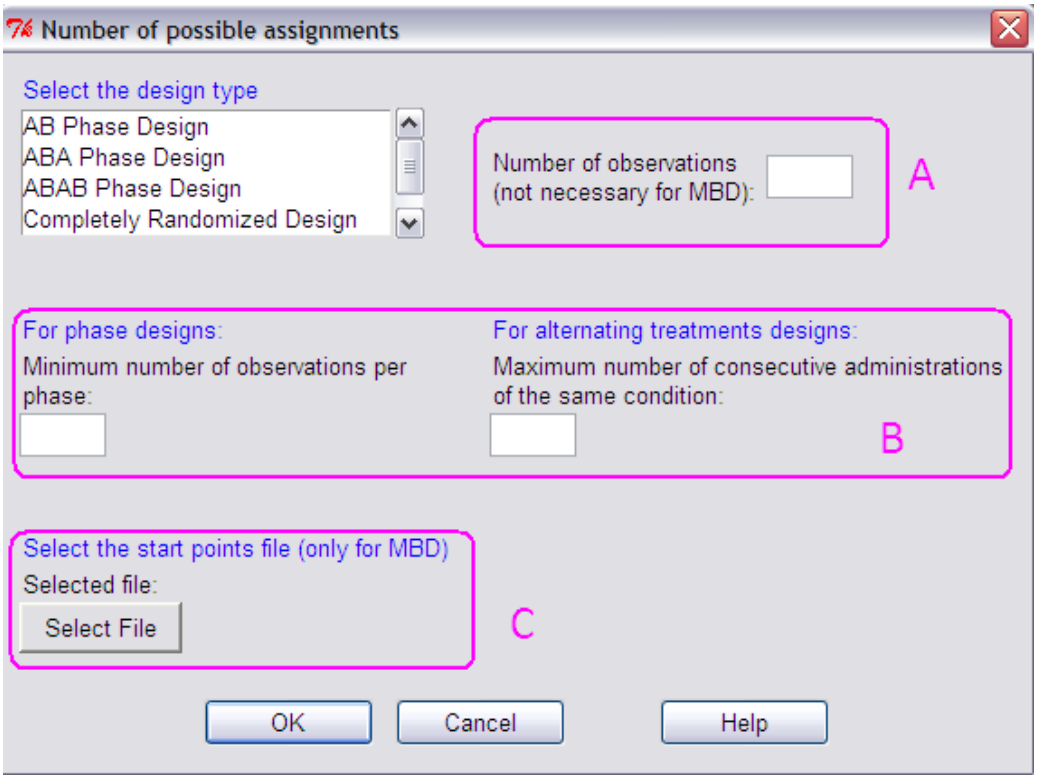



*Figure A7.* The dialogue box for the 'Number of possible assignments' menu.

**-Choose 1 possible assignment:** one assignment possibility is randomly selected from all theoretical possibilities. The dialogue box is similar to that of the 'Number of possible assignments' function.

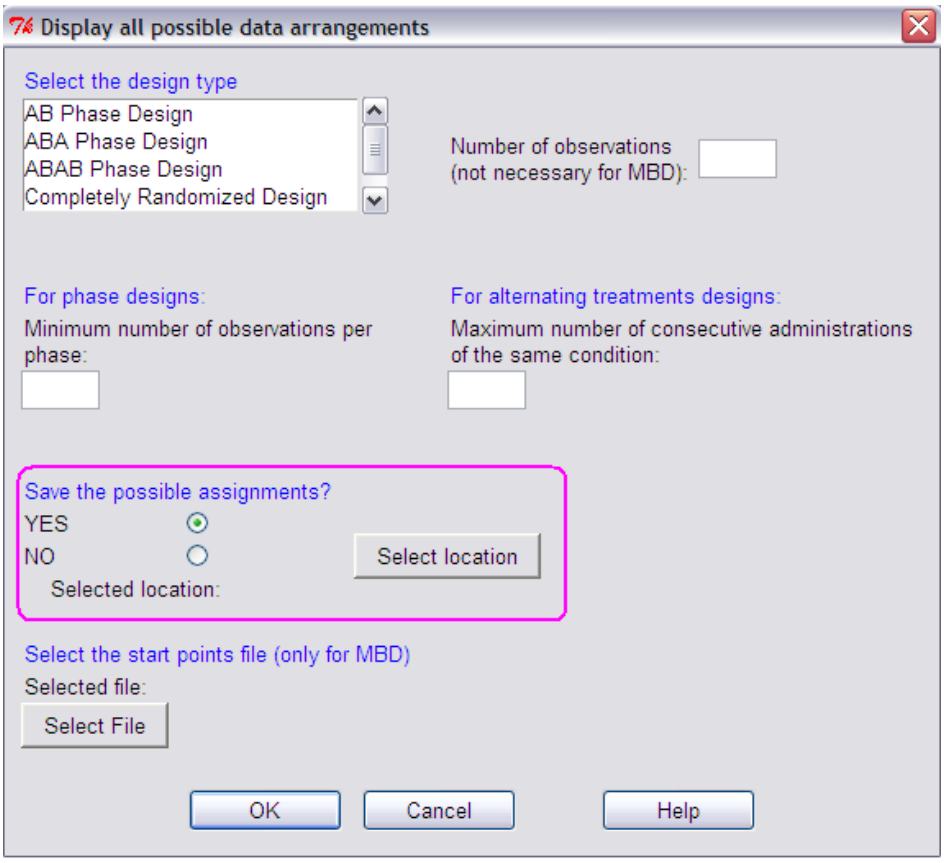



*Figure A8.* The dialogue box for the 'Display all possible assignments' menu.

- **Analyze your data**

  **-Observed test statistic:** the observed test statistic is calculated from the obtained raw data.

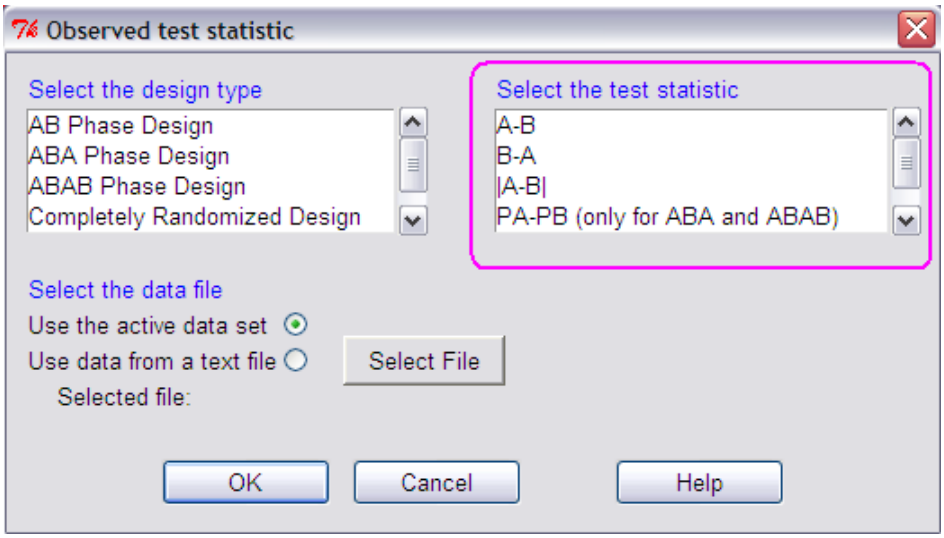



*Figure A9.* The dialogue box for the 'Observed test statistic' menu.

There are several built-in possibilities as test statistic (see Figure A9). For alternation designs, multiple-baseline designs and AB phase designs, there are three options: *"A-B", "B-A",* and *"|A-B|"*, which stand for the (absolute value of the) difference between the condition/phase 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 phase means, and *"AA-BB", "BB-AA",* and *"|AA-BB|"* represent the (absolute value of the) difference between the sums of phase means.

**-Randomization distribution:** the randomization distribution is generated.

One can choose between the exhaustive ('systematic') and the nonexhaustive ('Monte Carlo') randomization distribution (Figure A10 box A). For the exhaustive randomization distribution all assignment possibilities are enumerated, while the nonexhaustive randomization distribution is generated by a random sample of all assignment possibilities. The size of this random sample should be indicated in the box 'number of randomizations'. In box B (Figure A10) can be chosen if the randomization distribution should only be displayed in the output window of the R commander, or also should be saved to a file. The location to save the randomization distribution has to be indicated with the 'Select location' button.

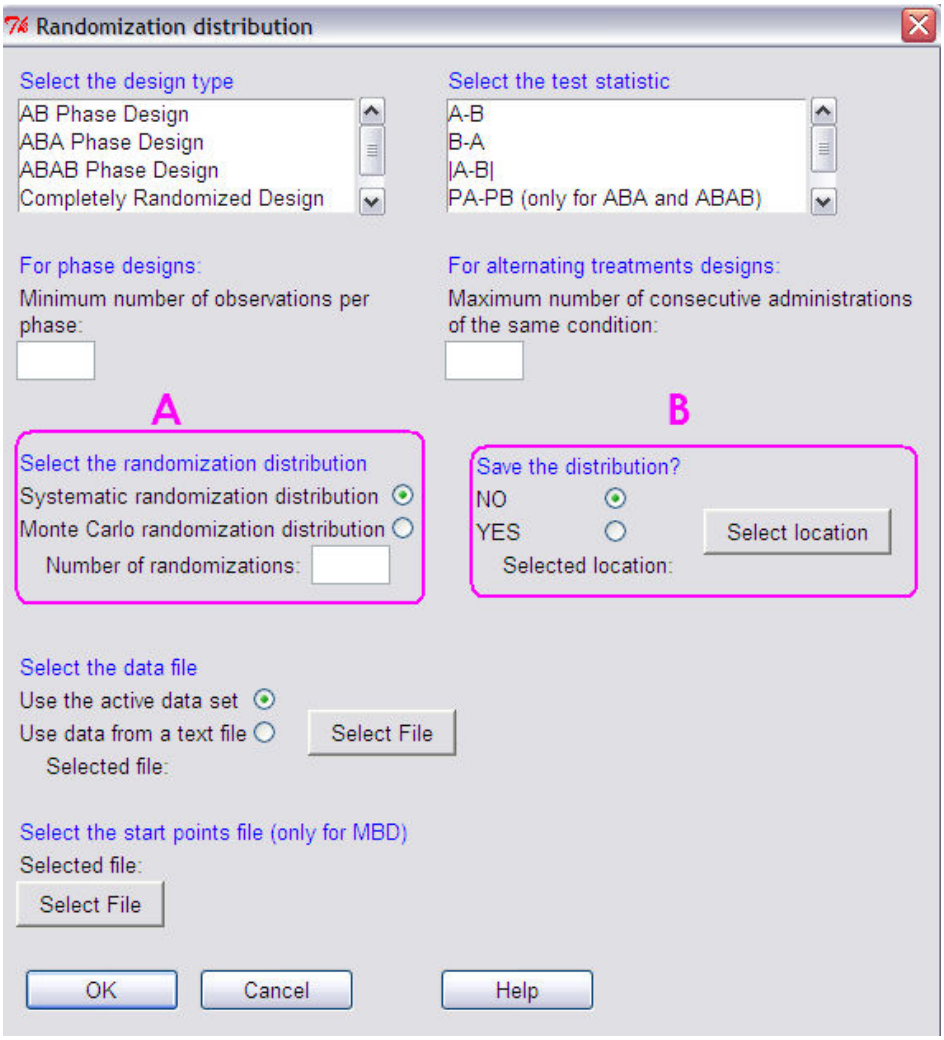



*Figure A10.* The dialogue box for the 'Randomization distribution' menu.

