OPEN ACCESS

ESCOLA DE MEDICINA PUCRS

SCIENTIA MEDICA

Scientia Medica Porto Alegre, v. 31, p. 1-9, jan.-dez. 2021
e-ISSN: 1980-6108 | **ISSN-L:** 1806-5562

http://dx.doi.org/10.15448/1980-6108.2021.1.40128

SEÇÃO: EDUCATION IN HEALTH SCIENCES

Small numbers are an opportunity, not a problem

Números pequenos são uma oportunidade, não um problema

Jimmie Leppink[1]
orcid.org/0000-0002-8713-1374
j.leppink@gmail.com

Received: Feb, 12th, 2021
Aproved: Apr, 21th, 2021
Published: June. 30th. 2021.

Abstract

Aims: outcomes of research in education and training are partly a function of the context in which that study takes place, the questions we ask, and what is feasible. Many questions are about learning, which involves repeated measurements in a particular time window, and the practical context is usually such that offering an intervention to some but not to all learners does not make sense or is unethical. For quality assurance and other purposes, education and training centers may have very locally oriented questions that they seek to answer, such as whether an intervention can be considered effective in their context of small numbers of learners. While the rationale behind the design and outcomes of this kind of studies may be of interest to a much wider community, for example to study the transferability of findings to other contexts, people are often discouraged to report on the outcomes of such studies at conferences or in educational research journals. The aim of this paper is to counter that discouragement and instead encourage people to see small numbers as an opportunity instead of as a problem.

Method: a worked example of a parametric and a non-parametric method for this type of situation, using simulated data in the zero-cost Open Source statistical program R version 4.0.5.

Results: contrary to the non-parametric method, the parametric method can provide estimates of intervention effectiveness for the individual participant, account for trends in different phases of a study. However, the non-parametric method provides a solution in several situations where the parametric method should be used.

Conclusion: Given the costs of research, the lessons to be learned from research, and statistical methods available, small numbers should be considered an opportunity, not a problem.

Keywords: mixed model, percentage of all non-overlapping data bayes, single case design, single case experimental design, time series.

Resumo

Objetivo: os resultados da pesquisa em educação e treinamento são, em parte, uma função do contexto em que esse estudo ocorre, das perguntas que fazemos e do que é viável. Muitas perguntas são sobre a aprendizagem, que envolve medições repetidas em uma janela de tempo específica, e o contexto prático, geralmente, é tal, que oferecer uma intervenção a alguns, mas não a todos os alunos, não faz sentido ou é antiético. Para garantia de qualidade e outros propósitos, os centros de educação e treinamento podem ter perguntas orientadas localmente que procuram responder, como, por exemplo, se uma intervenção pode ser considerada eficaz em seu contexto de pequeno número de alunos. Embora a justificativa por trás do projeto e dos resultados deste tipo de estudos possa ser do interesse de uma comunidade muito mais ampla, por exemplo, para estudar a possibilidade de transferência de resultados para outros contextos, as pessoas são frequentemente desencorajadas a relatar os resultados de tais estudos em conferências ou em revistas de pesquisa educacional. O objetivo deste artigo é combater esse desânimo e, em vez disso, incentivar as pessoas a verem os pequenos números como uma oportunidade em vez de um problema.

Método: realizado um exemplo de método paramétrico e não paramétrico para este tipo de situação, utilizando dados simulados no programa estatístico Open

[1] University of York, York, North Yorkshire, United Kingdom.

Source R versão 4.0.5 de custo zero.

Resultados: ao contrário do método não paramétrico, o método paramétrico pode fornecer estimativas da eficácia da intervenção para o participante individual, levando em conta as tendências em diferentes fases de um estudo. No entanto, o método não paramétrico fornece uma solução em várias situações, onde o método paramétrico deve ser usado.

Conclusão: dados os custos da pesquisa, as lições a serem aprendidas com a pesquisa e os métodos estatísticos disponíveis, pequenos números devem ser considerados uma oportunidade, não um problema.

