OPEN ACCESS

ESCOLA DE MEDICINA
PUCRS

SCIENTIA MEDICA

Scientia Medica Porto Alegre, v. 30, p. 1-10, jan.-dez. 2020
e-ISSN: 1980-6108 | ISSN-L: 1806-5562

http://dx.doi.org/10.15448/1980-6108.2020.1.38066

EDUCATION IN HEALTH SCIENCES

Statistics for N = 1: A Non-Parametric Bayesian Approach

Estatística para N = 1: uma abordagem Bayesiana não paramétrica

Jimmie Leppink[1]
orcid.org/0000-0002-8713-1374
hyjl17@hyms.ac.uk

Received on: May.11th, 2020.
Approved on: Oct. 03rd, 2020.

Abstract: Research in education is often associated with comparing group averages and linear relations in sufficiently large samples and evidence-based practice is about using the outcomes of that research in the practice of education. However, there are questions that are important for the practice of education that cannot really be addressed by comparisons of group averages and linear relations, no matter how large the samples. Besides, different types of constraints including logistic, financial, and ethical ones may make larger-sample research unfeasible or at least questionable. What has remained less known in many fields is that there are study designs and statistical methods for research involving small samples or even individuals that allow us to address questions of importance for the practice of education. This article discusses one type of such situations and provides a simple coherent statistical approach that provides point and interval estimates of differences of interest regardless of the type of the outcome variable and that is of use in other types of studies involving large samples, small samples, and single individuals.

Keywords: 95% Credible Interval; Percentage of All Non-Overlapping Data (PAND); Percentage of All Non-Overlapping Data Bayes (PAND-B); Single Case Design (SCD); Single Case Experimental Design (SCED).

Resumo: A pesquisa em educação é frequentemente associada à comparação de médias de grupo e relações lineares em amostras suficientemente grandes, e a prática baseada em evidências trata do uso dos resultados dessa pesquisa na prática educacional. No entanto, há questões importantes para a prática da educação que não podem ser realmente abordadas por comparações de médias de grupo e relações lineares, por maiores que sejam as amostras. Além disso, diferentes tipos de restrições, incluindo as logísticas, financeiras e éticas, podem tornar a pesquisa com amostras maiores inviável ou, pelo menos, questionável. O que tem ficado menos conhecido em muitos campos é que existem desenhos de estudos e métodos estatísticos para pesquisas envolvendo pequenas amostras ou mesmo indivíduos que nos permitem abordar questões de importância para a prática da educação. Este artigo discute um tipo de tais situações e fornece uma abordagem estatística coerente simples que fornece estimativas de ponto e intervalo de diferenças de interesse, independentemente do tipo de variável de resultado e que é útil em outros tipos de estudos envolvendo grandes amostras, pequenas amostras, e indivíduos solteiros.

Palavras-chave: Intervalo de credibilidade de 95%; Porcentagem de todos os dados não sobrepostos (PAND); Porcentagem de todos os Bayes de dados não sobrepostos (PAND-B); Projeto de caixa única (SCD); Projeto Experimental de Caso Único (SCED).

ABBREVIATIONS: PAND, percentage of all non-overlapping data; PAND-B, PAND Bayes; PAND-BC, PAND-B corrected; SCD, Single Case Design; SCED, Single Case Experimental Design.

[1] University of York, York, North Yorkshire (NY), United Kingdom

Introduction

Research in education most commonly involves samples of participants in actual education or artificial (e.g., laboratory) settings and its outcomes are generalized far beyond the samples studied. Whether we deal with a survey study that focuses on motivation to learn, a randomized controlled experiment that compares effects of different types of instruction on learning or a study aimed at developing an assessment tool, we generally deal with groups of participants for which linear relations between variables of interest are calculated or, where different groups are available, these different groups are compared in terms of average scores on some outcome variables of interest. Even in studies where no numbers appear to be involved (e.g., qualitative judgments from interviews, focus groups or observations), the abstract and discussion section of the resulting papers often clearly imply a wide generalizability (or as some 'qualitativists' prefer to name it: transferability). In some fields of education, researchers studying small groups or even individuals resort to qualitative methods partly because there is a common belief that quantitative methods are mainly or even only about linear relations and average comparisons in large samples. However, quantitative methods can be used for all kinds of relations and for samples as small as one single individual or subject (i.e., *N* = 1). Examples of clearly non-linear patterns include the evolvement of a pandemic like COVID-19 and changes in stock markets and price trajectories of many goods and services over time.

