6

ONE BY ONE

The design and analysis of replicated randomized single-case experiments

Patrick Onghena

FACULTY OF PSYCHOLOGY AND EDUCATIONAL SCIENCES, CENTRE FOR METHODOLOGY OF EDUCATIONAL RESEARCH, KU LEUVEN, BELGIUM

Introduction

A solution to the problem of small sample sizes for the design and analysis of empirical research can be found in at least two ways. The first, most common, way is by examining the behavior of large-sample inferential techniques if sample size becomes smaller and smaller, and by determining the sample size at which these techniques break down; see, for example, Chapter 17 (Smid & Rosseel). The guiding question is: "How low can you go?" A second, radically different, way is to intensively study a single participant (see also Chapter 7; Maric & Van der Werff) and to use replication logic or meta-analysis to arrive at the small sample. The present chapter takes the latter perspective. Because usually large numbers of repeated measurements for each participant are recommended, this perspective paradoxically offers a "large n" solution to the problem of small sample sizes.

Besides the repeated measurements, another part of the small sample size solution presented in this chapter lies in the addition of classical experimental design elements that can support the inference: randomization, blocking, and replication (Fisher, 1926). The result will be a "single-case experiment". In the first section of this chapter, this kind of experiment will be introduced, and in the subsequent sections some design and analysis options will be added.

The single-case experiment

A single-case experiment is defined as an experiment in which one unit is observed repeatedly during a certain period of time under different levels of at least one manipulated variable (Barlow, Nock, & Hersen, 2009). This definition is important to make the distinction between single-case experiments and other types of single-case research (e.g., qualitative case studies or observational time-series research).

Furthermore, this definition clarifies that randomization and replication are not intrinsic parts of a single-case experiment. Randomization and replication are considered in this chapter as additional tools that may substantially increase the internal and external validity, and hence the credibility, of single-case experimental research.

Example

Randy is taking sleeping pills for his sleeping problems but lately he has also started to experience disturbingly severe dizziness in the morning. He contacts his physician with this complaint and together they wonder whether the dizziness is caused by the sleeping pills. They decide to set up a two-week experiment in which Randy will register the severity of his dizziness daily on a seven-point scale, while taking the sleeping pills on some days and while replacing them with placebo pills on other days.

They contact a pharmacist, who prepares a strip with 14 identical capsules; seven of them contain the active sleeping medication and seven contain powdered sucrose. The capsules are unlabelled and randomly ordered to make sure that neither Randy nor the physician will know which pills are taken on which days, to rule out expectancy effects and confirmation bias. So together they set up a randomized placebo-controlled double-blind clinical trial on a single participant. After two weeks, Randy returns to his physician and shows his results. These can be found in Figure 6.1.

Without knowing the exact sequence in which the active medication and the placebo have been administered, these results only show the general distribution of the complaints during the two-week period. This is the point where the pharmacist is contacted again to disclose the randomization sequence. The sequence happens to be

E C E C C E C E E C C E E C

with *E* for *Experimental treatment* (sleeping pill) and *C* for *Control treatment* (placebo pill). If Figure 6.1 is transformed by connecting the dots for each experimental condition, Figure 6.2 is obtained.

Figure 6.2 shows that, by eyeballing, taking the sleeping pills is associated with higher severity ratings than taking the placebo pills. However, this is not always the case: the full line and the dotted line intersect. For example, on Day 10, a morning after taking a placebo pill, the complaints are more severe than on Day 9, a morning after taking the active sleeping medication. Now the patient and the physician may want to interpret and summarize these results even further. They may want to quantify the effect (using a measure of effect size) or they may want to know whether the effect is statistically significant