Palavras-chave: modelo misto, porcentagem de todos os dados bayes não sobrepostos, projeto de caso único, projeto experimental de caso único, séries temporais.

Abbreviations: PAND-B = percentage of all non-overlapping data Bayes; SCD = single case design; SCED = single case experimental design

Introduction

In research in education and training, at least where statistical analysis is involved, small numbers of participants are often considered a problem and a reason for either not carrying out a study that might yield important results or for not presenting the outcomes of a study carried out to a wider audience. This is unfortunate for several reasons. To start, even though virtually any study on education or training takes place in a particular context, the reasoning behind as well as the design and outcomes of a given study may have useful lessons to be learned for a much wider audience. In addition, for publication purposes and equally for internal quality assurance and accreditation purposes, we want education and training to be evidence based, and appropriately designed studies with numbers of participants large or small provide the best if not the only way to enable that. Finally, from an ethical and usually also financial and logistic perspective, research ought not be about getting ever larger numbers of people to participate in our studies; instead, a key principle should be to not use more resources than necessary. Some institutions and centers may have the luxury to involve hundreds of students in some of their research, whereas in other places the numbers are much smaller. For example, a surgical department in a hospital may have as few as three (i.e., N = 3) residents in a specific specialty in any period. Leitmotiv in conversations in such settings is that it is pointless to do research on the effectiveness of a type of intervention or component of a training otherwise let alone publish that research because the numbers are so small. However, even with such small numbers, the ability to establish evidence for a possible positive or negative effect of an intervention or component in such a setting may have tremendous implications for the growth of residents, for the delivery of healthcare and possibly patient outcomes. And when well documented, a report on the outcomes of a study in this setting – at a conference, in a journal, or otherwise – may help centers in a similar situation elsewhere to inform decision making in their own training program or possibly carry out the same or a very similar study in their own setting.

A simulated example

Suppose we find ourselves in a Colon and Rectal Surgery department in a hospital with seven new residents in training. The department has had a simulation training program in place, using a variety of methods to help residents develop the knowledge and skills they need to treat conditions in the colon, rectum and a number of other areas including the liver and reproductive systems. This year, the department is considering the introduction of virtual reality technology as a simulation method to see if the ability to view a complex structure and condition from different angles can facilitate simulation training performance. To start small, the department decide to introduce virtual reality halfway the training on liver surgery, a training that includes a total of ten performance measurements.

Probably not among the options to study this question is the classical randomized controlled experiment, where we would randomly allocate different residents to different groups, firstly because that kind of experiment would require substantially larger numbers of residents and secondly because in the context at hand not providing virtual reality probably makes no sense and could be questioned from an ethical perspective as well (i.e., why withhold a potentially effective intervention that could have positive outcomes for resident training, healthcare and patient outcomes?). However, the researchers can use a so-called *single case design* (SCD) or, in experimental form, *single case experimental design* (SCED) to study their question (e.g., (1)-(7). Characteristic of SCDs and SCEDs is that they involve small groups of participants, or even single participants (i.e., $N = 1$), that are measured repeatedly on the same outcome variable of interest (i.e., time series data). Contrary to a classical randomized controlled experiment, where the question is *if* a given participant receives treatment, the question in SCEDs is *when* a given participant receives a treatment. For instance, in the liver surgery training, if the starting point of the intervention can vary across participants and the starting point of the intervention (e.g., after three, after five or after seven measurements) is randomized, we are dealing with a form of SCED. In settings other than learning, an alternative form of SCED could be found in randomized *combinations* of intervention / no intervention for each of a series of trials, but in learning that is often not an option because learning at one point in time tends to carry over to next measurement occasions.

