BRAIN IMPAIRMENT VOLUME 19 NUMBER 1 MARCH pp. 33–58 © Australasian Society for the Study of Brain Impairment 2017
doi:10.1017/BrImp.2017.25

One by One: Accumulating Evidence by using Meta-Analytical Procedures for Single-Case Experiments

Single-case research

Patrick Onghena,[1] Bart Michiels,[1] Laleh Jamshidi,[1] Mariola Moeyaert[2] and Wim Van den Noortgate[1]

[1] *Faculty of Psychology and Educational Sciences, KU Leuven, University of Leuven, Leuven, Belgium*
[2] *Department of Educational and Counseling Psychology, Division of Educational Psychology and Methodology, School of Education, University at Albany, SUNY, Albany, USA*

This paper presents a unilevel and multilevel approach for the analysis and meta-analysis of single-case experiments (SCEs). We propose a definition of SCEs and derive the specific features of SCEs' data that have to be taken into account when analysing and meta-analysing SCEs. We discuss multilevel models of increasing complexity and propose alternative and complementary techniques based on probability combining and randomisation test wrapping. The proposed techniques are demonstrated with real-life data and corresponding R code.

Keywords: single-case experiment, single-case experimental design, *N*-of-1 trial, meta-analysis, multilevel model, probability combining, randomisation test, permutation test

Many handbooks and courses in research methods and statistics seem to imply that groups have to be compared in order to conduct proper scientific research in the health sciences (see e.g., Jacobson, 2017; Moore, McCabe, & Craig, 2017; Peat, 2002). In those handbooks and courses, treatment effects are inferred by applying treatments to one or more experimental groups and by comparing their results to one or more control groups. The mantra is: 'Take a random sample of patients and randomly assign these patients to treatment and control'. In medicine, this mantra is consolidated in the randomised controlled trial being the gold standard (Kaptchuk, 2001; Sackett, Rosenberg, Gray, Haynes, & Richardson, 1996; Turner et al., 2012).

The practical disadvantages and ethical concerns of this group-comparison approach are well documented (Carey & Stiles, 2016), but more importantly, questions regarding the epistemological status of the results, and therefore also the clinical relevance, can be raised (Molenaar, 2004; Onghena, 2007; Onghena & Edgington, 2005). Intra-patient variability differs fundamentally from inter-patient variability, and consequently there is a fundamental difference between the meaning of a treatment effect in group-comparison studies and the meaning of a treatment effect for an individual patient (Molenaar & Campbell, 2009; Velicer & Molenaar, 2013). The reckless generalisation of group-comparison treatment effects to individual-patient treatment effects can be considered as a prime example of the well-known ecological fallacy (Harrington & Velicer, 2015; Onghena, 2016). Group-comparison treatment effects are not necessarily representative of individual-patient treatment effects; it might well be that the average pattern of group differences does not apply to any single patient involved in the study (Barlow, Nock, & Hersen, 2009; Gast & Ledford, 2014; Kazdin, 2011). As Sidman (1952) already observed: 'It

Address for correspondence: Patrick Onghena, Faculty of Psychology and Educational Sciences, KU Leuven, University of Leuven, Tiensestraat 102, BE-3000 Leuven, Belgium. E-mail: patrick.onghena@kuleuven.be

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press

appears, then, that when different groups of subjects are used to obtain the points determining a functional relation, the mean curve does not provide the information necessary to make statements concerning the function for the individual' (p. 268).

The alternative approach that we discuss in this paper turns the group-comparison model upside down. Instead of focusing on group-comparison treatment effects, leaving us with the impossibility to generalise to individual-patient treatment effects, we start by testing the treatment effect in an individual patient and by investigating the replicability in other patients afterwards. As Shapiro (1966) put it:

> It seems, therefore, that when one investigates change one cannot combine data from a number of individuals unless he knows that the differences among the curves obtained from the individuals are due to experimental error. The first step in the investigation of processes must logically consist of investigation in a number of individual cases. (p. 5)

This approach has a closer resemblance to clinical practice than the group-comparison approach: Patients enter a hospital or clinical trial one by one, and clinical care is aimed at individual patients. This closer resemblance of the individualised approach to actual clinical practice may also foster the generalisation of the results to this practical setting (i.e., increase the ecological validity of the study results) (Onghena & Edgington, 2005; Tate, Aird, & Taylor, 2013).

Statistical models for the group-comparison approach and for the analysis of RCTs are well-established (see e.g., Jacobson, 2017; Peat, 2002). In the present paper, we want to propose a generic model for the analysis of data from replicated experiments with individual patients or, more generally, from replicated '*single-case experiments*' (SCEs) as they are called in the methodological literature. We will try to reach this aim by

- Examining the specific features of SCEs that preclude the mere adoption of the common statistical techniques known from the group-comparison approach;
- Looking for a statistical model that takes these features into account;
- Proposing the use of meta-analytical procedures for the aggregation of replicated SCEs.

We will demonstrate our proposal with real-life data and apply some alternative and complementary approaches on the same data.

Definition of Single-Case Experiment

If we want a generic model for the analysis and meta-analysis of SCEs, then it is crucial to be crystal clear about the kind of data we are dealing with. We need to know exactly what the experimental units are, which units are considered in the statistical analysis, and which units are considered in the meta-analysis. In this first section, we reflect on the SCE definition and comment on the essential ingredients and their implications.