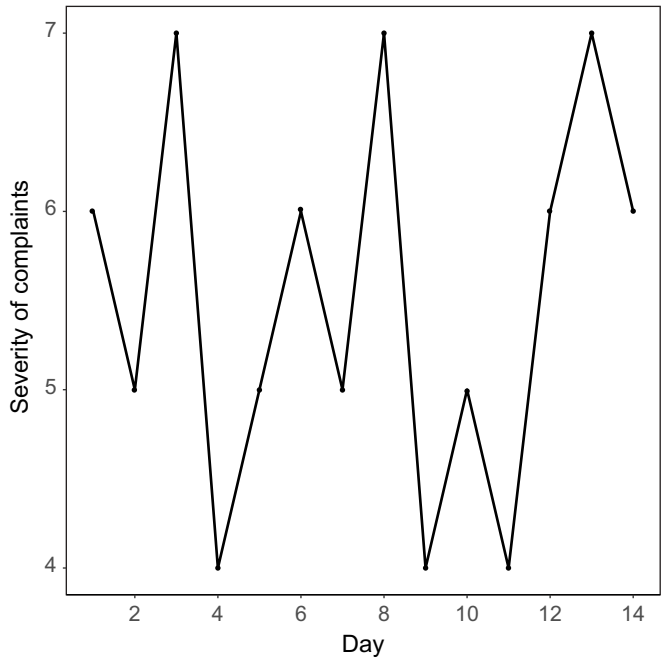

FIGURE 6.1 Example data for two weeks' daily registrations of the severity of dizziness on a seven-point scale by a single participant

(using inferential techniques). Ultimately, they have to decide on a course of action: continue or discontinue the sleeping pills, or eventually conducting some more follow-up experiments. These data and their analysis will be revisited in the Analysis section of this chapter.

Design of single-case experiments

Single-case experimental designs have been classified in many different ways, using divergent terminology (see, for example, Barlow et al., 2009). This section features one possible classification that is based on the two additional design elements: randomization and replication. Replicated single-case experiments represent one of the small sample size solutions that is central to this book.

The randomized experimental approach

In single-case experiments, *randomization* refers to the random assignment of measurement occasions to treatments. Random assignment of measurement occasions to treatments provides statistical control over both known and unknown variables that are time related (e.g., "history", "maturation"; Shadish,

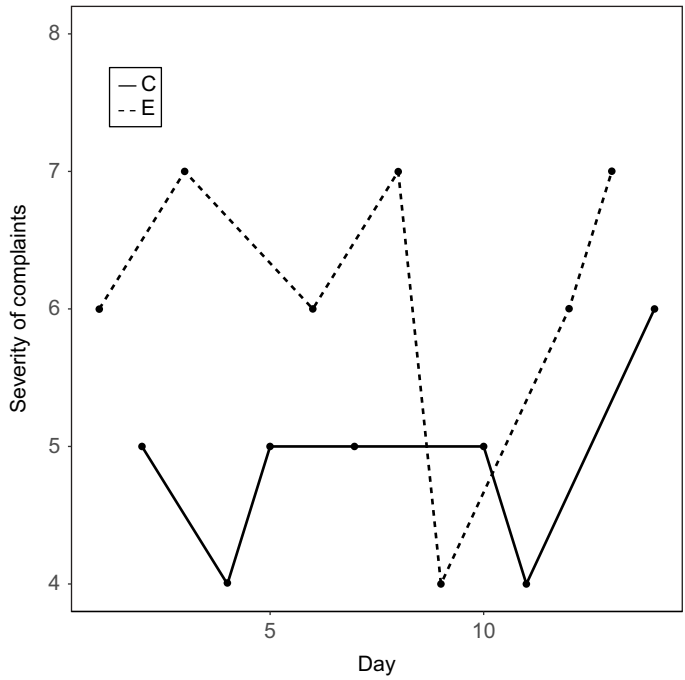

FIGURE 6.2 Example data of Figure 6.1 but distinguishing between the days when taking a sleeping pill (*E* for *Experimental treatment*) and days taking a placebo pill (*C* for *Control treatment*)

Cook, & Campbell, 2002). In addition, this random assignment is the basis for a statistical test, mentioned later in this chapter: the randomization test.

The specific random assignment procedure is intimately linked to the design of the experiment. Two broad design categories can be distinguished: alternation designs and phase designs. Alternation designs have rapid alternation of levels of the manipulated variable, just as in the Randy example. Their design characteristic is that each level of the manipulated variable can be implemented at each measurement occasion. Phase designs have consecutive measurement occasions under the same level of the manipulated variable. Their design characteristic is that the measurement occasions are divided in phases, and that the number of phases is preplanned and limited.