If variation in the starting point of an intervention is not considered feasible, for example because the seven residents in question take the training at the same time and it is considered important to give every resident five practice (i.e., measurement) occasions prior to and five occasions after the introduction of the intervention, we are dealing with a form of SCD that is sometimes also referred to as interrupted time series design but it is not a form of SCED. After all, in the latter case, although there is still a manipulation in the form of a *baseline* (i.e., prior to intervention) condition and an *intervention* condition, the moment when the intervention is introduced no longer varies between participants (i.e., it is after five occasions for all participants) and is not randomized. Nevertheless, the outcomes of this study can still provide useful insights for decision making in the department as well as for informing similar studies in other settings. Therefore, Table 1 presents a simulated example of what the data could look like for the seven hypothetical residents.

TABLE 1 – The (simulated) performance data of 7 students (ID #1-#7) before and during the intervention, with five measurement occasions within each phase (i.e., time_in_phase), and each occasion resulting in an integer performance score from 0 (min) to 10 (max)

Time	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Phase	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**1**	**1**
Time_in__phase	**4**	**3**	**2**	**1**	**0**	**4**	**3**	**2**	**1**	**0**
ID #1	2	3	2	3	3	6	7	8	7	8
ID #2	2	3	2	4	3	6	7	6	5	6
ID #3	3	2	4	3	3	5	7	6	7	8
ID #4	2	1	2	1	2	5	4	5	6	5
ID #5	2	2	3	2	2	5	4	5	4	5
ID #6	3	2	4	2	3	6	5	5	4	5
ID #7	1	3	2	4	3	5	7	6	8	7

Phase 0 = before intervention, Phase 1 = during the intervention.
Time_in_phase: 4, 3, 2, 1, 0 occasions prior to the last occasion in the given phase

The data matrix in **Table 1** presents the performance indicated by an integer score ranging from 0 (min) to 10 (max) for each of the seven residents for each of ten measurement occasions, with the first five measurement occasions being in the baseline phase (i.e., Phase 0) and the last five measurement occasions being in the intervention phase (i.e., Phase 1). Time in phase in Table 1 indicates the number of measurement occasions prior to the end of the phase.

A mixed model

To analyze these data, we need a method that can account for the fact that the seven residents times ten measurement occasions are not seventy independent observations but seven sets of correlated observations (also called a mixed model, e.g., (1), (7), with the correlation between occasions decreasing as time between occasions increases, and – since an intervention can have different effects for different residents – can be used for individual residents. Maric and Van der Werff (7) present a mixed regression model that does exactly that, using the *nlme* package (8) in the Open Source environment *R* (9). The outcomes of that model for the individual residents are presented in **Figure 1** and **Table 2** (*R* version used for this paper: 4.0.5).