**-P-value:** the *p*-value corresponding to the observed value of the test statistic is obtained by locating this value in the exhaustive or nonexhaustive randomization distribution. The dialogue box is similar to that of the 'randomization distribution' function.

## SCMA (Single-Case Meta-Analysis)

- **Calculate effect size:** the specified effect size measure is calculated (Figure A11). Four effect size measures are included in the RcmdrPlugin.SCDA:

  -*standardized mean difference*: the mean of the baseline condition is subtracted from the mean of the treatment condition, and this difference is divided by the baseline standard deviation.

  -*pooled standardized mean difference*: similar to the standardized mean difference, but with the pooled standard deviation used as denominator

  -*PND* (percentage of nonoverlapping data): the percentage of data points in the treatment phase that exceed the most extreme value in the baseline phase (i.e., the proportion lower than the lowest baseline point for interventions designed to decrease unwanted behavior, and the proportion higher than the highest baseline point for interventions intended to increase desired behavior).

  -*PEM (*percentage of data points exceeding the median): the overlap is calculated as the percentage of data points in the treatment phase that are located above the extended median line of the previous baseline phase (or below this line, if the undesired behavior is expected to decrease).

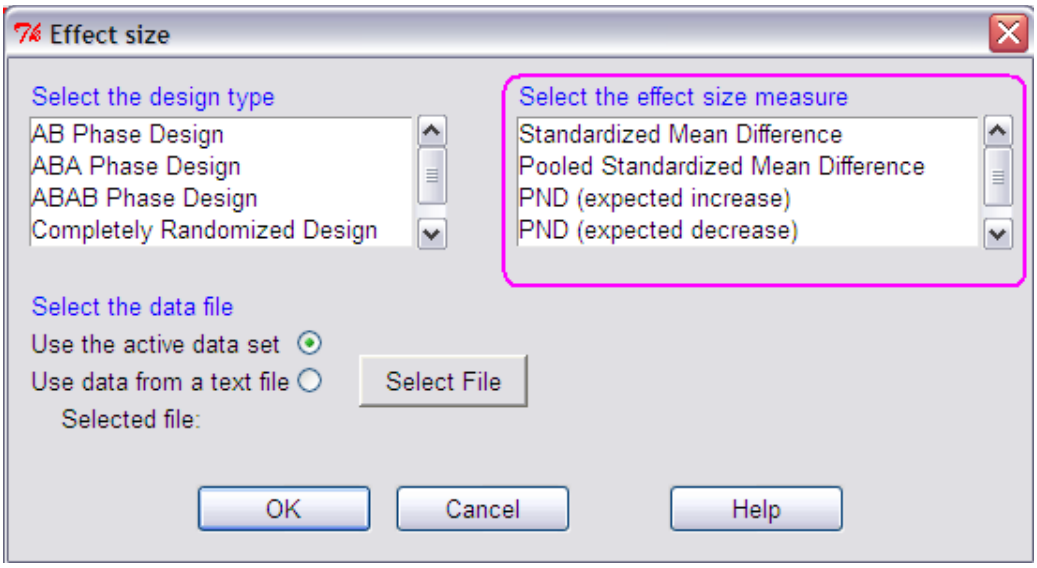



*Figure A11.* The dialogue box for the 'Calculate effect size' menu.

- **Combine *p*-values:** nonparametric combination of *p*-values, with the multiplicative approach using Pearson's formula or the additive approach using Edgington's formula (Figure A12 box A). In box B (Figure A12) the location of the file in which the *p*-values to b combined can be found should be selected.

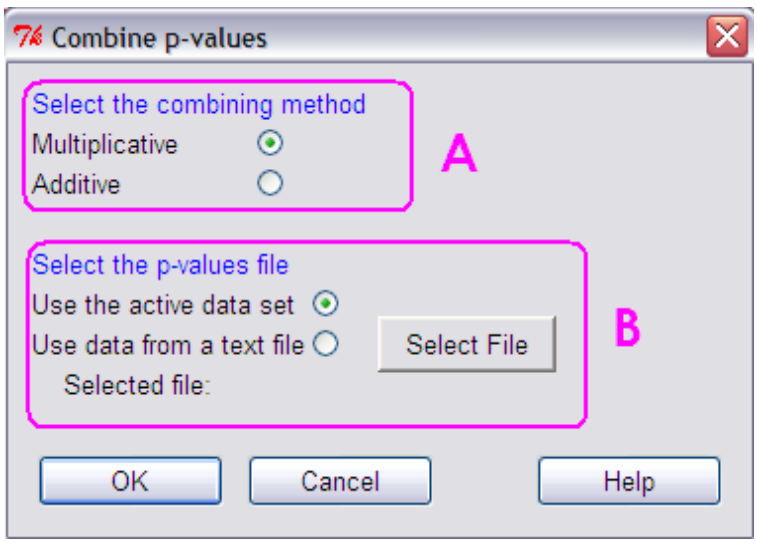



Figure A12: The dialogue box for the 'Combine *p*-values' menu.

# General discussion

Although in many instances group studies are the best choice, sometimes single-case experiments should be preferred. This could be the case when it is unnecessary to collect data from a large group of subjects (e.g., generating pilot data) or when this is impossible (e.g., studying rare conditions). Researchers are encouraged to choose the design that best fits their research question. This also holds when selecting the technique to analyze the data. Visual analysis is an important and necessary first step, but is not always sufficient to come to reliable conclusions. When the effect is not very obvious, statistical tests might be necessary. Which statistical test to use depends, among other things, on the nature of the data and the type of research design chosen.

## Critical appraisal of the research project

With the development of a software package for analyzing single-case data, we contributed to the availability and ease of use of several data analytical tools. This will hopefully encourage applied researchers to visually analyze their data in a systematical and structured manner, and, when designated, to also use statistical tests and calculate effect size measures to answer their research questions. A particular strength of this software package is that it already gathers various techniques for evaluating single-case data, so that researchers can turn to one product instead of using different programs. Furthermore, the packages can be attained at no cost from

the CRAN website (cran.r-project.org), which makes them also attractive for teaching purposes. Of course, the SCDA package is not yet comprehensive and developers are invited to expand the functionality.

*First*, adding extra options within the existing functions would be useful. The addition of more design options is needed, for example, adding a third (C) or a fourth (D) phase or condition, allow for a different phase order, include other design types, add the possibility of having a different minimum number of observations over the phases, … Also more possible test statistics should be appended, with the possibility to test for delayed effect, and including the possibility to devise one's own test statistic. This flexibility could also be built into the graphical possibilities of the software, allowing users to choose the axis labels and phase/condition names without having to alter the package's source code.

*Second*, several extensions of the existing functions can be considered. For example, other effect size measures for single-case data could be included. These could comprise more visual based approaches, such as the extended celeration line (ECL; White & Haring, 1980); the percentage of zero data (PZD; Scotti, Evans, Meyer, & Walker, 1991); the mean baseline reduction (MBR; Campbell, 2004); the percentage of all non-overlapping data (PAND; Parker, Hagan-Burke, & Vannest, 2007); the nonoverlap of all pairs (NAP; Parker & Vannest, 2009); a modification of the PND index (PNCD; Manolov & Solanas, 2009); the improvement rate difference (IRD; Parker, Vannest, & Brown, 2009); or the Tau-U (Parker, Vannest,

Davis, & Sauber, 2011). Also regression approaches to calculating effect sizes could be considered, including Gorsuch's (1983) trend analysis, the model presented by Center et al. (1985-1986), White et al.'s (1989) approach with the adjustment proposed by Faith et al. (1997), and Allison and Gorman's (1993) model. The software could also be extended with methods for combining effect sizes. The easiest method is probably taking the weighted average of the effect sizes to be combined. More refined are the three techniques proposed by Busk and Serlin (1992), differing in the assumptions made about the distribution and the variability of the data. And even more sophisticated is the suggestion of Van den Noortgate and Onghena (2003b) of using a hierarchical linear model for combining effect sizes of single-case data. By modeling the hierarchical structure, the possible dependence of the scores is taken into account. Furthermore, more methods for combining *p*-values would be useful. Possibilities are the binomial test (Rosenthal, 1978); Stouffer's method and Mosteller and Bush's and Winer's modifications to this approach (Rosenthal, 1978); different combination methods described in Pesarin (2001) and Pesarin and Salmaso (2010), for example Liptak, Tippett and Lancaster combining procedures; and Manolov and Solanas' (2012) proposal of taking the weighted combination of *p*-values. Also other statistical tests for analyzing single-case data should be added. It is not because randomization tests are valid without making distributional assumptions, that they are the best solution in every situation. Other plausible options are multilevel models (Van den Noortgate & Onghena, 2003a, 2003b) and time series analysis (Box et al., 1994).

*Third*, adding new tools could expand the functionality. An interesting path is the development of a tool to allow for power calculations. For group designs such software is already available (e.g., G*power, Optimal design), but for single-case designs this is not the case. It is, however, of great importance for example to determine a priori the number of observations necessary to obtain enough power to be able to detect significant effects. Another useful expansion would be to provide a tool for handling missing data. Most study designs are subjected to the problem of dropout, but with repeated observations the risk of missing data becomes even higher. This cannot always be reduced to intentional noncompliance; also forgetfulness and not having access to the data collection tools may play a part. Although several methods of managing noncompliance can be used (training, incentives, reminders, ...) missing data are often unavoidable (Smith, 2012). Several missing data procedures could be included (e.g., Smith et al., 2012; Velicer & Colby, 2005).

The beauty of R and the R commander is that this new code does not necessarily has to be included in the same (plug-in) package; the environment for analyzing single-case data can be extended by developing additional packages and loading them into R (or the R commander). For the slope and level change effect size index (SLC), for example, already an R package and a plug-in for the R commander exist (Solanas, Manolov, & Onghena, 2010).

# Perspectives for the future

Single-case experimental designs are clearly on the rise. This is not only obvious from several special journal issues in applied research areas (e.g., Remedial and Special Education (2013); the Journal of Applied Sports Psychology (2013)) and a proliferation in published empirical studies. Also stochastic simulation studies are conducted with focus on techniques for analyzing single-case data. However, additional research is definitely needed. The R scripts developed in this research project could be used for this purpose.

## Empirical research

In the applied field of exposure in vivo for pain patients, for example, already several single-case experimental studies were conducted. Exposure in vivo is one of the most effective treatments for people with anxiety disorders, and recently an adjusted program was applied with success in patients with elevated levels of fear of pain. Single-case experiments were used in the applied field of exposure in vivo, because of the suitability of this context for this particular research methodology.

We already used the R scripts presented in dissertation in three empirical studies. The first study concerned patients with acute *post-traumatic neck pain*, who suffered from substantial pain-related fear (de Jong et al., 2008). Using a replicated single-case phase design, with 8 sequentially treated patients reporting daily on 14 items with visual analog scales, it was shown that graded exposure in vivo results in decreased levels of pain intensity, pain-related fear, and

disability, and in an improvement in physical activity. In a second study (de Jong et al., submitted), a similar diary was filled in to check whether exposure indeed modified fear appraisals and activity goal achievement. Here, a replicated single-case experimental design was used with patients with *complex regional pain syndrome*, to test the generalization of graded exposure in vivo to threatening activities that were not addressed during the treatment. The third study investigated the effectiveness of therapist-aided exposure in the treatment of *primary vaginismus* (ter Kuile et al., 2009, 2010), by means of a replicated single-case phase design. Ten women daily indicated their intercourse ability for 24 weeks. Results demonstrated that 9 out of the 10 patients were able to have intercourse after the treatment. In these three studies, nonparametric randomization tests were used to analyze at least part of the data. The results from different patients, coming from sequentially replicated studies, were gathered using the meta-analytical technique of combining *p*-values.