Alteration designs

The random assignment procedure in alternation designs is most straightforward and comparable to the random assignment procedure in group comparison designs. Because each level of the manipulated variable can be implemented at each

measurement occasion, it is like tossing a coin to determine the level at each measurement occasion. A Completely Randomized Design (CRD) is obtained by randomly selecting a sequence of levels, given a prespecified total number of measurement occasions, number of levels, and number of measurement occasions for each level. To take the Randy example: there are 14 measurement occasions, two levels, and seven measurement occasions for each level. Therefore, the random assignment procedure consists of randomly selecting one sequence out of $\binom{14}{7} = 3432$ possibilities.

A CRD is rarely used in practice because it includes possibilities in which the levels are all grouped together. In the Randy example, this means that

C C C C C C C E E E E E E E

would be one of the possible designs to work with. In order to avoid such sequences, researchers use other designs, such as the Randomized Block Design (RBD). In an RBD, blocks of levels are constructed and random assignment only occurs within blocks. In the Randy example, this could mean that blocks of C E pairs are constructed and that it is randomly determined for each pair whether the order is "C E" or "E C". The obtained sequence mentioned before can be symbolized as

[EC] [EC] [CE] [CE] [EC] [CE] [EC]

with the square brackets indicating that random ordering only occurred within the brackets. The random assignment procedure for an RBD would consist of randomly selecting one sequence out of $2^7 = 128$ possibilities.

One could, however, argue that an RBD is overly restrictive regarding the number of permissible sequences. For example, there is no obvious reason why a sequence like

E E C E C C E C E E C C E C

should be avoided, although it is impossible given the RBD specification. A workable compromise between a CRD and an RBD is to use the specifications of the CRD but to limit the number of consecutive measurement occasions for the same level. In the Randy example, this could mean that there are 14 measurement occasions, two levels, and seven measurement occasions for each level, and that the maximum number of consecutive measurement occasions for the same level is specified at two (to avoid phases of three consecutive measurement occasions for the same level). This random assignment procedure implies that 518 possibilities of the original 3,432 CRD possibilities

are permissible, which is more than four times the number of RBD possibilities. The logic of this design is most similar to what Barlow et al. (2009) called an Alternating Treatments Design (ATD).

Phase designs

The random assignment procedure in phase designs is less obvious and has no corresponding group comparison design analogue. Because the measurement occasions are divided in phases and the number of phases is preplanned and limited, it is not possible to randomly determine the level at each measurement occasion. For example, phase designs are denoted as AB designs (two levels of the manipulated variable, two phases), ABA designs (two levels of the manipulated variable, three phases, with the third phase called a "reversal" or "withdrawal" phase), ABAB designs (two levels of the manipulated variable, four phases), ABACABA designs (three levels of the manipulated variable, seven phases), and so on. Phase designs need phases, so once a certain level of the manipulated variable is started, several measurement occasions for that level should follow.

The one component of phase designs that can be randomly determined is the moment of the first measurement occasion after a phase transition. Figure 6.3 shows