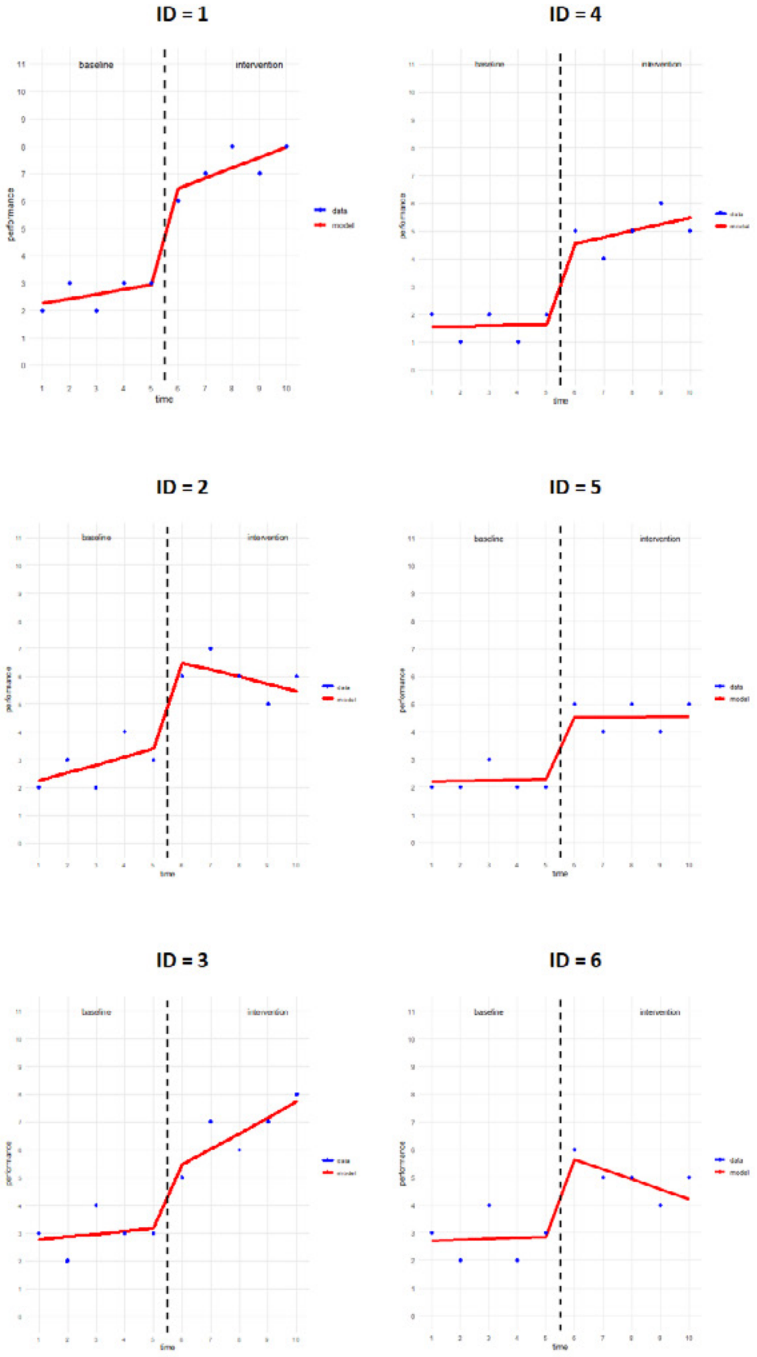

Figure 1 – Graphical presentation of the outcomes of mixed model analysis for the 7 students.

TABLE 2 – Intervention effectiveness results using a mixed model for N = 1: regression coefficient estimate (B), standard error (SE), p-value (p) and 95% confidence interval lower bound (LB) and upper bound (UB)

		B	***SE***	***p***	***LB***	***UB***
ID #1	b_0	2.941	0.369	< 0.001	2.039	3.844
	b_1	5.028	0.513	< 0.001	3.773	6.284
	b_2	-0.174	0.154	0.303	-0.551	0.204
	b_3	-0.205	0.211	0.369	-0.720	0.311
ID #2	b_0	3.383	0.412	< 0.001	2.375	4.391
	b_1	2.095	0.568	0.010	0.704	3.486
	b_2	-0.287	0.175	0.151	-0.715	0.140
	b_3	0.537	0.235	0.062	-0.037	1.111
ID #3	b_0	3.198	0.442	< 0.001	2.115	4.281
	b_1	4.532	0.612	< 0.001	3.034	6.030
	b_2	-0.109	0.187	0.581	-0.565	0.348
	b_3	-0.458	0.252	0.119	-1.075	0.159
ID #4	b_0	1.631	0.409	0.007	0.632	2.631
	b_1	3.864	0.566	< 0.001	2.480	5.249
	b_2	-0.024	0.172	0.895	-0.444	0.397
	b_3	-0.215	0.233	0.391	-0.785	0.354
ID #5	b_0	2.292	0.258	< 0.001	1.659	2.924
	b_1	2.253	0.355	< 0.001	1.383	3.122
	b_2	-0.022	0.111	0.847	-0.294	0.250
	b_3	0.018	0.147	0.905	-0.342	0.378
ID #6	b_0	2.856	0.314	< 0.001	2.088	3.623
	b_1	1.360	0.431	0.020	0.304	2.415
	b_2	-0.033	0.136	0.815	-0.366	0.300
	b_3	0.392	0.179	0.071	-0.047	0.830
ID #7	b_0	3.515	0.287	< 0.001	2.814	4.216
	b_1	4.311	0.395	< 0.001	3.344	5.277
	b_2	-0.422	0.126	0.015	-0.730	-0.115
	b_3	-0.155	0.165	0.382	-0.558	0.247

b_0 = intercept, b_1 = phase, b_2 = time in phase, b_3 = phase-by-time in phase interaction.