In interesting track for future research is conducting a meta-analysis on the effectiveness of exposure in vivo for the reduction of pain-related fear and disability in chronic pain. By reviewing all (replicated) single-case studies used so far in this research domain, the current state of affairs can be presented. Aggregating findings from several single-case studies serves several purposes: it helps strengthening conclusions about the impact of the intervention, it could help identifying sources of variation in outcome and as such unravel the conditions and level of generalizability, and it can increase the power of statistical tests (Gingerich, 1984).

## Stochastic simulation studies

To increase the popularity of (meta-analyses of) single-case experiments with applied researchers, it is important that consensus is reached on which data-analytical techniques are most suited under which circumstances. Therefore, a thorough examination of the statistical properties of the suggested effect size measures and statistical techniques for (meta-) analyzing data from single-case experiments is required. An in depth investigation of those techniques could be accomplished by stochastic simulation studies, which allow to derive more general statements. Simulation studies allow to control the population values and to assess the validity, accuracy and power of the statistical procedures in realistic but generic situations.

In the course of this research project we contributed in three stochastic simulation studies. The first study investigated the statistical properties of randomization tests for ABAB phase designs, in the special case that the desirable randomization is not possible (Manolov, Solanas, Bulté, & Onghena, 2010). Robustness and power were studied under several experimental conditions (varying levels of autocorrelation, effect sizes and phase lengths). It was shown that the presence of autocorrelation almost always distorted Type I error rates. Satisfactory Type II error rates were only obtained for large treatment effects. In the second simulation study, a permutation solution is suggested to analyze data from alternation single-case experiments (Solmi, Onghena, Salmaso, & Bulté, in press-a). The aim was to improve the existing proposals, by using a longitudinal data

analysis tool for the test statistic. The time functional shape of the observations is first estimated with a nonparametric smoother or a parametric time-series analysis tool, and then a summary statistic of the resulting curve is used, which is able to measure the difference between the smoothed time-processes. The reliability and the power of the proposed procedure was demonstrated with a Monte Carlo simulation study. In the third study, an extension to this permutation procedure is presented (Solmi, Onghena, Salmaso, & Bulté, in press-b). To special cases of multiplicity are considered: multivariate responses and replicated single-case alternation experiments. The procedure uses the nonparametric combination of dependent permutation tests. The approach also proves to be reliable under the null hypothesis and powerful under the alternative hypothesis.

Of course, more research is needed concerning techniques to analyze and meta-analyze single-case data sets in order to derive more general guidelines. Of all the research lines that would be interesting on this topic, three suggestions concerning the meta-analysis of single-case data are further described. First, no consensus has yet been reached concerning the use of effect-size measures in single-case studies. This has as a consequence that, in the rare cases that effect size estimates are reported, there is no consistency in the measures that are used, which is problematic for the comparability of the results across different studies and the combining of effect sizes over studies. Based on the importance of having an empirically guided selection of metrics, it would be interesting to contrast the suggested measures to study the conditions under which each

method is optimal. Their performance could be assessed under differing conditions, to determine the sensitivity to the level of serial dependence, the effect of data pattern on the detection of effects, and the effect of series length on the differentiation between data patterns. Second, the additive and the multiplicative approach for the combination of (randomization test's) *p*-values are considered as complementary, with the former method being more powerful for homogeneous *p*-values, and the latter method being more powerful for heterogeneous *p*-values (Edgington, 1972a; Edgington & Haller, 1984). This statement was however not proven, but merely illustrated with a few numerical examples. It would be interesting to test this hypothesis of differences in sensitivity by means of a stochastic simulation study. And third, the statistical power and robustness of the nonparametric and parametric meta-analytical approaches should be compared. The multilevel approach, suggested by Van den Noortgate and Onghena (2003a, 2003b), could be selected as the parametric technique to contrast with the nonparametric ones. Several parameters could be manipulated, such as the degree of serial dependence, the effect size, and the number of cases to be combined. Results could be compared for the case of homogeneous and heterogeneous effect sizes. See for example Ugille, Moeyaert, Beretvas, Ferron, and Van den Noortgate (2012) and Moeyaert, Ugille, Ferron, Beretvas, and Van den Noortgate (in press).

## To conclude

With this research, our aim was to contribute to making researchers more familiar with analytic methods for single-case experimental designs, and to provide them with a user-friendly software package. Although the development of a graphical user interface that combines analytical tools for single-case data has the advantages of being easily available and user-friendly, it is not without risks. By providing a window based software environment, statistical testing could be reduced to plain clicking in menus and dialogue boxes without any statistical background knowledge. This is of course not the purpose of data analysis and should be prevented. We tried to do this by describing the different packages in several manuscripts, and by providing extensive help files. Within the given timeframe of the project, we however did not solve the issue of missing straightforward guidelines concerning which technique to use under which circumstances, but tried to give some directions for future research concerning this topic. Finally, researchers should also bear in mind that the results of data analysis, in the form of *p*-values and measures of effect size do not tell the whole story. Especially in clinical settings, the clinical relevance of the effects should always be considered.

# 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.

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single-case. *Behavior Research and Therapy, 31,* 621-631.

Allison, D. B., Franklin, R. D., & Heshka, S. (1992). Reflections on visual inspection, response guided experimentation, and Type I error rate in single-case designs. The Journal of Experimental Education, 61, 45–51.

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.

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.

Bailey, D. B. Jr. (1984). Effects of lines of progress and semilogarithmic charts on ratings of charted data. Journal of Applied Behavior Analysis, 17, 359–365.

Baplu, A. (2005). *Rilatine, the smart drug. Een dubbelblind placebogecontroleerd N* 5 *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.

Barlow, D. H., Hayes, S. C., & Nelson, R. O. (1984). *The scientist practitioner: Research and accountability in clinical and educational settings.* New York: Pergamon.

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

Beretvas, S. N., & Chung, H. (2008). A review of meta-analyses of single-subject experimental designs: Methodological issues and practice. *Evidence-Based Communication Assessment and Intervention, 2,* 129-141.

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

Bliss, S. L., Skinner, C. H., Hautau, B., & Carroll, E. E. (2008). Articles published in four school psychology journals from 2000 to 2005: An analysis of experimental/intervention research. *Psychology in the Schools, 45,* 483-498.

Bobrovitz, C. D., & Ottenbacher, K. J. (1998). Comparison of visual inspection and statistical analysis of single-subject data in rehabilitation research. American Journal of Physical Medicine & Rehabilitation, 77, 94–102.

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

Brossart, D. F., Parker, R. I., Olsen, E. A., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification, 30,* 531-563.

Brossart, D. F., Parker, R. I., Olson, E. A., & Lakshmi, M. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification, 30*, 531–563.

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

Bulté, I., & Onghena, P. (2009). Randomization tests for multiple baseline designs: An extension of the SCRT-R package. *Behavior Research Methods, 41,* 477-485.

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

Bulté, I., & Onghena, P. (submitted). The single-case data analysis package: Analyzing single-case experiments with the R software.

Bulté, I., Van den Noortgate, W., Heyvaert, M., & Onghena, P. (submitted). Meta-analysis of single-case studies: The SCMA package. Manuscript submitted for publication.

Busk, P. L., & Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counterargument to the myth of no autocorrelation. *Behavioral Assessment, 10,* 229-242.

Busk, P. L., & Marascuilo, L. A. (1992). Statistical analysis in single-case research: Issues, procedures, and recommendations, with special applications to multiple behaviors. In T. R. Kratochwill &

J. R. Levin (Eds.), Single-case research designs and analysis: New directions for psychology and education (pp. 159–185). Hillsdale, NJ: Erlbaum.

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 187-212). Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

Campbell, J. M. (2004). Statistical comparison of four effect sizes for single-subject designs. *Behavior Modification, 28*, 234-246.

Campbell, J. M., & Herzinger, C. V. (2010). Statistics and single subject research methodology. In D. L. Gast (Ed.), Single subject research methodology in behavioral sciences (pp. 417–453). New York, NY: Routledge.

Carr, J. E., & Burkholder, E. O. (1998). Creating single-subject design graphs with Microsoft Excel™. *Journal of Applied Behavior Analysis, 31,* 245-251.

Carter, M. (2009). Effects of graphing conventions and response options on interpretation of small n graphs. Educational Psychology, 29, 643–658.

Center, B. A., Skiba, R. J., & Casey, A. (1985-1986). A methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education, 19,* 387-400.

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.

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

Crawley, M. J. (2005). *Statistics: An introduction using R*. London: John Wiley and Sons.

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.

de Jong, J. R., den Hollander, M. L., Bulté, I., Ruijgrok, J., Onghena, P., & Vlaeyen, J. W. S. (submitted). Generalisation of graded exposure in vivo in complex regional pain syndrome type I (CPRS-I). Attenuate renewal in pain-related fear learning. Manuscript submitted to *Pain*.

de Jong, J. R., Vangronsveld, K., Peters, M., Goossens, M. E. J. B., Bulté, I., Onghena, P., & Vlaeyen, J. W. S. (2008). Reduction of pain-related fear and disability in posttraumatic neck pain: A replicated single case experimental study of exposure in vivo. *Journal of Pain, 9,* 1123-1134.

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analyses of intrasubject data. Journal of Applied Behavior Analysis,12, 573–579.

Dugard, P., File, P., & Todman, J. (2012). *Single-case and small-n experimental designs: A practical guide to randomization tests* (2nd ed.). New York: Routledge Taylor & Francis Group.

Edgington, E. S. (1972a). An additive method for combining probability values from independent experiments. *Journal of Psychology, 80,* 351-363.

Edgington, E. S. (1972b). A normal curve method for combining probability values from independent experiments. *Journal of Psychology, 82,* 85-89.

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 and Applied, 90*, 57-68.

Edgington, E. S. (1980a). Overcoming obstacles to single-subject experimentation. *Journal of Educational Statistics, 5,* 261-267.

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

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. (1996). Randomized single-subject experimental designs. *Behaviour Research and 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., & Haller, O. (1984). Combining probabilities from discrete probability-distributions. *Educational and Psychological Measurement, 44,* 265-274.

Edgington, E. S., & Onghena, P. (2007). *Randomization tests* (4th ed.). London: Chapman & Hall/CRC.

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist, 33,* 517. doi:10.1037/0003-066X.33.5.517.a

Eysenck, H. J. (1994). Systematic reviews: Meta-analysis and its problems. *British Medical Journal, 309,* 789-792.

Faith, M. S., Allison, D. B., & Gorman, B. S. (1997). Meta-analysis of single-case research. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 119-158). Mahwah, NJ: Lawrence Erlbaum Associates.

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. *The Journal of Experimental Education, 75*, 66-81.

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

Ferron, J., & Scott, H. (2005). Multiple baseline designs. In B.S. Everitt & D.C. Howell (Eds.), *Encyclopedia of statistics in behavioural science* (Vol. 3) (pp. 1306-1309). Chichester: John Wiley & Sons.

Ferron, J., & Sentovich, C. (2002). Statistical power of randomization tests used with multiple-baseline designs. *The 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.

Fisch, G. S. (1998). Visual inspection of data revisited: Do the eyes still have it? The Behavior Analyst, 21, 111–123.

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

Fox, J. (2005). The R Commander: A basic-statistics graphical user interface to R. *Journal of Statistical Software, 14.*

Fox, J. (2007). Extending the R commander by "plug-in" packages. *R News*, 7(3), 46-52. URL http://CRAN.R-project.org/doc/Rnews/.

Franklin, R. D., Allison, D. B., & Gorman, B. S. (Eds.). (1997). *Design and analysis of single-case research.* Mahwah, NJ: Lawrence Erlbaum Associates.

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: Lawrence Erlbaum Associates.