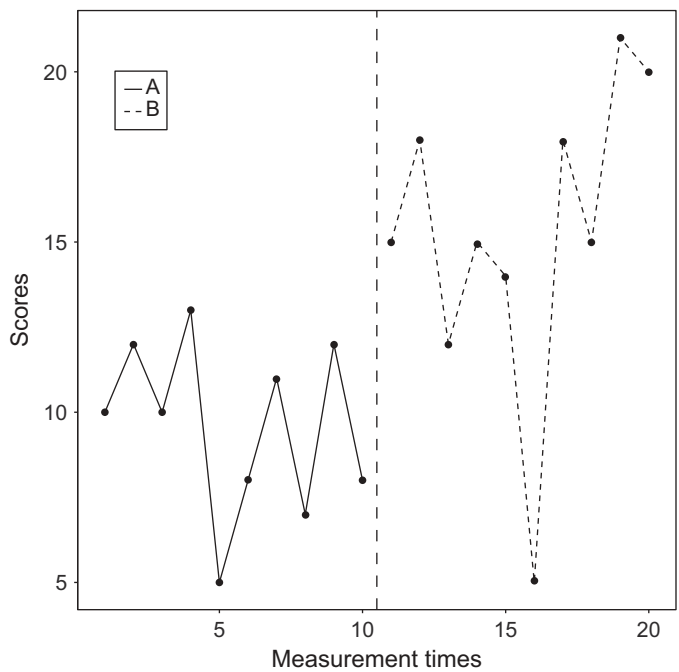

FIGURE 6.3 Hypothetical data for a single-case experiment using an AB design with 20 measurement occasions: 10 in the baseline phase (A) and 10 in the treatment phase (B)

hypothetical results for an AB design. There are 20 measurement occasions, 10 for level A and 10 for level B. The first measurement occasion after the (only) phase transition is the 11th. This moment might have been randomly determined. For example, given 20 measurement occasions and at least three measurement occasions in each phase, this results in 15 possibilities going from

A A A B B B B B B B B B B B B B B B B

to

A A A A A A A A A A A A A A A A A B B B

Random assignment in such an AB design involves randomly selecting one of the 15 possibilities out of this list. A similar procedure can be followed in ABA and ABAB designs, and other phase designs, with the list of possibilities for each of these designs generated by combining the possible moments for two or more phase transitions.

Phase logic can be combined with alternation logic if the order of the phases is allowed to be determined randomly. For example, if the previous specifications can be extended to include designs with a B phase before an A phase, then the previous 15 possibilities are combined with another 15 possibilities going from

B B B A A A A A A A A A A A A A A A A A

to

B B B B B B B B B B B B B B B B B A A A

The combination of phase logic and alternation logic results in even more combination possibilities for ABA designs, ABAB designs, ABACABA designs, and so on.

The replicated experimental approach

Replication is essential for many research endeavors, and replicability may even be considered as a demarcation criterion between science and non-science (Open Science Collaboration, 2015). In single-case research, replicability pertains to the results at the single-case level, and two basic procedures have been part of single-case research since its earliest inception (Barlow et al., 2009): direct replication and systematic replication. Direct replication is the repetition of the same experiment by

the same researcher; systematic replication is the repetition of the experiment with one or more variations as compared to the original experiment (e.g., another setting, another outcome variable, another treatment). From a design perspective, the distinction between sequential and simultaneous replications is crucial (Onghena & Edgington, 2005).

Sequential replication designs

The most common way to include replication in the design is by performing single-case experiments one by one. For example, single-case experiments in a hospital might be performed with consecutive patients. The results of the earlier experiments in the series can be used as prior information to optimize the later single-case experiments. This does not mean that in such a sequential replication design, the next experiment is only started if the previous one is completely finished; it just means that the timing is not preplanned and that the replications do not necessarily overlap in time.

Simultaneous replication designs

A special category of single-case designs is derived if the replications are preplanned and simultaneous. In a so-called "multiple-baseline across-participants design", several AB phase designs are conducted at the same time, and the B phases are started at different "staggered" moments. For example, Figure 6.4 shows the results of a four-week multiple-baseline design across three children for evaluating the effect of a behavioral intervention on the general distress level of the children as assessed daily by the staff at a daycare center. The intervention is started after the first week for the first child, after the second week for the second child, and after the third week for the third child. If there is a functional relation between the intervention and the distress, then the distress is expected to decrease after the intervention is implemented for the target child, but not for other children. In order to check for this differential effect, the results of all children are monitored simultaneously. If there would be a more general external cause of the decrease in distress, then all children would be affected (e.g., by a change in policy at the daycare center). Hence in multiple-baseline across-participants designs both within-AB and between-AB comparisons are possible and important. Simultaneous replication designs, in contrast to sequential replication designs, have the advantage that the confounding effects of "history" (external events) can be controlled by the between-series comparisons (Onghena & Edgington, 2005).