An example of a non-linear pattern with N = 1: Gold prices

Figure 1 presents an example of a clearly non-linear pattern of the gold price in American Dollars ($) in a five-year period (18 April 2015 – 18 April 2020) measured once a week, on Saturdays, when the gold price does not move because markets are closed.

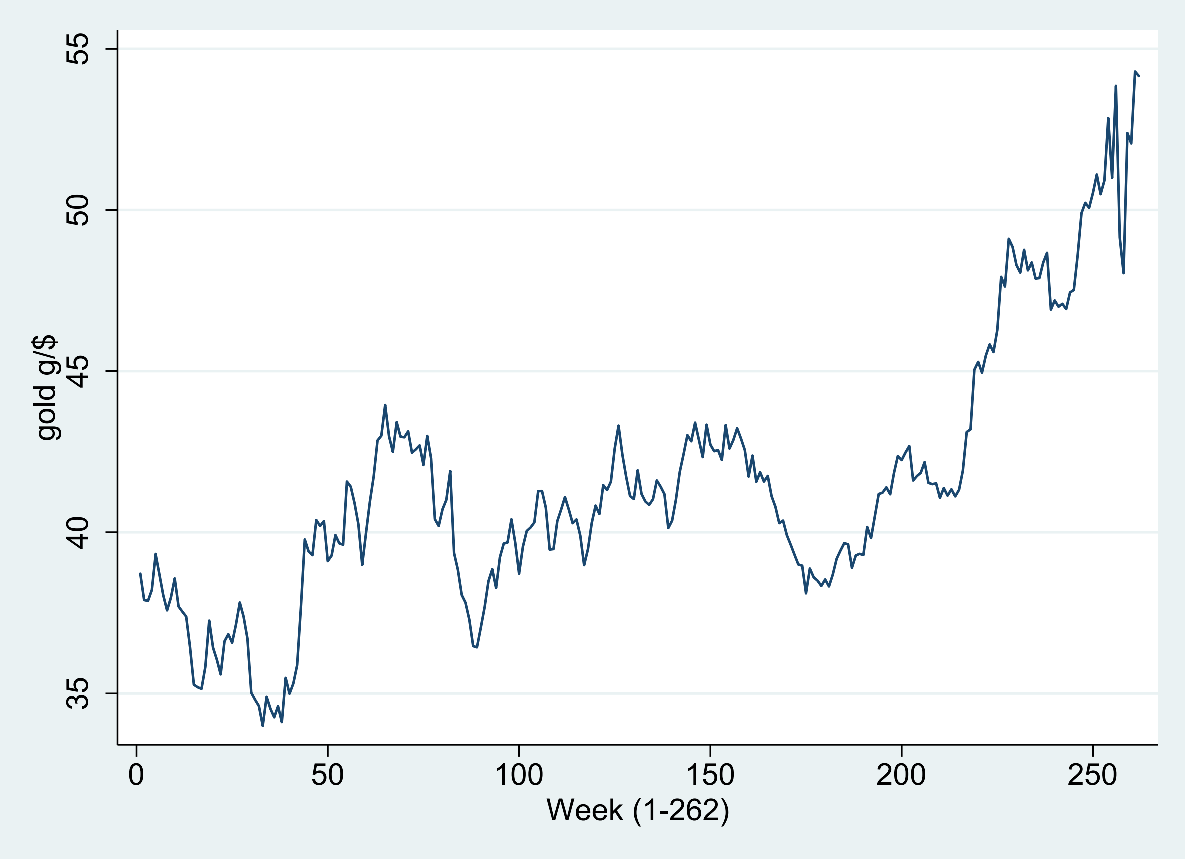

Figure 1 – Price of one gram of gold in US Dollars ($) from 18 April 2015 (week 1) until 18 April 2020 (week 262).