Furlong, M. J., & Wampold, B. E. (1982). Intervention effects and relative variation as dimensions in experts' use of visual inference. *Journal of Applied Behavior Analysis, 15,* 415-421.

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.

Gibson, G., & Ottenbacher, K. (1988). Characteristics influencing the visual analysis of single-subject data: An empirical analysis. The Journal of Applied Behavioral Science, 24, 298–314.

Gingerich, W. J. (1984). Meta-analysis of applied time series data. *Journal of Applied Behavioral Science, 20,* 71-79.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3-8.

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: Lawrence Erlbaum Associates.

Gorsuch, R. L. (1983). Three methods for analyzing limited time-series (N of 1) data. *Behavioral Assessment, 5,* 141-154.

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.

Hammond, D., & Gast, D. L. (2010). Descriptive analysis of single-subject research designs: 1983-2007. *Education and Training in Autism and Developmental Disabilities, 45,* 187-202.

Harbst, K. B., Ottenbacher, K. J., & Harris, R. S. (1991). Interrater reliability of therapists' judgments of graphed data. Physical Therapy, 71, 107–115.

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.

Hedges, L. V. (1992). Meta-analysis. *Journal of Educational Statistics, 17,* 279-296.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York: Academic Press.

Hojem, M. A., & Ottenbacher, K. J. (1988). Empirical investigation of visual-inspection versus trend-line analysis of single-subject data. Physical Therapy, 68, 983–988.

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

Horner, R. D., & Baer, D. M. (1978). Multiple-probe technique: A variation of the multiple baseline. *Journal of Applied Behavior Analysis, 11,* 189-196.

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

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

Hornik, K. (2010). The R FAQ*: Frequently asked questions on R.* Retrieved November 12, 2010, from CRAN.R-project. org/doc/FAQ/

Hornik, K. (2012). *The R FAQ: Frequently asked questions on R.* Retrieved February 17, 2013 from cran.r-project.org/doc/FAQ/R-FAQ.html.

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

Huber, P., & Ronchetti, E. (2009). *Robust statistics* (2nd ed.). Hoboken, NJ: John Wiley and Sons.

Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment, 7,* 107-118.

Jeng, H. (2010). On small samples and the use of robust estimators in loss reserving. Casualty Actuarial Society E-Forum, 1, 1–27.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (2nd ed.). New York: Wiley.

Jones, L. V., & Fiske, D. W. (1953). Models for testing the significance of combined results. *Psychological Bulletin, 50,* 375-382.

Jones, R.J., Weinrott, M.R., & Vaught, R.S. (1978). Effects of serial dependency on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis, 11,* 277-283.

Kahng, S. W., Chung, K.-M., Gutshall, K., Pitts, S. C., Girolami, J. K., & Girolami, K. (2010). Consistent visual analyses of intrasubject data. Journal of Applied Behavior Analysis, 43, 35–45.

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. (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.

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York: Oxford University Press.

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.

Knapp, T. J. (1983). Behavior analysts' visual appraisal of behavior change in graphic display. Behavioral Assessment, 5, 155–164.

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.

Kratochwill, T. R., & Brody, G. H. (1978). Single subject designs: A perspective on the controversy over employing statistical inference and implications for research and training in behavior modification. Behavior Modification, 2, 291–307.

Kratochwill, T. R., & Levin, J. R. (1992). *Single-case research designs and analysis: New directions for psychology and education.* Mahwah, NJ: Erlbaum.

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods, 15,* 124-144.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M & Shadish, W. R. (2010). Single-*case designs technical documentation*. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf.

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 and Psychotherapy, 2*, 177-191.

Loughin, T. M. (2004). A systematic comparison of methods for combining p-values from independent tests. *Computational Statistics and Data Analysis, 47,* 467-485.

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* 5 1) experimental designs. *Journal of Counseling & Development*, 78, 92-102.

Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject research: Percentage of data points exceeding the median. Behavior Modification, 30, 598–617.

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

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

Manolov, R., & Solanas, A. (2008). Comparing N = 1 effect size indices in presence of autocorrelation. *Behavior Modification, 32,* 860-875.

Manolov, R., & Solanas, A. (2009). Percentage of nonoverlapping corrected data. *Behavior Research Methods, 41,* 1262-1271.

Manolov, R., & Solanas, A. (2012). Assigning and combining probabilities in single-case studies. *Psychological Methods, 17,* 495-509.

Manolov, R., Solanas, A., Bulté, I., & Onghena, P. (2010). Data-division-specific robustness and power of randomization tests for ABAB designs. *Journal of Experimental Education,78,* 191-214.

Manolov, R., Solanas, A., & Leiva, D. (2010). Comparing "visual'' effect size indices for single-case designs. *Methodology, 6,* 49-58.

Manolov, R., Solanas, A., Sierra, V., & Evans, J. J. (2011). Choosing among techniques for quantifying single-case intervention effectiveness. *Behavior Therapy, 42,* 533-545.

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

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.

Matson, J. L., Turygin, N. C., Beighley, J., & Matson, M. L. (2012). Status of single-case research designs for evidence-based practice. *Research in Autism Spectrum Disorders, 6,* 931-938.

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 Behavior Analysis, 23,* 341-351.

Matyas, T. A., & Greenwood, K. M. (1991). Problems in the estimation of autocorrelation in brief time series and some implications for behavioral data. *Behavioral Assessment, 13,* 137-157.

Matyas, T. A., & Greenwood, K. M. (1997). Serial dependency in single-case time-series. In R.D. Franklin, D.B. Allison, & B.S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 215-244). Mahwah, NJ: Lawrence Erlbaum Associates.

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S. N., & Van den Noortgate, W. (in press). Modeling external events in the three-level analysis of multiple-baseline across-participants designs: A simulation study. *Behavior Research Methods.*

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

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

Mosteller, F., Siegel, A. F., Trapido, E., & Youtz, C. (1981). Eye fitting straight lines. The American Statistician, 35, 150–152.

Mullen, B., & Rosenthal, R. (1985). *BASIC Meta-analysis: Procedures and programs*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

Normand, M. P., & Bailey, J. S. (2006). The effects of celeration lines on visual data analysis. Behavior Modification, 30, 295–314.

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. Everitt & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science: Vol. 4* (pp. 1850-1854). New York: 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.

Ottenbacher, K. J. (1990). When is a picture worth a thousand p values? A comparison of visual and quantitative methods to analyze single subject data. Journal of Special Education, 23, 436–449.

Paramore, N. W., & Higbee, T. S. (2005). An evaluation of a brief multiple-stimulus preference assessment with adolescents with emotional-behavioral disorders in an educational setting. *Journal of Applied Behavior Analysis, 38,* 399-403.

Park, H. S., Marascuilo, L., & Gaylord-Ross, R. (1990). Visual inspection and statistical analysis in single-case designs. The Journal of Experimental Education, 58, 311–320.

Parker, R. I., & Brossart, D. F. (2003). Evaluating single-case research data: A comparison of seven statistical methods. *Behavior Therapy, 34,* 189-211.

Parker, R. I., Brossart, D. F., Vannest, K. J., Long, J. R., Garcia De-Alba, R., Baugh, F. G., & Sullivan, J. R. (2005). Effect sizes in single case research: How large is large? *School Psychology Review, 34,* 116–132.

Parker, R. I., & Hagan-Burke, S. (2007). Useful effect size interpretations for single case research. *Behavior Therapy, 38,* 95-105.

Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data (PAND): An alternative to PND. *The Journal of Special Education, 40,* 194-204.

Parker, R. I., & Vannest, K. J. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy, 40,* 357-367.

Parker, R. I., Vannest, K. J., & Brown, L. (2009). The improvement rate difference for single-case research. *Exceptional Children, 75,* 135-150.

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

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy, 42,* 284-299.

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: Lawrence Erlbaum Associates.

Pesarin, F. (2001). *Multivariate permutation tests. With applications in biostatistics.* New York: John Wiley & Sons.

Pesarin, F., & Salmaso, L. (2010). *Permutation tests for complex data: Theory, applications and software.* Chichester, UK: Wiley.

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.

Revusky, S. H. (1967). Some statistical treatments compatible with individual organism methodology. *Journal of the Experimental Analysis of Behavior, 10,* 319-330.

Richards, S. B., Taylor, R. L., & Ramasamy, R. (1997). Effects of subject and rater characteristics on the accuracy of visual analysis of single subject data. Psychology in the Schools, 34, 355–362.

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

Rojahn, J., & Schulze, H. H. (1985). The linear regression line as a judgmental aid in visual analysis of serially dependent A-B

time-series data. Journal of Psychopathology and Behavioral Assessment, 7, 191–206.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85,* 185-193.

Scotti, J. R., Evans, I. M., Meyer, L. H., & Walker, P. (1991). A meta-analysis of intervention research with problem behavior: Treatment validity and standards of practice. *American Journal of Mental Retardation, 96,* 223-256.

Scruggs, T. E., & Mastropieri, M. A. (1994). The utility of the PND statistic: A reply to Allison and Gorman. *Behavior Research and Therapy, 32,* 879-883.

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987a). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education, 8,* 24-33.

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987b). Reply to Owen White. *Remedial and Special Education, 8, 40-42*.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental design and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin Company.

Shadish, W. R., & Rindskopf, D. M. (2007). Methods for evidence-based practice: Quantitative synthesis of single-subject designs**.** *New Directions for Evaluation, 113,* 95-109.

Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental

designs. *Evidence-Based Communication Assessment and Intervention, 2,* 188-196.

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43,* 971-980.

Sideris, G. D., & Greenwood, C. R. (1997). Is human behavior autocorrelated? An empirical analysis. *Journal of Behavioral Education, 7,* 273-293.

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

Skiba, R., Deno, S., Marston, D., & Casey, A. (1989). Influence of trend estimation and subject familiarity on practitioners judgments of intervention effectiveness. Journal of Special Education, 22, 433–446.

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

Smith, J. D., Borckhardt, J. J., & Nash, M. R. (2012). Inferential precision in single-case time-series data streams: How well does the EM procedure perform when missing observations occur in autocorrelated data. *Behavior Therapy, 43,* 679-685.

Solanas, A., Manolov, R., & Onghena, P. (2010). Estimating slope and level change in N = 1 designs. *Behavior Modification, 34,* 195-218.

Solmi, F. Onghena, P., Salmaso, L., & Bulté, I. (in press-a). A permutation solution to test for treatment effects in alternation design single-case experiments. Manuscript accepted for publication in *Communication in Statistics – Simulation and Computation*.

Solmi, F., Onghena, P., Salmaso, L, & Bulté, I. (in press-b). Extensions of permutation solutions to test for treatment effects in replicated single-case alternation experiments with multivariate response. Manuscript accepted for publication in *Communication in Statistics – Simulation and Computation*.

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.

Strube, M. J. (1985). Power analysis for combining significance levels. *Psychological Bulletin, 98,* 595-599.

Strube, M. J., & Miller, R. H. (1986). Comparison of power rates for combined probability procedures – A simulation study. *Psychological Bulletin, 99,* 407-415.

Swaminathan, H., & Rogers, H.J. (2007). Statistical reform in school psychology research: A synthesis. *Psychology in the Schools, 44,* 543-549.

ter Kuile, M. M., Bulté, I., Weijenborg, P. T. M., Beekman, A., Melles, R., & Onghena, P. (2009). Therapist-aided exposure for women

with lifelong vaginismus: A replicated single-case design. *Journal of Consulting and Clinical Psychology, 77,* 149-159.

ter Kuile, M., Bulté, I., Weijenborg, P., Beekman, A., Melles, R., & Onghena, P. (2010). Begeleide exposure voor vrouwen met primair vaginisme: Een gerepliceerd single-case design. *Gezinstherapie Wereldwijd, 21*, 117-144.