Classical multiple-baseline designs, with AB phase designs as building blocks, have also been proposed across behaviors (within one participant) and across settings (within one participant), but simultaneous replication designs can be constructed with other phase and alternation designs as well.

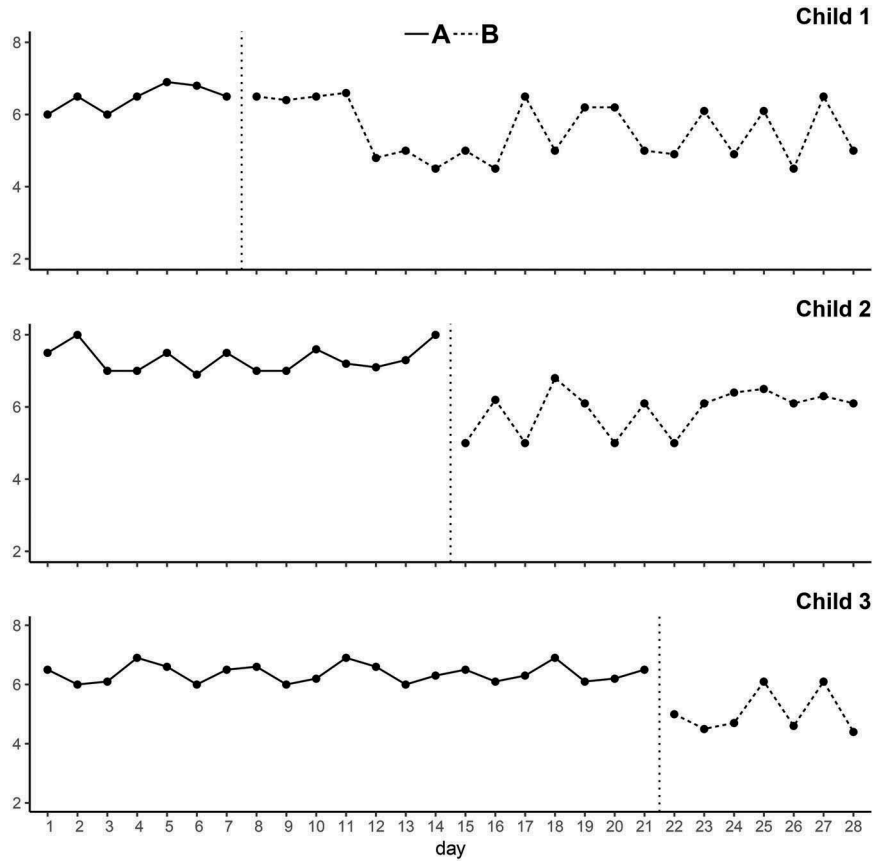

FIGURE 6.4 Hypothetical data for a single-case experiment using a four-week multiple-baseline design across three children for evaluating the effect of a behavioral intervention on the general distress level of the children as assessed daily by the staff at a daycare center

Analysis of single-case experiments

Several authors have contended that there is controversy regarding the analysis of data collected in single-case experiments and that these controversies should be resolved before making recommendations (see, for example, Kratochwill et al., 2010), but this is not the point of departure in the present section. Controversies in data analysis are intrinsic to making sense of data and to scientific inference in general, and single-case research is no exception (see, for example, discussions on Bayesian versus frequentist statistics, Neyman–Pearson versus Fisher approaches, and parametric versus nonparametric techniques; Silberzahn et al., 2018). Our point of departure is that all techniques that have been proposed for general data analysis can, in one way or another, be modified

to suit single-case data analysis, taking into account the defining characteristics of single-case data (for an elaboration of this perspective, see Onghena, Michiels, Jamshidi, Moeyaert, & Van Den Noortgate, 2018).

Without going into the details of all possible data-analytical techniques, it is important to highlight three different dichotomies that have to be taken into account when analyzing (single-case) data. These dichotomies are sometimes muddled up in applications or discussions and lead to dispensable controversy. The first dichotomy is descriptive versus inferential use of statistics. The other two dichotomies fall within the realm of inferential statistics: model- versus design-based inference on the one hand, and unilevel versus multilevel inference on the other hand.