Whether we express the gold price in $, in Euros (€), in British Pounds (£) or in another currency, gold prices tend to gradually go up in times of inflation and may peak substantially after major regional or global geopolitical, financial or health-related events that contribute to uncertainty, even more so when multiple events come together (e.g., COVID-19 paralyzing the economy in many countries and postponing crucial negotiations for trade deals such as between the United States and China or between the United Kingdom and the European Union). Although a change in price after a specific event does not imply a causal relation between that event and the price change, it is known that investment tends to move away from uncertainty. With significant events such as war, large-scale economic downturn and/or a pandemic, many stocks may (temporarily) go down, except for stocks in specific sectors that gain importance in such times, but gold tends to go up (and may come down substantially once the geopolitical, financial or health storm lays down). While it is impossible to tell exactly where the gold price is headed in the short term (e.g., a few weeks from now) let alone in the medium to long term, statistical time-series methods (e.g., [1-2]) that use the information of historical gold prices, combined with knowledge of important regional or global geopolitical, financial or health-related events coming up can help us to predict to some extent where gold and other prices are headed, at least in the short run.

The remainder of this article

For simplicity, the example in Figure 1 uses gold prices observed on a weekly basis, but where we have gold prices on a daily or even hourly basis, we can use the same time-series methods and knowledge to make forecasts about next hours or days. Further, where learning (or behavior) among humans or animals is concerned and many carefully timed measurements are available, we can use the same time-series methods, in combination with learning theory, to model, understand, and predict future learning (or behavior) [3].

In hardly any practical education setting, we will be able to collect hundreds of measurements about learning or behavior of interest from the same individuals. However, even if numbers of measurements per individual are much smaller (e.g., fifteen or twenty), with the right study designs, we can use statistical methods to address questions that matter for practice and research in education. This article discusses one type of such situations and provides a simple coherent statistical approach that provides point and interval estimates of differences of interest regardless of the type of the outcome variable and that is of use in other types of studies involving large samples, small samples, and single individuals as well.

Questions and constraints drive methodological choices

Some readers may wonder why not just aim for large-sample experiments and quasi-experiments. After all, randomized controlled experiments have, at least in some fields, been considered a kind of 'gold standard' and quasi-experiments a kind of second-best alternative. However, while larger-sample experiments and quasi-experiments can certainly address a wide range of questions that are relevant and important for research and practice in education, there are questions that cannot really be addressed with such studies or can be addressed more efficiently with studies using smaller samples. Larger-sample experiments and quasi-experiments are typically intended to address questions of relevance to a wide range of research and practical settings. However, from a practical perspective, a common question is whether a specific type of instruction, assessment or intervention is effective for a given individual or small group of individuals, and average comparisons from large samples may not adequately address that question. Besides, random sampling, and in the case of experiments also random allocation to the available conditions, is important in large-sample research, but what if that is not an option? And what if logistic and financial constraints researchers and practitioners often deal with do not allow for large-sample research? Or what if withholding a treatment is considered unethical? This is where studies using a single case design (SCD) or, in experimental form, a single case experimental design (SCED), come in (e.g., [3-7]). There are many different types of SC(E)Ds and, as for larger-scale experiments and quasi-experiments, which type is to be considered depends on the question(s) asked. A full overview of all possible SC(E)Ds is beyond the scope of this article, but some common ones are discussed in the next sections.

Interrupted time series

A first common type of SCD is found in so-called *interrupted time-series designs*: studies where two or more sequences of measurements or observations are carried out. Perhaps the simplest form is found in two sequences equal in number of measurements. For example, in a classroom, students may have 10 practice trials with linear algebra and then face 10 testing trials with linear algebra. Alternatively, in a weight loss program, a client's weight may be registered weekly prior to intervention A for a total of 20 times and then for a total of 20 times after intervention A. To return to education, suppose that practitioners who developed a six-week online training to deliver education during COVID-19 lockdown discover that statistics of daily study time in the first three weeks of the training are not as high as anticipated and therefore decide to make a small change hoping that the statistics in the second half of the training will indicate an increased study time. Given that a combination of factors may contribute to a difference in algebra performance, weight or study time between the phases, we cannot just interpret that difference as a causal link between the manipulated change and more (or less) favourable outcomes. However, if we are pragmatic and interested in achieving better outcomes regardless of causal inference, any sufficiently substantial change in weight or study time for the better may be worth the change.