Timothée. (2010, March 5). Data visualization (in R). Getting data from an image (introductory post) [Web log message]. Retrieved from http://rdataviz.wordpress.com/2010/03/05/getting-data-from-an-image-introductory-post/

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

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

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

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

Ugille, M., Moeyaert, M., Beretvas, S. N., Ferron, J., & Van den Noortgate, W. (2012). Multilevel meta-analysis of single-subject experimental designs: A simulation study. *Behavior Research Methods, 44*, 1244-1254.

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.

Velicer, W. F., & Colby, S. M. (2005). A comparison of missing-data procedures for ARIMA time-series analysis. *Educational and Psychological Measurement, 65,* 596-615.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*, 1-48.

Vlaeyen, J. W. S, Kole-Snijders, A. M., Boeren, R. G., & van Eek, H. (1995). Fear of movement/(re)injury in chronic low back pain and its relation to behavioral performance. *Pain, 62,* 363-372.

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

Wampold, B. E., & Furlong, M. J. (1981). The heuristics of visual inference. Behavioral Assessment, 3, 79–92.

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

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.

White, D. M., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P. (1989). Applications of meta-analysis in individual subject research. *Behavioral Assessment, 11,* 281-296.

White, O. R. (1987). Some comments concerning 'the quantitative synthesis of single-subject research'. *Remedial and Special Education, 8,* 34-39.

White, O. R., & Haring, N. G. (1980). *Exceptional teaching* (2nd ed.). Columbus, OH: Merrill.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). San Diego, CA: Elsevier.

Wilkinson, L. (2005). *The grammar of graphics* (2nd ed.). New York, NY: Springer.

Wilkinson, L., the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604.

Wilson, G. T., & Rachman, S. J. (1983). Meta-analysis and the evaluation of psychotherapy outcome: limitations and liabilities. *Journal of Consulting and Clinical Psychology, 51,* 54-64.

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.

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *Journal of Special Education, 44,* 18-29.

Ximenes, V. M., Manolov, R., Solanas, A., & Quera, V. (2009). Factors affecting visual inference in single-case designs. The Spanish Journal of Psychology, 12, 823–832.

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.

Zwahlen, M., Renehan, A., & Egger, M. (2008). Meta-analysis in medical research: Potentials and limitations. *Urologic Oncology – Seminars and Original Investigations*, *26*, 320-329.

# Appendix

## Appendix A - References of Articles that used a Multiple Baseline Design

1. Ahearn, W.H. (2003). Using simultaneous presentation to increase vegetable consumption in a mildly selective child with autism. *Journal of Applied Behavior Analysis, 36,* 361-365.

2. Ahearn, W.H., Kerwin, M.L.E., Eicher, P.S., Shantz, J., & Swearingin, W. (1996). An alternating treatments comparison of two intensive interventions for food refusal. *Journal of Applied Behavior Analysis, 29,* 321-332.

3. Allday, R.A., & Pakurar, K. (2007). Effects of teacher greetings on student on-task behavior. *Journal of Applied Behavior Analysis, 40,* 317-320.

4. Allen, K.D. (1998). The use of an enhanced simplified habit-reversal procedure to reduce disruptive outbursts during athletic performance. *Journal of Applied Behavior Analysis, 31,* 489-492.

5. Austin, J., Hatfield, D.B., Grindle, A.C., & Bailey, J.S. (1993). Increasing recycling in office environments – the effects of specific, informative cues. *Journal of Applied Behavior Analysis, 26,* 247-253.

6. Austin, J., Weatherly, N.L., & Gravina, N.E. (2005). Using task clarification graphic feedback and verbal feedback to increase closing-task completion in a privately owned restaurant. *Journal of Applied Behavior Analysis, 38,* 117-120.

7. Barker, M.R., Bailey, J.S., & Lee, N. (2004). The impact of verbal prompts on child safety-belt use in shopping carts. *Journal of Applied Behavior Analysis, 37,* 527-530.

8. Bayhinitz, A.K., Peterson, R.F., & Quilitch, H.R. (1994). Cooperative games – a way to modify aggressive and cooperative behaviors in young-children. *Journal of Applied Behavior Analysis, 27,* 435-446.

9. Berens, N.M., & Hayes, S.C. (2007). Arbitrarily applicable comparative relations: Experimental evidence for a relational operant. *Journal of Applied Behavior Analysis, 40,* 45-71.

10. Bicard, D.F., & Neef, N.A. (2002). Effects of strategic versus tactical instructions on adaptation to changing contingencies in children with ADHD. *Journal of Applied Behavior Analysis, 35,* 375-389.

11. Borrero, C.S.W., & Vollmer, T.R. (2006). Experimental analysis and treatment of multiply controlled problem behavior: A systematic replication and extension. *Journal of Applied Behavior Analysis, 39,* 375-379.

12. Brothers, K.J., Krantz, P.J., & McClannahan, L.E. (1994). Office paper recycling – a function of container proximity. *Journal of Applied Behavior Analysis, 27,* 153-160.

13. Bullock, C., & Normand, M.P. (2006). The effects of a high-probability instruction sequence and response-independent reinforcer delivery on child compliance. *Journal of Applied Behavior Analysis, 39,* 495-499.

14. Camarata, S. (1993). The application of naturalistic conversation training to speech production in children with speech abilities. *Journal of Applied Behavior Analysis, 26,* 173-182.

15. Charlop-Christy, M.H., Carpenter, M., Le, L., LeBlanc, L.A., & Kellet, K. (2002). Using the picture exchange communication system (PECS) with children with autism assessment of PECS acquisition, speech, social-communicative behavior, and problem behavior. *Journal of Applied Behavior Analysis, 35,* 213-231.

16. Clayton, M., Helms, B., & Simpson, C. (2006). Active prompting to decrease cell phone use and increase seat belt use while driving. *Journal of Applied Behavior Analysis, 39,* 341-349.

17. Codding, R.S., Feinberg, A.B., Dunn, E.K., & Pace, G.M. (2005). Effects of immediate performance feedback on implementation of behavior support plans. *Journal of Applied Behavior Analysis, 38,* 205-219.

18. Coleman-Martin, M.B., & Heller, K.W. (2004). Using a modified constant prompt-delay procedure to teach spelling to students with physical disabilities. *Journal of Applied Behavior Analysis, 37,* 469-480.

19. Craft, M.A., Alber, S.R., & Heward, W.L. (1998). Teaching elementary students with developmental disabilities to recruit teacher attention in a general education classroom: Effects on teacher praise and academic productivity. *Journal of Applied Behavior Analysis, 31,* 399-415.

20. Davis, C.A., Brady, M.P., Hamilton, R., McEvoy, M.A., & Williams, R.E. (1994). Effects of high-probability requests on the social interactions of young-children with severe disabilities. *Journal of Applied Behavior Analysis, 27,* 619-637.

21. Davis, C.A., & Reichle, J. (1996). Variant and invariant high-probability requests: Increasing appropriate behaviors in children with emotional-behavioral disorders. *Journal of Applied Behavior Analysis, 29,* 471-482.

22. DeQuinzio, J.A., Townsend, D.B., Sturmey, P., & Poulson, C.L. (2007). Generalized imitation of facial models by children with autism. *Journal of Applied Behavior Analysis, 40,* 755-759.

23. Drasgow, E., Halle, J.W., & Ostrosky, M.M. (1998). Effects of differential reinforcement on the generalization of a replacement mand in three children with severe language delays. *Journal of Applied Behavior Analysis, 31,* 357-374.

24. Ducharme, D.E., & Holborn, S.W. (1997). Programming generalization of social skills in preschool children with hearing impairments. *Journal of Applied Behavior Analysis, 30,* 639-651.

25. Eikeseth, S., & Nesset, R. (2003). Behavioral treatment of children with phonological disorder: The efficacy of vocal imitation and sufficient-response-exemplar training. *Journal of Applied Behavior Analysis, 36,* 325-337.

26. Engelman, K.K., Altus, D.E., & Mathews, R.M. (1999). Increasing engagement in daily activities by older adults with dementia. *Journal of Applied Behavior Analysis, 32,* 107-110.

27. Engelman, K.K., Altus, D.E., Mosier, M.C., & Mathews, R.M. (2003). Brief training to promote the use of less intrusive prompts by nursing assistants in a dementia care unit. *Journal of Applied Behavior Analysis, 36,* 129-132.

28. Faw, G.D., Davis, P.K., & Peck, C. (1996). Increasing self-determination: Teaching people with mental retardation to evaluate residential options. *Journal of Applied Behavior Analysis, 29,* 173-188.

29. Fueyo, V., & Bushell, D. (1998). Using number line procedures and peer tutoring to improve the mathematics computation of low-performing first graders. *Journal of Applied Behavior Analysis, 31,* 417-430.

30. Gena, A., Krantz, P.J., McClannahan, L.E., & Poulson, C.L. (1996). Training and generalization of affective behavior displayed by youth with autism. *Journal of Applied Behavior Analysis, 29,* 291-304.

31. Gilbert, L.M., Williams, R.L., & McLaughlin, T.F. (1996). Use of assisted reading to increase correct reading rates and decrease error rates of students with learning disabilities. *Journal of Applied Behavior Analysis, 29,* 255-257.

32. Gillat, A., & Sulzerazaroff, B. (1994). Promoting principals managerial involvement in instructional improvement. *Journal of Applied Behavior Analysis, 27,* 115-129.

33. Gross, A., Miltenberger, R., Knudson, P., Bosch, A., & Breitiwieser, C.B. (2007). Preliminaryevallation of a parent training program to prevent gun play. *Journal of Applied Behavior Analysis, 40,* 691-695.

34. Hagopian, L.P., Fisher, W.W., & Legacy, S.M. (1994). Schedule effects of noncontingent reinforcement on attention-maintained destructive behavior in identical quadruplets. *Journal of Applied Behavior Analysis, 27,* 317-325.

35. Harris, T.A., Peterson, S.L., Filliben, T.L., Glassberg, M., & Favell, J.E. (1998). Evaluating a more cost-efficient alternative to providing in-home feedback to parents: The use of spousal feedback. *Journal of Applied Behavior Analysis, 31,* 131-134.

36. Heck, A., Collins, J., & Peterson, L. (2001). Decreasing childrens risk taking on the playground. *Journal of Applied Behavior Analysis, 34,* 349-352.

37. Hendrickson, J.M., Gardner, N., Kaiser, A., & Riley, A. (1993). Evaluation of a social-interaction coaching program in an integrated day-care setting. *Journal of Applied Behavior Analysis, 26,* 213-225.

38. Hernandez, E., Hanley, G.P., Ingvarsson, E.T., & Tiger, J.H. (2007). A preliminary evaluation of the emergence of novel mand forms. *Journal of Applied Behavior Analysis, 40,* 137-156.

39. Himle, M.B., Miltenberger, R.G., Flessner, C., & Gatheridge, B. (2004). Teaching safety skills to children to prevent gun play. *Journal of Applied Behavior Analysis, 37,* 1-9.

40. Hughes, C., Harmer, M.L., Killian, D.J., & Niarhos, F. (1995). The effects of multiple-exemplar self-instructional training on high-school-students generalized conversational interactions. *Journal of Applied Behavior Analysis, 28,* 201-218.

41. Ingvarsson, E.T., & Hanley, G.P. (2006). An evaluation of computer-based programmed instruction for promoting teachers' greetings of parents by name. *Journal of Applied Behavior Analysis, 39,* 203-214.