Descriptive versus inferential statistics

Statistics can be used for description and statistics can be used for inference (Moore & Notz, 2017). *Descriptive use* refers to summarizing the observations at hand and *inferential use* refers to making statements (inferences) about unobserved quantities or patterns. In the latter case it is important to be explicit about the kind of inferences one wants to make: population inferences, causal inferences, or inferences to future observations.

If description is the only purpose, then tables, graphs, and descriptive measures suffice. Visual inspection and analysis, which is the most common and popular technique for analyzing single-case data, falls within this category. Visual analysis involves a time-series plot of the results and a systematic interpretation of six data aspects that should be evident in the plot: the level of the measurements within and between phases, possible trends, the variability within and between phases, the immediacy of the effect after the introduction of the treatment, the overlap of the measurements in different phases, and the consistency of data patterns across similar phases (Kratochwill et al., 2010). In addition, each of these data aspects may be quantified using a descriptive measure or an effect size statistic (Tanious, De, Michiels, Van Den Noortgate, & Onghena, 2019).

Model- versus design-based inference

If statistics are used for inferential purposes, then one of the most basic distinctions is the distinction between model- and design-based inference (Koch & Gillings, 1984). In model-based inference, a statistical model is fitted to the observed data and the parameters and their uncertainty are estimated from the data. In design-based inference, the data are analyzed in accordance with the way the data were collected; the only stochastic element entering the calculations refers to the sampling scheme or the assignment procedure. Although in practice both model and design information might be integrated and combined, the results of model-based inference and design-based inference in their purest forms do not necessarily coincide (Gregoire, 1998). A prototypical example of model-based inference for

a single-case experiment is time-series analysis (Tarlow & Brossart, 2018). A prototypical example of design-based inference for a single-case experiment is a randomization test and a confidence interval based on randomization test inversion (Michiels, Heyvaert, Meulders, & Onghena, 2017).

Unilevel versus multilevel inference

In replication designs, more than one level of inference is involved: the repeated measurements are nested within participants, and the participants can be conceptualized as a random sample from a target population (Moeyaert, Ferron, Beretvas, & Van Den Noortgate, 2014; Onghena et al., 2018). In this setting, multilevel inference can be conceptualized as a meta-analytical model (Van Den Noortgate & Onghena, 2003) or can be the result of using combination methods following a design-based approach (Levin, Ferron, & Gafurov, 2018; Tyrell, Corey, Feldman, & Silverman, 2013). Figure 6.5 shows the fourfold classification that arises by crossing the two inferential dichotomies, together with the prototypical examples.

An example of unilevel design-based inference

Unilevel design-based inference will be used for the Randy example and the data in Figures 6.1 and 6.2. This approach was selected because it is appealing for a properly designed experiment with a limited number of repeated measurements, and because it is relatively easy to explain and to understand. Furthermore, the approach stays close to the data and the data description, and makes minimal statistical assumptions, which is most compatible with the behavior analytical perspective from which the single-case experimental design tradition grew (Jacobs, 2019).

Take the data in Figure 6.2. In the Rubin model for causal inference, the causal effect is defined by E − C at a given measurement occasion and the Average Causal Effect is obtained by averaging over all measurement occasions (Holland, 1986). The fundamental problem of causal inference is that on a given

	Model-based	**Design-based**
Unilevel	Time series models	Randomization-based inference
Multilevel	Meta-analytical models	Combination methods

FIGURE 6.5 Classification of data-analytic strategies for (replicated) single-case experiments

measurement occasion only one observation (for one of the treatment levels) can be made; the observation for the other treatment level (the counterfactual) is missing. For example, on Day 1, there is an observation for E, but there is no observation for C. On Day 2, it is the other way around, and so on. The observed data were:

E	C	E	C	C	E	C	E	E	C	C	E	E	C
6	5	7	4	5	6	5	7	4	5	4	6	7	6

but the counterfactual is missing:

C	E	C	E	E	C	E	C	C	E	E	C	C	E
?	?	?	?	?	?	?	?	?	?	?	?	?	?