Experimental designs

If additional to achieving better outcomes causal inference *is* of interest, we need stronger types of SCDs, and these can be found in a range of SCEDs. SCEDs are sometimes mistaken as non-experimental research, because most researchers associate experiments with random allocation of a sufficiently large random sample to control and treatment conditions. However, like in larger-sample experiments, there is *manipulation* (e.g., before and after the intervention in the weight loss program or in the COVID-19 online training). In addition, *randomness* is found in the *timing* of treatment, that is: while in many larger-sample experiments the question is *if* a given participant receives treatment, in SCEDs the question is *when* a participant receives treatment, and that 'when' is the result of some form of random allocation. In the weight loss or online training example, for instance, we could let the timing of (the start of) the intervention be a result of randomness. Whether we have 15 clients signing up for our weight loss intervention or we have 15 students taking our online training in times of COVID-19, we can *randomize the start* of the intervention across participants.

For some types of outcomes, an alternative to randomizing the start of the treatment can be found in randomizing the *occurrence* of treatment for each individual trial or randomizing blocks of treatment-no-treatment sequences to different individuals. For instance, in a study on online learning of statistics where the interest lies in comparing a traditional text-only (i.e., no treatment) and innovative infographic (i.e., treatment) condition in terms of the time needed to complete the section where the information is provided, and the online training requires students to complete a total of 40 short (i.e., a few minutes) sections, we may choose a study design where the occurrence of one format (treatment) vs. the other (traditional, no treatment) is randomized either for each trial or for blocks of trials. Although randomizing for each trial yields many more possible sequences than randomizing for blocks of trials, a problem with randomizing for each trial is that it includes sequences with few or no observations in one of the two conditions as well as sequences in which most or all observations of one condition are grouped together. Randomizing for blocks of trials avoids such grouping together and allows us to have equal numbers of observations for each condition. For instance, if we divide the 40 trials (i.e., the 40 short sections) into 20 blocks of 2 trials and use two possible random orders for each block – AB and BA – we have $2^{20} = 1048576$ possible sequences, in which at most two consecutive measurements are from the same condition (i.e., two times A or two times B in a row).

From questions to study designs and statistical methods

Which type of SC(E)D should be used largely depends on the question(s) of interest as well as on the nature of the phenomenon studied. For instance, while the type of randomized block designs with rapidly alternating AB/BA sequences may be useful in many health-related settings, it is problematic in many settings where learning takes place, because treatment received at one stage may influence outcomes of many if not all later measurements in more than one way even if treatment is not continued, and in most practical education settings withdrawing a potentially effective intervention is considered both unusual and unethical.

As for larger-sample experiments and quasi-experiments, which statistical methods to use depends on the question of interest, the type of design used, the level of measurement of the outcome variables, and distributional features of the outcomes. However, regardless of the design and outcomes of a given study, findings from individuals can be combined into models using data from groups of individuals as in larger-sample experiments and quasi-experiments and meta-analytic studies combining outcomes of different studies [3, 8].

Example: improved task performance after an intervention?