The model that is used consists of four regression coefficients (B_0-B_3):

$$\textit{performance at occasion } i = B_0 + (B_1 * \textit{phase at occasion } i) + (B_2 * \textit{time in phase at occasion } i) + (B_3 * \textit{phase-by-time-in-phase interaction}) + \text{residual.}$$

In plain language, these four coefficients mean the following:

- B_0: the model's score (i.e., red line in Figure 1) at the end of the baseline phase, which is in this case at occasion $i = 5$;
- B_1: the model's difference between the end of the intervention phase and the end of the baseline phase;
- B_2: the model's slope in the baseline phase (which given the coding is negative when scores go up in the baseline phase, and vice versa); and
- B_3: the difference between the model's slopes for baseline and intervention (in statistical terms, the interaction effect).

Thus, in Table 2, the outcomes of B_2 and B_0 respectively indicate the linear trend throughout the baseline phase and the extent to which performance at the end of the baseline phase differs from zero. While for outcome variables where '0' outcomes are either unlikely or impossible B_0 is only necessary for getting the model right, when dealing with outcomes where '0' or negative outcomes are well possible interpreting B_0 may be useful as well. Further, while both B_2 and B_0 are needed in the model in order to get the interpretation of intervention effects right, B_1 and B_3 are our indicators of intervention effects.

For the seven residents under study, we find a statistically significant baseline trend (B_2) in only one case (resident #7) but no statistically significant interaction effects (B_3). However, the difference between end-of-intervention (i.e., occasion $i = 10$) and end-of-baseline (i.e., occasion $i = 5$) performance (B_1) is statistically significant for all seven residents, albeit slightly weaker for residents #2 and #6 (where performance seems slightly better earlier instead of later in the intervention phase) than for the other residents. In other words, the intervention has had a positive effect of some kind for all seven residents.

Combining outcomes from different individuals

While a great feature of the presented mixed model is that through its application to the individual it can help us understand differences between individuals in response to an intervention, keeping other factors constant smaller numbers of observations come with lower statistical power than larger numbers of observations. In addition, the number of coefficients like in Table 2 quickly increases with increasing numbers of participants in a study and it can be very useful to combine the outcomes of different participants in a meta-analysis. For example, a meta-analysis using restricted maximum likelihood estimation (10) provides a useful way to obtain an overall estimate for a given regression coefficient with a 95% confidence interval around it. Doing so for B_1, B_2 and B_3, using the meta-analysis module in JASP (11), treating the individual resident as study unit, we find the following results.

For B_1, we find an estimate of 3.329 with a 95% confidence interval of [2.269; 4.389], indicating a clearly positive gain from end-of-baseline to end-of-intervention in line with what we already saw in Figure 1 and Table 2. For B_2, which represents the baseline trend, we find an estimate of -0.153 and a 95% confidence interval of [-0.284; -0.022]. This interval does not include zero and given the coding of time in phase a negative difference indicates lower scores earlier than later in the baseline phase. In other words, although at the level of the individual we find a statistically significant outcome for B_2 only for resident #7, at group level we find a statistically significant baseline trend, which makes sense in the context of repeated practice probably resulting in some increase in knowledge or skill. Finally, for B_3, which denotes the interaction effect, we find an estimate of -0.003 with a 95% confidence interval of [-0.249; 0.242], in short nicely around zero.

And what if the proposed mixed model does not work?