The design-based solution to the fundamental problem of causal inference is to examine what happens if there was no relation between x and y whatsoever (i.e., null hypothesis reasoning). In that case, the missing data are known. If the null hypothesis were true, then exactly the same series of data would be observed irrespective of the treatment level to which that occasion was assigned.

This null hypothesis reasoning can now be used to evaluate whether the observed data are extreme or unexpected in any way, given that the null hypothesis is true. A first step is to summarize the data in a test statistic. Suppose Randy and the physician agreed to take the absolute difference in means as their primary outcome (absolute difference because they were not sure about the direction of the effect). For the observed data, this test statistic amounts to $\mid 6.1428 - 4.8571 \mid = 1.2857$ on the seven-point scale. The second step is to calculate this test statistic for all random assignments that could have been chosen. This calculation makes sense, because if the null hypothesis were true, then the data are a fixed time series of numbers whatever random assignment was selected.

Suppose an RBD was used, then a total of 128 test statistic values are calculated (with the observed value as one of them). These 128 values form the reference distribution in which the observed value is located. A dot plot representation of the reference distribution for the difference in means between T and C is shown in Figure 6.6. The calculations needed to construct this reference distribution are tedious, but fortunately user-friendly software is available (see Bulté & Onghena, 2008, 2009, 2013; De, Michiels, Vlaeyen, & Onghena, 2017).

A third step is to calculate the probability to arrive at a test statistic value that is equal to or even more extreme than the observed value if the null hypothesis were true. This is the p-value of the randomization test. Figure 6.6 shows that the observed value of 1.2857 is in the right tail. Actually, there are four positive values equal to 1.2857 and one even larger, and there are four negative values

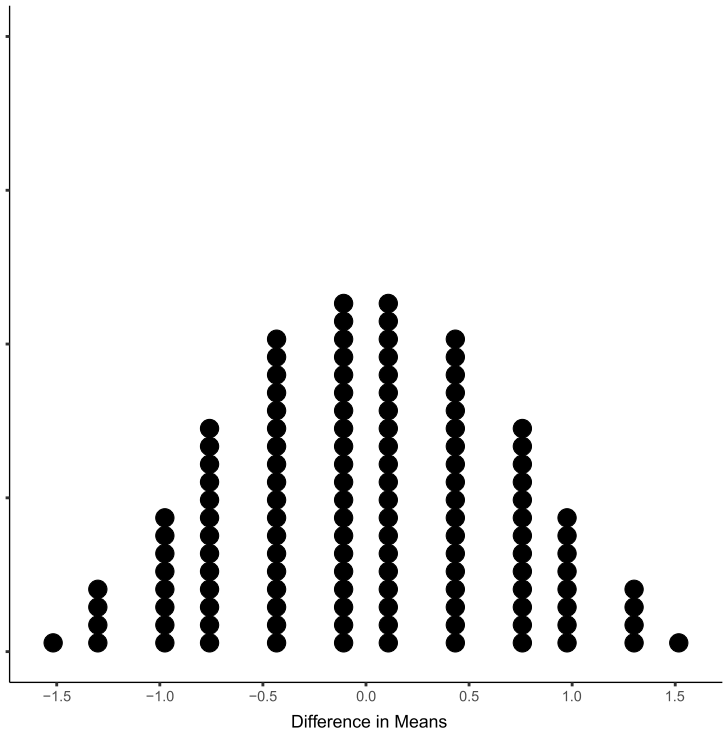

FIGURE 6.6 Dot plot of the randomization distribution for the difference in means between the E- and the C-scores in Figure 6.2 if an RBD was used

equal to −1.2857 and one even smaller. Therefore, the *p*-value is $10/128 = .0781$. If a traditional 5% significance level were used, this result would be labeled as "not statistically significant".

However, as was mentioned before, in design-based inference the data are analyzed in accordance with the way the data were collected. If the data were collected using a CRD, then the randomization test *p*-value would have been $154/3432 = .0449$. If the data were collected using a more prudent ATD with a maximum of two consecutive measurement occasions for the same level, then the randomization test *p*-value would have been $24/518 = .0463$.