To revisit our COVID-19 online training example, suppose we have a new cohort of 6 health science students completing a training on statistical inference that comprises a total of 40 short sessions of about 15 minutes each, each of which involves studying a piece of information followed by completing a short task that results in correct (1) or incorrect (0) result. Each of the 40 sessions focuses on slightly different content, but the difficulty level of each of the assignments is such that historically they have resulted in about 40% correct response. Students have complained that the tasks are difficult, and the training developers have developed alternative versions of the last 20 tasks (i.e., the second half of the 40 sessions) that present the same questions on the same content but with an additional instructional support in the form of a brief explanation what is expected from the student in the task at hand. Before developing alternatives for the other 20 sessions as well, the training developers want to run the training with this new cohort of 6 students, who will complete the first 20 sessions in the usual format and the subsequent 20 sessions in the new format. This is an example of an interrupted time-series design SCD with 20 measurements in a baseline condition (A) and 20 measurements in a treatment condition (B). It is *not* an SCED because there is no randomization of the start or occurrence of treatment; instead, treatment starts at the same time for all 6 students. **Table 1** presents the task performance outcomes for each of the 6 students in each of the two conditions.

TABLE 1 – Performance per trial (0 = incorrect, 1 = correct) and per condition (%) for the 6 students in the online training study.

Student	First half (A)		Second half (B)	
	Per trial	%	Per trial	%
#1	0 0 1 0 0 0 1 0 1 0 1 1 0 0 1 0 0 0 1 0	35	0 1 0 1 0 1 1 0 0 1 0 1 0 1 0 1 0 1 0 0	45
#2	0 1 0 0 1 0 1 0 0 1 0 1 0 1 0 0 1 0 0 1	40	0 1 0 1 0 0 0 1 1 0 1 0 1 0 0 1 0 0 0 1	40
#3	1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 1	40	0 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 1 0 0	30
#4	0 0 0 1 0 0 0 1 0 1 0 1 0 1 1 0 0 1 0 0	35	1 0 1 0 1 1 0 0 1 1 0 1 1 0 1 0 1 1 1 1	65
#5	0 0 1 0 0 1 0 0 0 1 0 1 0 0 0 1 0 1 0 0	30	1 1 1 0 1 1 0 1 1 1 1 0 1 1 1 1 0 1 1 1	80
#6	0 1 0 0 1 0 1 0 0 0 1 0 1 0 0 0 1 1 0 0	35	1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1	95

Percentages such as in Table 1, or mean or median differences when dealing with scale or multicategory ordinal outcome variables, can be interpreted as effect size estimates at the level of the individual. While these simple measures do not account for possible trends in the baseline phase (in the case of dichotomous outcome variables, for example more correct performance towards the end), with small numbers of observations such as in the example at hand such trends are difficult to estimate, at least at the level of the individual, and like in larger-sample studies, one should not interpret statistics without graph and/or table inspection anyway. Some might prefer an effect size statistic such as Cohen's *d* [9], Pearson's correlation *r* or in squared form R^2, but these statistics still do not adjust for baseline trends, can be quite sensitive to outliers, and assume independent residuals while in time-series data residuals tend to be correlated (e.g., [1-3]).

Non-parametric point and interval estimates for individual treatment effects

There are non-parametric approaches to the effectiveness of interventions that reduce the problem of correlated residuals, including the percentage of all non-overlapping data (PAND, a better alternative) [10], and a Bayesian modification of the percentage of all non-overlapping data (PAND-B, where B stands for 'Bayes') which applies a correction for smaller samples and provides a 95% credible interval as an interval estimate (i.e., none of the other two overlap approaches provides interval estimates which makes statistical testing difficult) [3]. While none of these non-parametric approaches resolve the potential problem of baseline trends, combined with graph and/or table inspection they do enable researchers to draw tentative conclusions regarding the effectiveness of interventions for individuals as well as for groups of individuals [3, 10].

The rationale behind PAND and PAND-B is to determine how many observations (scores) would need to be 'swapped' between conditions in order to have 100% separation (i.e., no overlap at all). In addition, given that the effect of an intervention can often be either positive or negative, a more precise way to conceptualize PAND and PAND-B is how many percent of the observations are in favor of the treatment. When dealing with scale outcome variables such as counts, time, and quantitative performance, this can be done through careful visual inspection of time-series graphs (for an example, see Chapter 16 in Leppink [3]), and for all types of outcome variables two-way tables can be used. In our example, where we have a baseline condition (A) of 20 observations followed by a treatment condition (B) of 20 observations, all '0' in the baseline condition and all '1' in the treatment condition are observations in favor of the treatment, while all '1' in the baseline condition and all '0' in the treatment condition are observations pointing against the treatment. PAND and PAND-B are functions of these two types of outcomes (i.e., in favor vs. against the treatment). **Table 2** summarizes the outcomes for each of the 6 students and for the group in our example study.