42. Iwata, B.A., Wallace, M.D., Kahng, S.W., Lindberg, J.S., Roscoe, E.M., Conners, J., Hanley, G.P., Thompson, R.H., & Worsdell, A.S. (2000). Skill acquisition in the implementation of functional analysis methodology. *Journal of Applied Behavior Analysis, 33,* 181-194.

43. Johnson, M.D., & Fawcett, S.B. (1994). Courteous service – its assessment and modification in a human-service organization. *Journal of Applied Behavior Analysis, 27,* 145-152.

44. Kahng, S.W., Iwata, B.A., Thompson, R.H., & Hanley, G.P. (2000). A method for identifying satiation versus extinction effects under noncontingent reinforcement schedules. *Journal of Applied Behavior Analysis, 33,* 419-431.

45. Kamps, D.M., Barbetta, P.M., Leonard, B.R., & Delquadri, J. (1994). Classwide peer tutoring – an integration strategy to improve reading-skills and promote peer interactions among students with autism and

general-education peers. *Journal of Applied Behavior Analysis, 27,* 49-61.

46. Koegel, R.L., & Frea, W.D. (1993). Treatment of social-behavior in autism through the modification of pivotal social skills. *Journal of Applied Behavior Analysis, 26,* 369-377.

47. Krantz, P.J., & McClannahan, L.E. (1993). Teaching children with autism to initiate to peers – effects of a script-fading procedure. *Journal of Applied Behavior Analysis, 26,* 121-132.

48. LeBlanc, L.A., Coates, A.M., Daneshvar, S., Charlop-Christy, M.H., Morris, C., Lancaster, B.M. (2003). Using video modeling and reinforcement to teach perspective-taking skills to children with autism. *Journal of Applied Behavior Analysis, 36,* 253-257.

49. Lee, R., McComas, J.J., & Jawor, J. (2002). The effects of differential and lag reinforcement schedules on varied verbal responding by individuals with autism. *Journal of Applied Behavior Analysis, 35,* 391-402.

50. Lindberg, J.S., Iwata, B.A., Kahng, S.W., & DeLeon, I.G. (1999). DRO contingencies: An analysis of variable-momentary schedules. *Journal of Applied Behavior Analysis, 32,* 123-136.

51. Ludwig, T.D., Gray, T.W., & Rowell, A. (1998). Increasing recycling in academic buildings: A systematic replication. *Journal of Applied Behavior Analysis, 31,* 683-686.

52. Maag, J.W., Reid, R., & Digangi, S.A.(1993). Differential-effects of self-monitoring attention, accuracy, and productivity. *Journal of Applied Behavior Analysis, 26,* 329-344.

53. MacDuff, G.S., Krantz, P.J., & McClannahan, L.E. (1993). Teaching children with autism to use photographic activity schedules –

maintenance and generalization of complex response chains. *Journal of Applied Behavior Analysis, 26,* 89-97.

54. Magee, S.K., & Ellis, J. (2000). Extinction effects during the assessment of multiple problem behaviors. *Journal of Applied Behavior Analysis, 33,* 313-316.

55. Mayfield, K.H., & Vollmer, T.R. (2007). Teaching math skills to at-risk students using home based peer tutoring. *Journal of Applied Behavior Analysis, 40,* 223-237.

56. McComas, J.J., Lalli, J.S., & Benavides, C. (1999). Increasing accuracy and decreasing latency during clean intermittent self-catheterization procedures with young children. *Journal of Applied Behavior Analysis, 32,* 217-220.

57. Miltenberger, R.G., Flessner, C., Gatheridge, B., Johnson, B., Satterlund, M., & Egemo, K. (2004). Evaluation of behavioral skills training to prevent gun play in children. *Journal of Applied Behavior Analysis, 37,* 513-516.

58. Mozingo, D.B., Smith, T., Riordan, M.R., Reiss, M.L., & Bailey, J.S. (2006). Enhancing frequency recording by developmental disabilities treatment staff. *Journal of Applied Behavior Analysis, 39,* 253-256.

59. Neef, N.A., Marckel, J., Ferreri, S., Jung, S., Nist, L., & Armstrong, N. (2004). Effects of modeling versus instructions on sensitivity to reinforcement schedules. *Journal of Applied Behavior Analysis, 37,* 267-281.

60. Neef, N.A., Nelles, D.E., Iwata, B.A., & Page, T.J. (2003). Analysis of precurrent skills in solving mathematics story problems. *Journal of Applied Behavior Analysis, 36,* 21-33.

61. Nikopoulos, C.K., & Keenan, M. (2004). Effects of video modeling on social initiations by children with autism. *Journal of Applied Behavior Analysis, 37,* 93-96.

62. O'Callaghan, P.M., Allen, K.D., Powell, S., & Salama, F. (2006). The efficacy of noncontingent escape for decreasing children's disruptive behavior during restorative dental treatment. *Journal of Applied Behavior Analysis, 39,* 161-171.

63. Pace, G.M., Iwata, B.A., Cowdery, G.E., Andree, P.J., & McIntyre, T. (1993). Stimulus (instructional) fading during extinction of self-injurious escape behavior. *Journal of Applied Behavior Analysis, 26,* 205-212.

64. Petscher, E.S., & Bailey, J.S. (2006). Effects of training, prompting, and self-monitoring on staff behavior in a classroom for students with disabilities. *Journal of Applied Behavior Analysis, 39,* 215-226.

65. Phaneuf, L., & Mcintyre, L.L. (2007). Effects of individualized video feedback combined with group parent training on inappropriate maternal behavior. *Journal of Applied Behavior Analysis,40,* 737-741.

66. Porritt, M., Burt, A., & Poling, A. (2006). Increasing fiction writers' productivity through an internet-based intervention. *Journal of Applied Behavior Analysis, 39,* 393-397.

67. Reeve, S.A., Reeve, K.F., Townsend, D.B., & Poulson, C.L. (2007). Establishing a generalized repertoire of helping behavior in children with autism. *Journal of Applied Behavior Analysis, 40,* 123-136.

68. Rhymer, K.N., Skinner, C.H., Henington, C., D'Reaux, R.A., & Sims, S. (1998). Effects of explicit timing on mathematics problem completion rates in African-American third-grade elementary students. *Journal of Applied Behavior Analysis, 31,* 673-677.

69. Roscoe, E.M., Iwata, B.A., & Goh, H.L. (1998). A comparison of noncontingent reinforcement and sensory extinction as treatments for self-injurious behavior. *Journal of Applied Behavior Analysis, 31,* 635-646.

70. Ryan, S., Ormond, T., Imwold, C., & Rotunda, R.J. (2002). The effects of a public address system on the off-task behavior of elementary physical education students. *Journal of Applied Behavior Analysis, 35,* 305-308.

71. Sharpe, T., Brown, M., & Crider, K. (1995). The effects of a sportsmanship curriculum intervention on generalized positive social behavior of urban elementary school students. *Journal of Applied Behavior Analysis, 28,* 401-416.

72. Shore, B.A., Iwata, B.A., Vollmer, T.R., Lerman, D.C., & Zarcone, J.R. (1995). Pyramidal staff training in the extension of treatment for severe behavior disorders. *Journal of Applied Behavior Analysis, 28,* 323-332.

73. Solberg, K.M., Hanley, G.P., Layer, S.A., & Ingvarsson, E.T. (2007). The effects of reinforcer pairing and fading on preschoolers' snack selections. *Journal of Applied Behavior Analysis, 40,* 633-644.

74. Stark, L.J., Knapp, L.G., Bowen, A.M., Powers, S.W., Jelalian, E., Evans, S., Passero, M.A., Mulvihill, M.M., & Hovell, M. (1993). Increasing calorie consumption in children with cystic-fybrosis – replication with 2-year follow-up. *Journal of Applied Behavior Analysis, 26,* 435-450.

75. Taylor, B.A., Hughes, C.E., Richard, E., Hoch, H., & Coello, A.R. (2004). Teaching teenagers with autism to seek assistance when lost. *Journal of Applied Behavior Analysis, 37,* 79-82.

76. Twohig, M.P., Shoenberger, D., & Hayes, S.C. (2007). A preliminary 1nvestigation of acceptance and commitment therapy as a treatment for marijuana dependence in adults. *Journal of Applied Behavior Analysis, 40,* 619-632.

77. Twohig, M.P., & Woods, D.W. (2001). Habit reversal as a treatment for chronic skin picking in typically developing adult male siblings. *Journal of Applied Behavior Analysis, 34,* 217-220.

78. Van Houten, R., & Malenfant, J.E.L. (2004). Effects of a driver enforcement program on yielding to pedestrians. *Journal of Applied Behavior Analysis, 37,* 351-363.

79. Van Houten, R., Van Houten, J., & Malenfant, J.E.L. (2007). Impact of a comprehensive safety program on bicycle helmet use among middle-school children. *Journal of Applied Behavior Analysis, 40,* 239-247.

80. Vaughn, B.J., Clarke, S., & Dunlap, G. (1997). Assessment-based intervention for severe behavior problems in a natural family context. *Journal of Applied Behavior Analysis, 30,* 713-716.

81. Vintere, P., Hemmes, N.S., Brown, B.L., & Poulson, C.L. (2004). Gross-motor skill acquisition by preschool dance students under self-instruction procedures. *Journal of Applied Behavior Analysis, 37,* 305-322.

82. Wagaman, J.R., Miltenberger, R.G., & Arndorfer, R.E. (1993). Analysis of a simplified treatment for stuttering in children. *Journal of Applied Behavior Analysis, 26,* 53-61.

83. Ward, P., & Carnes, M. (2002). Effects of posting self-set goals on collegiate football players' skill execution during practice and games. *Journal of Applied Behavior Analysis, 35,* 1-12.

84. Watson, T.S. (1996). A prompt plus delayed contingency procedure for reducing bathroom graffiti. *Journal of Applied Behavior Analysis, 29*, 121-124.

85. Werle, M.A., Murphy, T.B., & Budd, K.S. (1993). Treating chronic food refusal in young-children – home-based parent training. *Journal of Applied Behavior Analysis, 26,* 421-433.

86. Wetterneck, C.T., & Woods, D.W. (2006). An evaluation of the effectiveness of exposure and response prevention on repetitive behaviors associated with Tourette's syndrome. *Journal of Applied Behavior Analysis, 39,* 441-444.

87. Williams, G., Donley, C.R., & Keller, J.W. (2000). Teaching children with autism to ask questions about hidden objects. *Journal of Applied Behavior Analysis, 33,* 627-630.

88. Woods, D.W., Miltenberger, R.G., & Lumley, V.A. (1996). Sequential application of major habit-reversal components to treat motor tics in children. *Journal of Applied Behavior Analysis, 29,* 483-493.

89. Wu, H., & Miller, L.K. (2007). A tutoring package to teach pronunciation of Mandarin Chinese characters. *Journal of Applied Behavior Analysis, 40,* 583-586.

90. Young, J.M., Krantz, P.J., McClannahan, L.E., & Poulson, C.L. (1994). Generalized imitation and response-class formation in children with autism. *Journal of Applied Behavior Analysis, 27,* 685-697.

91. Zarcone, J.R., Iwata, B.A., Smith, R.G., Mazaleski, J.L., & Lerman, D.C. (1994). Reemergence and extinction of self-injurious escape behavior during stimulus (instructional) fading. *Journal of Applied Behavior Analysis, 27,* 307-316.

92. Zarcone, J.R., Iwata, B.A., Vollmer, T.R., Jagtiani, S., Smith, R.G., & Mazaleski, J.L. (1993). *Journal of Applied Behavior Analysis, 26,* 353-360.