One caveat is that *p*-values should always be interpreted cautiously. For a randomized single-case experiment, such as in this example, neither random sampling nor a population of participants is involved. Hence classical population inference is out of the question. The only inference that is possible is *causal inference* for the particular participant involved in the study. Furthermore, this causal inference is probabilistic and tentative. There is no 100% certainty that the sleeping pills are responsible for the increase in dizziness, and the results of

this experiment cannot guarantee that a reduction of the medication in the future will have the same beneficial effect as in the actual experiment.

Conclusion

In this chapter, a small sample size solution was provided by turning the problem upside down. Instead of looking at the population of participants from which only a small number of participants can be sampled, the starting point can be an intensive study of a single participant that can be repeated in other participants. In this perspective, a single participant is considered as the population and repeated measurements constitute the sample from that population. If there is more than one participant, then the results for the other participants are considered replications. In addition, the chapter demonstrated how randomized single-case experiments can be designed and analyzed. The resulting inference is about establishing a functional relation between a manipulated variable and an outcome variable for a single participant, which can be accomplished by using counterfactual reasoning, as illustrated with an example.

References

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

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

Bulté, I., & Onghena, P. (2009). Randomization tests for multiple baseline designs: An extension of the SCRT-R package. *Behavior Research Methods*, *41*(2), 477–485. doi:10.3758/BRM.41.2.477.

Bulté, I., & Onghena, P. (2013). The single-case data analysis package: Analysing single-case experiments with R software. *Journal of Modern Applied Statistical Methods*, *12* (2), 450–478.

De, T. K., Michiels, B., Vlaeyen, J. W. S., & Onghena, P. (2017). One by one: A Shiny web-app for the design and analysis of single-case experiments. Poster presented at the useR! annual conference, Brussels, Belgium. www.user2017.brussels/uploads/Tamal-Kumar-De_170623_103005.pdf.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, *33*, 503–513.

Gregoire, T. G. (1998). Design-based and model-based inference in survey sampling: Appreciating the difference. *Canadian Journal of Forest Research*, *28*(10), 1429–1447.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960. doi:10.1080/01621459.1986.10478354.

Jacobs, K. W. (2019). Replicability and randomization test logic in behavior analysis. *Journal of the Experimental Analysis of Behavior*, *111*(2), 329–341. doi:10.1002/jeab.501.

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

Kratochwill, T. R., Hitchcock, J. R., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). Single-case design technical documentation. What Works Clearinghouse. https://files.eric.ed.gov/fulltext/ED510743.pdf.

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

Maric, M., & Van der Werff, V. (2020). Single-case experimental designs in clinical intervention research. In R. van de Schoot and M. Miočević (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners*. Abingdon: Taylor & Francis.

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

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

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

Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *Clinical Journal of Pain*, *21*(1), 56–68. doi:10.1097/00002508-200501000-00007.

Onghena, P., Michiels, B., Jamshidi, L., Moeyaert, M., & Van Den Noortgate, W. (2018). One by one: Accumulating evidence by using meta-analytical procedures for single-case experiments. *Brain Impairment*, *19*(1), 33–58. doi:10.1017/BrImp.2017.25.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. 10.1126/science.aac4716.

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

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., … Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1* (3), 337–356. doi:10.1177/2515245917747646.

Smid, S. C., & Rosseel, Y. (2020). SEM with small samples: Two-step modeling and factor score regression versus Bayesian estimation with informative priors. In R. van de Schoot and M. Miočević (Eds.), *Small sample size solutions: A how to guide for applied researchers and practitioners*. Abingdon: Taylor & Francis.

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

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

Tyrell, P. N., Corey, P. N., Feldman, B. M., & Silverman, E. D. (2013). Increased statistical power with combined independent randomization tests used with multiple-baseline design. *Journal of Clinical Epidemiology*, *66*(6), 691–694. doi:10.1016/j.jclinepi.2012.11.006.

Van Den Noortgate, W., & Onghena, P. (2003). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments & Computers*, *35*(1), 1–10. doi:10.3758/BF03195492.