TABLE 2 – PAND and PAND-B outcomes for each of the 6 students and for the group of 6 students in our online training example study.

	Counts		PAND	PAND-B		
Student	**Favor**	**Against**	**Point**	**Point**	**95% LB**	**95% UB**
#1	22	18	0.550	0.548	0.397	0.693
#2	20	20	0.500	0.500	0.351	0.649
#3	18	22	0.450	0.452	0.307	0.603
#4	26	14	0.650	0.645	0.494	0.779
#5	30	10	0.750	0.742	0.597	0.858
#6	32	8	0.800	0.790	0.651	0.894
All, U	148	92	0.617	0.616	0.554	0.676
All, C	44.835	27.870	0.617	0.615	0.501	0.720

All', U, uncorrected; All', C = intraclass correlation (0.059) corrected; 95% UB = 95% credible interval upper bound; 95% LB, 95% credible interval lower bound.

PAND is found by dividing the frequency of observations in favor of the treatment by the total number of observations, in this case 40 for the individual student. PAND values of 0.55, 0.65 and 0.75 correspond with 10%, 30%, and 50% difference in performance in favor of the treatment condition, and values of 0.45, 0.35 and 0.25 correspond with 10%, 30%, and 50% difference in performance in favor of the baseline condition.

Although traditionally PAND has been treated as a variable with a range of [0.5, 1] with 0.5 indicating no treatment effect and 1 being the maximum possible value, given that a treatment can be positive or negative, the range of PAND is actually quite a different one and depends on the outcome variable of interest. For scale outcome variables, PAND-values can in theory range from (almost) 0 to (almost) 1, a value of 0.5 indicates no effect, values below 0.5 indicate negative effects (i.e., the outcome being worse in the treatment condition) and values above 0.5 indicate positive effects. When dealing with dichotomous outcome variables, more positive outcomes in the treatment condition result in a higher lower bound of PAND while more positive outcomes in the baseline condition result in a lower upper bound of PAND. For example, in the study at hand, 30% positive outcomes in the baseline condition implies a maximum possible PAND value (i.e., upper bound) of 0.85. Likewise, for multicategory nominal outcome variables, where the interest lies for example in a change in choice from a three or more options the ordering of which is arbitrary, the proportion of observations resulting in a given option in the baseline condition influences the maximum possible shift away from that option in the treatment condition. Finally, for multicategory ordinal outcome variables, the PAND-range is influenced by the proportion of outcomes in the best category in the baseline condition (lower upper bound) and the proportion of outcomes in the best category in the treatment condition (higher lower bound). For dichotomous and scale outcome variables, PAND values of 0.5 indicate no effect; for multicategory nominal and ordinal outcome variables, in designs where the baseline and treatment condition differ in length, PAND sometimes cannot be exactly 0.5 and the outcomes nearest to 0.5 (i.e., on either side of 0.5) are to be interpreted as no effect.

PAND-B differs from PAND in that it uses a Beta(1,1) *prior* distribution that is updated with the data coming in to obtain a Beta *posterior* distribution [3]:

Prior + Data = Posterior.

For instance, for Student #1, the data is Beta(22,18), and therefore the posterior is Beta(23,19):

Beta(1,1) + Beta(22,18) = Beta(23,19).