Apart from its applicability to individual data, an important strength of the presented mixed model is that it accounts for baseline trends (through B_2) and different trends between phases (through B_3) and can be extended to more than two phases if more than two phases are present, for instance in a study with more than one intervention. However, one requirement for this model to work is that we deal with scale outcome variables such as the integer performance score in the example. When we deal with dichotomous outcomes (e.g., correct vs. incorrect, or checked vs. not checked), multicategory nominal outcomes (e.g., different options for subjective choice) or ordinal outcomes (e.g., performance judged as poor, satisfactory or good), this model will not work. In addition, very clear non-linear trends (e.g., a U-shape or inverted U-shape) also pose a threat to the validity of the model and with numbers of observations as small as in the example at hand (which is quite common in educational settings) a more complex model accounting for that trend are unlikely to be an option. For such situations, there is a non-parametric alternative that in cases where the presented mixed model does work is inferior to the mixed model because it does not account for trends in phases but does not assume linear trends or require scale outcome variables: the Bayesian percentage of all non-overlapping data (PAND-B) (12). Succinctly put, PAND-B is a function of how many data points for a given participant would need to be moved from one phase to another in order to achieve perfect non-overlap of data (PAND (13)) and uses a Bayesian Binomial prior distribution to avoid extreme estimates (in particular '0' or '100' percent) based on very small numbers and to provide an interval estimate that cannot be provided when using PAND (for the rationale behind this non-parametric Bayesian method an example of its application with dichotomous outcomes, see (12)).

For the seven residents at hand, the only resident that does not have perfect non-overlap of data is resident #6, where we would need to move the '4' score in the baseline phase to the intervention phase – or move the '4' score in the intervention phase to the baseline phase – in order to achieve perfect non-overlap. In other words, we have 10 successes out of 10 measurement occasions for all residents except for resident #6 where we have 9 successes out of 10 measurement occasions. For each resident, PAND-B works as follows:

$$\text{Prior} + \text{Data} = \text{Posterior}.$$

For resident #6, this means:

$$\text{Beta}(1,1) + \text{Beta}(9,1) = \text{Posterior}(10,2).$$

This corresponds with a posterior median of 0.852 and a 95% posterior interval of [0.587; 0.977]. This interval completely exceeds 0.5 and therefore the intervention can be considered effective for resident #6. For the other six residents, we find:

$$\text{Beta}(1,1) + \text{Beta}(10,0) = \text{Posterior}(11,1).$$

This corresponds with a posterior median of 0.939 and a 95% posterior interval of [0.715; 0.998], again indicative of the intervention being effective. If the outcomes were less clear than in the current example, the outcomes of different residents could be combined into an overall estimate accounting for the time series data structure (for an example, see (12)). Finally, to provide an estimate of the proportion of residents for which this intervention could be effective, we can use the same Binomial procedure:

$$\text{Beta}(1,1) + \text{Beta}(7,0) = \text{Posterior}(8,1).$$

In this formula, Beta(7,0) comes from the intervention having an effect for all seven residents. The resulting posterior median is 0.917 and the 95% posterior interval is [0.631; 0.997]. This posterior can be updated with new residents participating in a study (i.e., this posterior becomes the prior for the next study).

To conclude: with the right approach, small numbers are not a problem

The mixed model presented in this paper can provide estimates of intervention effectiveness for the individual even where numbers of measurements per individual are relatively small, accounting for trends in different phases of a study. For scale outcome variables where numbers are too small to model non-linear trends and the data does not show clear non-linear trends that would render linear models useless (e.g., U or inverted U, which in the context of human learning is unlikely), the mixed model is stronger than non-parametric alternatives. However, PAND-B provides a possible non-parametric alternative to the mixed for model where the mixed model falls short (i.e., dichotomous, multicategory nominal or ordinal outcomes, or clearly non-linear trends in scale outcomes) and can also be considered in addition to the mixed model, albeit that where both the mixed model and PAND-B work the mixed model is more powerful (i.e., PAND-B comes with a higher risk of failing to detect an intervention effect) and better in the face of linear trends.