93. 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 B - Master Table: Populations and Topics Studied in the Articles

| | Population | Problem/Topic |
|---|---|---|
| 1 | 14-year-old boy with autism and mental retardation | food rejection: effects of adding condiments on consumption of previously rejected foods (vegetables |
| 2 | 2 to 3-year old children with a history of food refusal and with medical and developmental problems | comparing two treatments for food refusal |
| 3 | 6th, 7th, and 8th grade students with consistent difficulty remaining on task during the First 10 minutes of class | effects of teacher greetings on student on-task behavior |
| 4 | 14-year-old tennis player wit hanger control problems during tennis matches | use of a habit-reversal procedure for disruptive, angry outbursts during tennis matches |
| 5 | members of two academic departments at a large university | effects of prompts on recycling behavior |
| 6 | dishwashers (average age = 20) and servers (average age = 23) in a dining restaurant | use of task clarification and graphic and verbal feedback to increase closing-task completion |
| 7 | shoppers who enter a supermarket and place a child in the seat of a shopping cart | impact of verbal prompts on shopping cart safety-belt use |
| 8 | preschool children (4 to 5 years old) | effects of competitive and cooperative games on aggressive and cooperative behaviors |
| 9 | typically developing girls (3 to 4 years old) who could not perform a series of problem-solving tasks | establish derived relational responding in accordance with a comparative frame |
| 10 | 10-year-old boys who met the diagnostic | effects of tactical and strategic |

| | | |
|---|---|---|
| | criteria for ADHD | instructions on academic responding (adaptation to changing contingencies) |
| 11 | 7-year old boy with mental retardation who engaged in aggression and disruption | reduce multiply controlled problem behavior |
| 12 | employees of a private nonprofit research and treatment program for children and adults with autism | effects of proximity of containers on pounds of office paper recycled |
| 13 | 2-year-old typically developing boys | effects of a high-probability instruction sequence and a fixed-time schedule of reinforcement on compliance |
| 14 | 3 to 4 year old children with speech impairments | test the effectiveness of including speech production into naturalistic conversation training |
| 15 | 3-, 5-, and 12-year old boys with autism | acquisition of Picture Exchange Communication System and effects on emergence of speech |
| 16 | faculty, staff and students of a university who drive a car | active prompting to increase seat belt use and decrease cell phone use among drivers exiting a university parking lot |
| 17 | 10 to 19-year-old male students with acquired brain injury who exhibited significant behavior problems | effects of immediate performance feedback on special education teachers' treatment integrity for implementing antecedent and consequence procedures in an ongoing behavior support plan |
| 18 | 10 to 12-year-old 4$^{th}$ graders with developmental disabilities | modified constant prompt-delay procedure to teach spelling |
| 19 | 10 to 11-year-old students with orthopedic impairments | training to recruit teacher attention |
| 20 | 5 to 6-year-old boys with severe disabilities (autistic, mentally retarded and speech handicapped) | use of high-probability requests to increase social interactions |

| | | |
|---|---|---|
| 21 | 4 to 5-year-old children with emotional-behavioral disorders | effects of variant versus invariant high-probability requests on the performance of requests to initiate a social bid |
| 22 | 3- to 6-year-old children with autism | teaching imitation training of facial models |
| 23 | 3-year-old children with autism or pervasive developmental disorder with severe language delays | effects of differential reinforcement on the generalization of a replacement mand |
| 24 | 52 to 72-months-old children with moderate to severe hearing loss (and language retardation) | efficacy of a social skills training package in producing stimulus generalization |
| 25 | 5- to 6-year-old children with phonological disorder | does sufficient-response-exemplar training of vocal imitation result in improved articulation |
| 26 | 85- to 95-year old residents of a special care unit for persons with dementia and their certified nursing assistants | effects of nursing assistants' use of prompts and praise to increase the engagement in daily living activities |
| 27 | 76 to 80-year-old residents of a dementia care unit and their nursing assistants | effect of a brief staff-training procedure to increase the use of graduated prompting while helping to dress patients |
| 28 | 22- to 25-year old institutionalized adults with mental retardation | self determination: learning skills at asking questions, reporting information, evaluating homes: evaluate residential options |
| 29 | First-grade students low-performing in mathematics | using number line procedures and peer tutoring to improve mathematics computation |
| 30 | 11 to 18-year old children with autism | teach conceptually appropriate affective behavior |
| 31 | 7-year-old elementary students with learning disabilities | effects of assisted reading on correct reading |

| | | |
|---|---|---|
| 32 | 50-year-old principal of an elementary school and 1 classroom with children ranging from 8 to 9-years old | promote principals' visiting classrooms, emphasizing achievement and training, and supporting teachers |
| 33 | 4 to 7-year old children and their parents | parent training program to prevent gun play and teach children safety skills needed if they encounter an unattended firearm |
| 34 | 5-year-old female quadruplets with pervasive developmental disorder and mental retardation | effects of noncontingent reinforcement on attention-maintained destructive behavior |
| 35 | husband-and-wife dyads of parents of children with autism (2 to 3-year-old boys) | contribution of spousal feedback to a parent education curriculum |
| 36 | 1st graders, 2nd graders and kindergarten/third-grade students | decreasing risky playground behaviors |
| 37 | 21 to 23-year-old female day-care providers and 2 to 4-year-old children with severe social interaction deficits (and developmental disorder(s)) | effects of structured coaching procedure on teaching behavior and on social interaction |
| 38 | 4-year-old boys with nonspecific developmental delays and a 1.7-year-old typically developing girl who made few vocal requests in the classrooms | demonstrate the extent to which differential reinforcement of single-word mands or framed mands would result in the emergence of other single-word and framed mands |
| 39 | 4- to 5-year-old preschool children | effectiveness of a behavioral skills training program supplemented with in situ training for teaching children safety skills to use when they find a gun |
| 40 | 17- to 21-year old female high school students with moderate mental retardation | effects of multiple-exemplar self-instructional training on acquisition and generalization of conversational interaction |
| 41 | 20- to 23-year old undergraduate student | effects of fully automated computerized assessment and programmed instruction |

| | | |
|---|---|---|
| | teachers in a preschool | intervention on promoting frequent and accurate use of parent's names by teachers |
| 42 | upper-level undergraduate students in a laboratory course in applied behavioral analysis and 8 graduate students | examination of skill acquisition in implementing functional analyses by evaluating an instructional program |
| 43 | 38 to 63-year old receptionists at an antipoverty agency | strategies to courteous provision of service |
| 44 | institutionalized adults (25, 31 and 43 years old) with developmental disabilities who engaged in SIB or aggression | evaluation of one method for determining whether response suppression under noncontingent reinforcement is a function of satiation or extinction |
| 45 | 8 to 9-year-old male students with autism (high functioning) and their peers | effects of classwide peer tutoring relative to traditional instruction on reading skills and social interaction time |
| 46 | 13 to 16-year-old children with autism | acquisition of individual social communicative behaviors |
| 47 | 9 to 12-year-old children with autism | teach peer initiations about recently completed, current and future activities through a script fading procedure |
| 48 | 7 to 13-year-old boys with autism | evaluation of video modeling and reinforcement for teaching perspective-taking skills |
| 49 | 7-year-old boys and a 27-year-old man with autism | effects of a lag schedule of differential reinforcement on varied verbal responding to social questions |
| 50 | 28 to 50-year old institutionalized women with developmental disabilities (mental retardation) who engaged in SIB | comparative analyses to determine the relative effectiveness of variable-momentary differential-reinforcement of other-behavior schedules |
| 51 | patrons of two academic buildings at a | effect of location of receptacles on |

| | | |
|---|---|---|
| | university | recycling behavior |
| 52 | fourth and sixth graders with learning disabilities | effects of self-monitoring on-task behavior, academic productivity and academic accuracy |
| 53 | 9 to 14-year-old boys with autism | learn to follow photographic activity schedules to increase on-task and on-schedule behavior |
| 54 | 7 year old boy with ADHD and 8-year-old boy with mental retardation, with multiple problem behaviors | evaluation of extinction effects after one of several target behaviors was observed during functional analysis |
| 55 | 9 to 16-year-old children with a history of maltreatment | using home-based peer tutoring to teach math skills |
| 56 | 5-year-old girls with normal intelligence and that needed catherization (bladder dysfunctions) | effects of simulation training on performance of clean intermittent self-catherization procedures |
| 57 | 6- to 7-year-old children | evaluate behavioral skills training for teaching firearm safety skills to children |
| 58 | 22 to 60-year old female paraprofessional direct-care instructors in a residential treatment setting for persons with developmental disabilities and their patients (adults with profound mental retardation) | evaluation of staff training and management package for increasing accuracy of recording frequency of problem behavior |
| 59 | 10 to 13-year-old typically developing children and children with ADHD | examination of effects of modeling versus instructions on the choices of children whose academic responding showed insensitivity to reinforcement schedules |
| 60 | 19 to 23-year-old students with developmental disabilities | analysis of responses that increase the effectiveness of a subsequent or current behavior in obtaining a reinforcer to facilitate the solution of arithmetic word (story) problems |

| | | |
|---|---|---|
| 61 | 7 to 9-year-old children with autism | effects of a video modeling intervention on social initiation and play behaviors |
| 62 | 4 to 7-year-old children with significant problems with disruptive behavior during restorative dentistry | effectiveness of a dentist-implemented behavioral intervention with brief noncontingent escape for decreasing disruptive behavior |
| 63 | 2 to 17 year old children with mental retardation | combined treatment of extinction and instructional fading to reduce self-injurious behaviour |
| 64 | female instructional assistants in a classroom for students with disabilities (third to fifth graders with severe behavior problems) | effects of a treatment package (training, prompting and self-monitoring) on implementation of a token economy by instructional assistants (staff behavior) |
| 65 | parents of 2 to 4-year old children with developmental disabilities | effects of adding individualized video feedback to group-based parent training program to reduce inappropriate maternal behavior |
| 66 | members of an international online fiction writing community | treatment package to increase the writing productivity of fiction writers |
| 67 | 5 to 6-year-old children with autism | establish a generalized repertoire of helping adults with different tasks through the use of a multicomponent teaching package |
| 68 | African-American third-grade students | effects of explicit timing procedure on mathematics problem competition rates and accuracy levels |
| 69 | 20 to 35-year-old institutionalized adults with developmental disabilities (mental retardation) who engaged in SIB | comparison of effects of noncontingent reinforcement and sensory extinction on self-injurious behavior |
| 70 | teacher and students in 3 coeducational physical education classes at an elementary school (first and second grade) | effects of teacher feedback delivered via a public address system on off-task behavior |

| | | |
|---|---|---|
| 71 | students in third-grade physical education classes (average age: 8 years old) | effects of a sportsmanship curriculum intervention on generalized positive social behavior |
| 72 | 22 to 44-year-old institutionalized adults with mental retardation and behavior disorders (SIB) | pyramidal training procedure for staff working with individuals who exhibited self-injurious behaviour, aggression and disruption |
| 73 | 3 to 4-year-old typically developing preschoolers | effect of reinforcement pairing and fading on snack selection |
| 74 | 3 to 8-year-old children with cystic fybrosis and their parents | behavioral group-treatment program to promote and maintain increased calorie consumption |
| 75 | 13 to 17-year-old children with autism | teaching to respond to a vibrating pager to seek assistance in community settings when physically separated from parents or teachers |
| 76 | 19 to 43-year-old adults with marijuana dependence | effect of an acceptance and commitment therapy for marijuana dependence |
| 77 | 20 to 22-year-old typically developing male sibling who engage in skin picking | evaluation of habit reversal as a treatment for skin picking |
| 78 | drivers travelling on two high-crash corridors | evaluation of a driver-yielding enforcement program on yielding to pedestrians |
| 79 | middle school students who come to schools on their bicycles | evaluation of bicycle helmet program |
| 80 | 8-year-old boy with severe disabilities including agenesis of the corpus callosum with severe behavior problems | functional assessments and assessment-based interventions for severe problem behavior in a natural family context |
| 81 | 3 to 5-year-old students in a preballet class at a private school | effects of two training procedures (modeling and praise; self-instruction, modeling and praise) on complex gross-motor chain acquisition |

| | | |
|---|---|---|
| 82 | 6 to 10-year-old stutterers | treatment for stuttering through awareness training, response incompatibility and social support |
| 83 | 19 to 21-year-old linebackers of a Division II football team | effects of self-set goals and public posting on athletic performance |
| 84 | visitors of men's restrooms on a college campus | effectiveness of posting signs for reducing graffiti |
| 85 | 21 to 54-months-old boys with selective food refusal and their mothers | effects of behavioral parent training program on parent and child feeding-related behaviors in the home |
| 86 | 11-year-old boy with Tourette's syndrome | evaluation of exposure and response prevention as treatment for repetitive behaviors |
| 87 | 4-year-old girls with autism | teaching to ask questions about hidden objects |
| 88 | 8 to 12-year-old children with chronic motor tics | sequential administration of habit-reversal components to treat motor tics |
| 89 | 25-year-old male American college student | effects of a tutoring package on Chinese pronunciations |
| 90 | 2 to 4-year-old children with autism | influence of response topography on generalization of imitation |
| 91 | 27 to 40-year-old institutionalized adults with mental retardation who engage in SIB | effects of stimulus fading on self-injurious behavior |
| 92 | 26 to 39-year-old institutionalized women with profound mental retardation who engage in SIB | treatment of self-injurious behavior maintained by negative reinforcement |
| 93 | male collegiate soccer players from a Division I team (mean age: 20-years-old) | effectiveness of an attentional training program on the execution of targeted soccer skills |