The Beta(23,19) distribution has a posterior median (point estimate) of 0.548 and a 95% credible interval of [0.397; 0.693]. The more data, the more the posterior median approaches the PAND estimate; in small samples, it is slightly pulled towards 0.5 to avoid ridiculous estimates like 0% or 100% based on very small numbers of observations. Another advantage of PAND-B over PAND is that it comes with an interval estimate, in the form of the 95% credible interval, which can be used for hypothesis testing as follows. To start, intervals excluding [0; 0.5] indicate positive effects, while intervals excluding [0.5; 1] indicate negative effects. In the example study, we find sufficient evidence in favor of a treatment effect only for Student #5 and Student #6. Further, if in a specific context we only consider differences of at least 10% of practical importance, the PAND-B region of [0.45; 0.55] indicates differences that are not of practical importance, and a difference of practical importance can be concluded if the 95% credible interval excludes [0.45; 0.55]. In the example study, we find sufficient evidence in favor of such a practically important treatment effect for Student #5 and Student #6, in both cases in favor of the treatment (i.e., positive treatment effect).

From individual to group

PAND-B estimates for individuals can be combined to obtain group estimates, and we can correct for the intraclass correlation due to the same individuals being measured repeatedly and some individuals performing better than others. As PAND-B does not deal with the actual scores or order of measurements, the intraclass correlation is much lower than in time-series models. However, there is still *some* intraclass correlation, due to which simply summing the frequencies in favor of and against the treatment of the different individuals (i.e., 'All, U' in Table 2) yields numbers for the group larger than the effective sample size, and a correction factor is needed to correct the numbers for group downward [3]. Given *k* number of measurements per individual, the correction factor is:

Correction factor = 1 + [(*k* – 1) * intraclass correlation].

We can estimate the intraclass correlation in a two-level logistic regression model which treats student (upper) and observation (lower) as hierarchical levels, includes the student-level intercept as random effect and condition as fixed effect. The outcome variable is a dichotomous variable, with for each observation either '0' (against treatment) or '1' (in favor of treatment). For the data at hand, we find an intraclass correlation coefficient of 0.059. Given *k* = 40 (i.e., 40 observations per individual), the correction factor is 3.301. This explains the numbers of 44.835 and 27.870 in 'All, C' in Table 2 and the resulting 95% credible interval being slightly wider than the interval in 'All, U' which incorrectly assumes zero intraclass correlation. The estimates in 'All, C' in Table 2 are also referred to as PAND-BC, where C stands for 'corrected'. As new cohorts of students come in, more observations become available, intraclass correlation estimates become more accurate, and 95% credible intervals become smaller. Although in large samples and in SCDs involving much larger numbers of measurements than in the example study discussed in this article time-series methods that account for baseline (and other trends) will normally provide more powerful methods (e.g., [3]), PAND-B and PAND-BC – both point and interval estimates – can be used for any sample size from very large to as small as a single individual (N = 1).

The need for combining individual outcomes into group outcomes depends on the setting and question of interest. Where the question of interest is whether a given treatment is effective for a *given individual* – which is a legitimate question in many practical education settings – the PAND-B point estimate and 95% credible interval of the individual at hand provide the outcome, and there is no need to merge outcomes from different individuals. Besides, while in the example study at hand it may make sense to treat the 6 students as one cohort or group, where different individuals undergo very different procedures in potentially quite distinct settings, it may make little sense to combine findings from different individuals into single 'group' estimates even if the (type of) treatment is the same across individuals. Finally, in any case, two other summarizing statistics for the effectiveness of an intervention at group level are the frequency and/or proportion of individuals for which we find sufficient evidence for a treatment effect (or for a treatment effect of practical importance if that is the question of interest). For the example study, that number of students is 2, and that corresponds with 33%. These statistics do not require any kind of correction for intraclass correlation and have another attractive feature: the Bayesian updating procedure with Beta-distributions applies to these statistics as well. Suppose, for example, that we deal with a cohort of 35 students and we find sufficient evidence for a positive treatment effect for 25 students. Using a Beta(1,1) prior distribution, this results in a Beta(26,11) posterior distribution, which yields a point estimate of 0.706 and a 95% credible interval of [0.548; 0.837]. This interval indicates that we have sufficient evidence to assume that the treatment works for more than 50% of the individuals (i.e., it exceeds [0; 0.5]).