Either way, in addition to research questions and a study design that make sense in the context at hand, we have a statistical solution for data acquired in an SCD or SCED. Where measuring the same individuals repeatedly over time is logical and feasible, such as in many settings involving learning, a major advantage of SCDs and SCEDs over classical group comparison studies is that the numbers of participants required to obtain meaningful estimates of intervention effects are much smaller. Research ought not be solely about getting as many participants as possible participate in our studies; rather, we should design our studies such that they make sense in the context at hand, help to address meaningful research questions and use no more resources than needed. Given the costs of research, the lessons to be learned from research, and statistical tools available to combine findings from different studies (including studies of *N* = 1) small numbers should be considered an opportunity, not a problem.

References

1. Leppink J. The art of modelling the learning process: Uniting educational research and practice. Cham: Springer; 2020. https://doi.org/10.1007/978-3-030-43082-5

2. Michiels B, Heyvaert M, Meulders A, Onghena P. Confidence intervals for single-case effect size measures based on randomization test inversion. Behav Res Meth. 2017;49:363-81. https://doi.org/10.3758/s13428-016-0714-4

3. Michiels B, Onghena P. Randomized single-case AB phase designs: prospects and pitfalls. Behav Res Meth. 2018;51:2454-76. https://doi.org/10.3758/s13428-018-1084-x

4. Parker RI, Hagan-Burke S, Vannest KJ. Percentage of all non-overlapping data (PAND): An alternative to PND. J Spec Educ. 2007;40:194-204. https://doi.org/10.1177/00224669070400040101

5. Pérez-Fuster P, Sevilla J, Herrera G. Enhancing daily living skills in four adults with autism spectrum disorder through an embodied digital technology-mediated intervention. Res Aut Spect Dis. 2019;58:54-67. https://doi.org/10.1016/j.rasd.2018.08.006

6. Tanious R, De TK, Onghena P. A multiple randomization testing procedure for level, trend, variability, overlap, immediacy, and consistency in single-case phase designs. Behav Res Therap. 2019;119:103414. https://doi.org/10.1016/j.brat.2019.103414

7. Maric M, Van der Werff V. Single-case experimental designs in clinical intervention research. In: R Van de Schoot & M Milocević, Small sample size solutions: A guide for applied researchers and practitioners. OAPEN Home; 2020. p. 102-11. https://library.oapen.org/bitstream/handle/20.500.12657/22385/9780367221898_text%20(1).pdf?sequence=1#page=116

8. Pinheiro J, Bates D, DebRoy S, Sarkar D, Team RC. nlme: Linear and nonlinear mixed effects models. R Package Ver. 2013;3:111.

9. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna: R Foundation for Statistical Computing (version 4.0.5); 2021 March 31 [cited 2021 May 6]. Available from: https://www.r-project.org

10. Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. J Educ Behav Stat. 2005;30:261-93. https://doi.org/10.3102/10769986030003261

11. Love J, Selker R, Marsman M, et al. JASP version 0.14.1.0 [Internet]; 2020 Dec 17 [cited 2021 May 6]. Available from: https://jasp-stats.org

12. Leppink J. Statistics for N = 1: A non-parametric Bayesian approach. Scientia Med. 2020;30:1-10. https://doi.org/10.15448/1980-6108.2020.1.38066

13. Parker RI, Hagan-Burke S, Vannest KJ. Percentage of all non-overlapping data (PAND): An alternative to PND. J Spec Educ. 2007;40:194-204. https://doi.org/10.1177/00224669070400040101

Jimmie Leppink

PhD in Statistics Education, LLM in Forensics, Criminology and Law, and MSc in Psychology and Law from Maastricht University, the Netherlands; MSc in Statistics from Catholic University of Leuven, Belgium; currently Senior Lecturer in Medical Education and Director of Assessment at Hull York Medical School, University of York, United Kingdom.

Mailing address

Jimmie Leppink

University of York, Heslington

York, Y010 5DD

Os textos deste artigo foram revisados/conferidos pela Poá Comunicaçãoe submetidos para validação do(s) autor(es) antes da publicação.