# Appendix C - Master Table: Type of Multiple Baseline Design, Median Number of Units, Median Number of Observations, Placement of the Intervention Points, and Analyses.

| | Type | Placement | Units | Observations | Analyses |
|---|---|---|---|---|---|
| 1 | Settings | systematic (apparent) | 3 | 38 | visual |
| 2 | Subjects | response-guided (explicit) | 3 | 60 | visual; descriptive stats |
| 3 | Subjects | response-guided (apparent) | 3 | 10 | visual; descriptive stats |
| 4 | Settings | response-guided (explicit) | 2 | 18.5 | visual |
| 5 | Settings | systematic (apparent) | 2 | 50 | visual; descriptive stats |
| 6 | Subjects | systematic (explicit) | 2 | 35 | visual; descriptive stats |
| 7 | Settings | systematic (apparent) | 2 | 38 | visual |
| 8 | Subjects | response-guided (explicit) | 2 | 39 | visual; descriptive stats |
| 9 | Subjects | response-guided (explicit) | 2 | 67 | visual |
| 10 | Subjects | response-guided (explicit) | 2 | 53 | visual; descriptive stats |
| 11 | Settings | systematic (apparent) | 3 | 48 | visual; descriptive stats |
| 12 | Settings | systematic (apparent) | 3 | 54 | visual; descriptive stats |
| 13 | Subjects | response-guided (apparent) | 2 | 55 | visual; descriptive stats |
| 14 | Behaviors | response-guided (apparent) | 2.5 | 28 | visual; descriptive stats |
| 15 | Subjects | response-guided (explicit) | 3 | 18 | visual; descriptive stats |
| 16 | Behaviors | systematic (apparent) | 2 | 65 | visual; descriptive stats |
| 17 | Subjects | response-guided (explicit) | 2.5 | 17 | visual; descriptive stats |
| 18 | Settings | response-guided (explicit) | 3 | 23 | visual; descriptive stats |
| 19 | Subjects | systematic (apparent) | 4 | 34.5 | visual; descriptive stats |
| 20 | Subjects | response-guided (apparent) | 3 | 25 | visual; descriptive stats |
| 21 | Subjects | response-guided (apparent) | 2 | 38 | visual; descriptive stats |
| 22 | Subjects | systematic (apparent) | 3 | 36 | visual |
| 23 | Settings | response-guided (explicit) | 3 | 9 | visual |
| 24 | Subjects | systematic (apparent) | 2 | 36 | visual; descriptive stats |

| | | | | | |
|---|---|---|---|---|---|
| 25 | Behaviors | response-guided (explicit) | 3 | 9 | visual |
| 26 | Settings | systematic (apparent) | 2 | 24 | visual; descriptive stats |
| 27 | Subjects | systematic (apparent) | 2 | 15 | visual; descriptive stats |
| 28 | Subjects | response guided (explicit) | 4 | 12.5 | visual; descriptive stats |
| 29 | Subjects | systematic (explicit) | 3 | 24.5 | visual; descriptive stats |
| 30 | Behaviors | response-guided (explicit) | 3.5 | 100 | visual; descriptive stats |
| 31 | Subjects | systematic (apparent) | 3 | 41 | visual; descriptive stats |
| 32 | Subjects | systematic (explicit) | 2 | 15 | visual; descriptive stats; inferential stats |
| 33 | Subjects | response-guided (explicit) | 4 | 8.5 | visual |
| 34 | Subjects | response-guided (explicit) | 4 | 79 | visual; descriptive stats |
| 35 | Subjects | systematic (explicit) | 3 | 30 | visual; descriptive stats |
| 36 | Subjects | systematic (apparent) | 3 | 22 | visual |
| 37 | Subjects | systematic (apparent) | 3 | 30 | visual; descriptive stats |
| 38 | Behaviors | response-guided (explicit) | 5 | 80 | visual; descriptive stats |
| 39 | Subjects | systematic (apparent) | 8 | 11.5 | visual |
| 40 | Subjects | response-guided (explicit) | 4 | 48.5 | visual; descriptive stats; inferential stats |
| 41 | Subjects | systematic (explicit) | 4 | 49.5 | visual; descriptive stats |
| 42 | Subjects | systematic (explicit) | 5.5 | 15 | visual; descriptive stats |
| 43 | Subjects | systematic (explicit) | 3 | 113 | visual; descriptive stats |
| 44 | Subjects | systematic (apparent) | 3 | 26 | visual; descriptive stats |
| 45 | Subjects | systematic (apparent) | 3 | 44 | visual; descriptive stats |
| 46 | Beh + subs | systematic (apparent) | 3 | 88 | visual |
| 47 | Subjects | systematic (apparent) | 4 | 43 | visual; descriptive stats |
| 48 | Behaviors | response-guided (explicit) | 2 | 17 | visual |
| 49 | Subjects | response-guided (explicit) | 3 | 39 | visual |
| 50 | Subjects | systematic (apparent) | 2 | 71.5 | visual |
| 51 | Subjects | systematic (explicit) | 2 | 73.5 | visual; descriptive stats |
| 52 | Subjects | systematic (explicit) | 2 | 100 | visual; descriptive stats |
| 53 | Subjects | systematic (apparent) | 4 | 105 | visual; descriptive stats |
| 54 | Behaviors | systematic (apparent) | 3.5 | 33 | visual |
| 55 | Behaviors | response-guided (explicit) | 3.5 | 33.5 | visual; descriptive stats |
| 56 | Behaviors | response-guided (explicit) | 3 | 45 | visual; descriptive stats |

| | | | | | |
|---|---|---|---|---|---|
| 57 | Subjects | response-guided (explicit) | 6 | 11.5 | visual |
| 58 | Settings | response-guided (apparent) | 2 | 25 | visual; descriptive stats |
| 59 | Subjects | response-guided (explicit) | 3 | 23 | visual; descriptive stats |
| 60 | Behaviors | systematic (apparent) | 5 | 11 | visual; descriptive stats |
| 61 | Subjects | response-guided (explicit) | 3 | 25 | visual |
| 62 | Subjects | response-guided (explicit) | 5 | 32 | visual; descriptive stats |
| 63 | Subjects | systematic (apparent) | 3 | 110 | visual; descriptive stats |
| 64 | Behaviors | response-guided (explicit) | 3 | 36 | visual; descriptive stats |
| 65 | Subjects | systematic (explicit) | 4 | 9.5 | visual; descriptive stats |
| 66 | Subjects | systematic (explicit) | 2 | 52.5 | visual |
| 67 | Subjects | response-guided (explicit) | 4 | 51 | visual; descriptive stats |
| 68 | Subjects | systematic (apparent) | 3 | 9 | visual; descriptive stats; inferential stats |
| 69 | Subjects | response-guided (apparent) | 3 | 40 | visual; descriptive stats |
| 70 | Subjects | response-guided (apparent) | 3 | 27 | visual; descriptive stats |
| 71 | Subjects | systematic (apparent) | 3 | 32 | visual; descriptive stats; inferential stats |
| 72 | Subjects | systematic (apparent) | 2 | 43.5 | visual |
| 73 | Subjects | response-guided (explicit) | 3 | 60 | visual; descriptive stats |
| 74 | Settings | response-guided (apparent) | 4 | 90 | visual; descriptive stats; inferential stats |
| 75 | Subjects | systematic (apparent) | 3 | 45 | visual |
| 76 | Subjects | response-guided (explicit) | 3 | 110 | visual; descriptive stats |
| 77 | Subjects | response-guided (explicit) | 2 | 29.5 | visual; descriptive stats |
| 78 | Settings | response-guided (explicit) | 2 | 31 | visual; descriptive stats |
| 79 | Subjects | response-guided (explicit) | 3 | 26 | visual; descriptive stats |
| 80 | Settings | systematic (apparent) | 2 | 20 | visual; descriptive stats |
| 81 | Behaviors | response-guided (explicit) | 3 | 12 | visual; descriptive stats |
| 82 | Subjects | response-guided (explicit) | 4 | 40 | visual; descriptive stats |
| 83 | Behaviors | systematic (apparent) | 3 | 43 | visual |
| 84 | Settings | response-guided (apparent) | 3 | 53 | visual; descriptive stats |
| 85 | Subjects | response-guided (explicit) | 3 | 17 | visual; descriptive stats |
| 86 | Behaviors | response-guided (explicit) | 3 | 9 | visual; descriptive stats |
| 87 | Behaviors | systematic (apparent) | 3 | 23 | visual; descriptive stats |
| 88 | Subjects | response-guided (explicit) | 6 | 28.5 | visual; descriptive stats |

| | | | | | |
|---|---|---|---|---|---|
| | + behaviors | | | | |
| 89 | Settings | systematic (apparent) | 2 | 17.5 | visual; descriptive stats |
| 90 | Behaviors | response-guided (apparent) | 3 | 125 | visual |
| 91 | Subjects | systematic (apparent) | 3 | 200 | visual; descriptive stats |
| 92 | Subjects | response-guided (apparent) | 3 | 110 | visual |
| 93 | Subjects | systematic (explicit) | 4 | 24 | visual; descriptive stats |

# Appendix D – Overview of R functions.

| 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)<br>-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<br>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)<br>-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)<br>-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)<br>-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<br>2) start points<br>3) save |
| distribution.random | -design: “MBD” (multiple baseline design)<br>-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)<br>-save: “yes” (save the distribution to a | Constructs the random randomization distribution under the null hypothesis, where all the | 1) data<br>2) start points<br>3) save |

| | | | |
|---|---|---|---|
| | file), or “no” (see the distribution as output in the R console)<br>-number: how many randomizations are required | test statistics are calculated. | |
| pvalue.systematic | -design: “MBD” (multiple baseline design)<br>-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)<br>-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<br>2) start points<br>3) save |
| pvalue.random | -design: “MBD” (multiple baseline design)<br>-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)<br>-save: “yes” (save the distribution to a file), or “no” (see the distribution as output in the R console)<br>-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<br>2) start points<br>3) save |