To conclude: a consistent non-parametric approach to individual treatment effects

Although when dealing with much larger numbers of measurements and/or much larger samples of individuals, more powerful parametric methods should be used, PAND-B provides a point estimate and a 95% credible interval that can be used to answer questions regarding the effectiveness of an intervention for any given individual under study as well as for a group of individuals. Of course, there is no free lunch; as for any statistical method, detecting treatment effects of interest is more difficult with small numbers of observations than with larger numbers of observations. For instance, with 40 measurements (cf. the example study), any number of observations in favor of the treatment for an individual of 27 (67.5%) or higher results in a 95% credible interval completely above 0.5; with 50 measurements, the minimum number in favor of the treatment needed for an individual is 32 (64%), while with 30 measurements that number is 21 (70%) and with 20 measurements that number is 15 (75%). Given that the PAND-B procedure is the same for all possible outcome variables, these numbers are no different for scale than for categorical outcome variables. It is important to keep this in mind when planning your study. With relatively strong treatment effects, it may be fairly easy to achieve such high percentages, but with somewhat weaker treatment effects which may still have important implications for practice lower percentages are likely and 95% credible intervals are then more likely to include 0.5 unless the number of observations is increased.

While research in education is often associated with linear relations and average comparisons in large samples, study designs and statistical methods for research involving individuals and small samples are available, and when dealing with specific practical questions or facing logistic, financial and/or ethical constraints, well-designed studies involving individuals or small samples may be more appropriate than larger-sample studies. Although larger-sample experiments and quasi-experiments will most probably remain valuable for research and practice in education, SCDs and SCEDs can help to address both questions concerning what works for larger groups and what likely does or does not work for specific individuals and can as such help to bridge possible gaps between education research and practice.

Notes

Funding

This study did not receive financial support from external sources.

Conflicts of interest disclosure

The author declares no competing interests relevant to the content of this study.

Author contributions

The author declares to have made substantial contributions to the conception, or design, or acquisition, or analysis, or interpretation of data; and drafting the work or revising it critically for important intellectual content; and to approve the version to be published.

Availability of data and responsibility for the results

The author declares to have had full access to the available data and they assume full responsibility for the integrity of these results

References

1. Box GEP, Jenkins GM, Reinsel GC. Time series analysis: Forecasting and control. 3.rd ed. Upper Saddle River, NJ: Prentice Hall; 1994.

2. Brockwell PJ, Davis RA. Time series: Theory and methods. 2.nd ed. New York: Springer; 2009.

3. Leppink J. The art of modelling the learning process: Uniting educational research and practice. [Internet]. Cham: Springer; 2020 [citado 2020 Nov 13]. Disponível em: https://doi.org/10.1007/978-3-030-43082-5

4. Michiels B, Heyvaert M, Meulders A, Onghena P. Confidence intervals for single-case effect size measures based on randomization test inversion. Behav Res Meth [Internet]. 2017 [citado 2020 Nov 13];49:363-81. Disponível em: https://doi.org/10.3758/s13428-016-0714-4

5. Michiels B, Onghena P. Randomized single-case AB phase designs: Prospects and pitfalls. Behav Res Meth [Internet]. 2018 [citado 2020 Nov 13];51:2454-76. Disponível em: https://doi.org/10.3758/s13428-018-1084-x

6. 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 [Internet]. 2019 [citado 2020 Nov 13];58:54-67. Disponível em: https://doi.org/10.1016/j.rasd.2018.08.006

7. 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 [Internet]. 2019 [citado 2020 Nov 13];119:103414. Disponível em: https://doi.org/10.1016/j.brat.2019.103414

8. Van de Schoot R, Milocević M. Small sample size solutions: A guide for applied researchers and practitioners [Internet]. OAPEN Home; 2020. [citado 2020 Nov 13]. Disponível em: http://library.oapen.org/handle/20.500.12657/22385

9. Cohen J. Statistical power analysis for the behavioural sciences. New York: Routledge; 1988.

10. Parker RI, Hagan-Burke S, Vannest KJ. Percentage of all non-overlapping data (PAND): An alternative to PND. J Spec Educ [Internet]. 2007 [citado 2020 Nov 13];40:194-204. Disponível em: 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, Y010 5DD, York

North Yorkshire, NY, United Kingdom