SCEs are defined as 'experiments in which one entity is observed repeatedly during a certain period of time, under different levels ('treatments') of at least one independent variable' (Onghena, 2005, p. 1850). There are two essential features in this definition. First, only one entity is involved (it is a single case), and second, there is a manipulation of the independent variable(s) (it is an experiment). Two other features are consequences of this set-up: the entity is exposed to all levels of the independent variable (like in a within-subject design) and there are repeated observations or measurements (like in a longitudinal or time series design).

For the purpose of this paper it is important to reflect on the main terms and implications of this definition. We summarise our reflection in seven comments:

1. The definition refers to the generic term '*entity*' to emphasise that any experimental unit at any level of aggregation may be involved. In the health sciences, the entity usually is a patient but the entity can also be a group of patients in a hospital ward, a part of a patient (e.g., a brain region), or a more abstract unit, such as a rehabilitation centre as an organisation (Barlow, Nock, & Hersen, 2009; Kazdin, 2011; Onghena, 2005).
2. The definition and the following discussion of statistical techniques are restricted to '*experiments*'. By referring to one or more manipulated variables, SCEs are distinct from qualitative and descriptive case studies and from observational time series studies (Onghena & Edgington, 2005; Onghena & Struyve, 2015).
3. Notice that the '*observed repeatedly*' in the definition does not need to refer to measurements of quantitative attributes. Also qualitative and naturalistic observations are possible without losing the experimental nature of the SCE. However, if the data have to be analysed statistically, all observations have to be quantitatively coded in one or more outcome variables. In the remainder of this article, we assume that measurement of a quantitative attribute is at the basis of at least one outcome variable. This is

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press

the most characteristic way that SCEs are presented, and is most feasible with the rising popularity of health care wearables and ecological momentary assessment (McDonald & Davidson, 2016; Piwek, Ellis, Andrews, & Joinson, 2016).

4. The internal and statistical-conclusion validity of an SCE can be strengthened by including randomisation in the design (Dugard, File, & Todman, 2011; Edgington, 1996; Heyvaert, Wendt, Van den Noortgate, & Onghena, 2015; Kratochwill & Levin, 2010), but randomisation is not a necessary feature in the SCE definition. Randomisation can be performed with respect to the treatment order or the moments of phase change or a combination of both (Levin, Ferron, & Gafurov, 2014; Onghena, Vlaeyen, & de Jong, 2007).
5. External validity and generalisability to other patients is not always the first aim of an SCE (Barlow et al., 2009; Gast & Ledford, 2014; Kazdin, 2011). In a purely idiographic approach, the clinician might just be interested to understand and help the single patient under investigation, and may use a customised treatment and a customised symptoms checklist. However if generalisability is an aim, it can be accomplished by replications using identical (or similar) treatments and outcome measures. Replicated demonstrations of a similar treatment effect in consecutive patients provide evidence for an identical underlying process or a general law.
6. Many experimental design schedules are possible for SCEs. A useful classification has been proposed by Shamseer et al. (2015), Tate et al. (2016a, 2016b), and Vohra et al. (2015). They distinguish:
   - Withdrawal or reversal designs (such as ABA, ABAB, ABACAD with each letter representing a phase of repeated measurements for the same level of the manipulated variable),
   - Alternating treatments designs
   - Multiple baseline designs, and
   - Changing criterion designs.

   A two-phase AB design is not included in this classification, although the use of this design in an experimental context would match the definition of an SCE. Our proposal would be to consider studies using an AB design as SCEs, but to acknowledge that this design is the poorest design in terms of methodological quality. Two-phase AB designs suffer from the fact that there is only one demonstration of the treatment effect and that effect might be due to an external event accidentally coinciding with the moment of phase change (Kratochwill et al., 2010, 2013; Tate et al., 2013). This should not mean that applied studies using an AB design are, by definition, poorest in terms of methodological quality. SCEs using an AB design may involve methodological add-ons to be able to provide solid evidence, such as a randomly timed intervention, a large number of reliable observations, and replications across patients to obtain sufficient statistical power (Barlow et al., 2009; Heyvaert et al., 2017; Heyvaert, Wendt, Van den Noortgate, & Onghena, 2015).
7. In the medical literature, the term '*N-of-1 trial*' has been adopted to refer to a prospectively planned, multiple crossover trial in a single patient to determine the effect of an intervention (Guyatt, Jaeschke, & McGinn, 2002; Schork, 2015; Shamseer et al., 2015; Vohra et al., 2015). *N*-of-1 trials are '*biomedical*' SCEs, usually using a withdrawal or reversal design. Furthermore, *N*-of-1 trials often evaluate pharmacological treatments and include placebo control, randomisation, and blinding (Onghena, 2007, 2016). Consequently *N*-of-1 trials can be considered a subset of SCEs (Tate et al., 2016a, 2016b).

An Example

An example of a replicated SCE with two patients was published in this journal by Douglas, Knox, De Maio, and Bridge (2015). They used an AB design with follow-up to evaluate the effectiveness of a new treatment, Communication-specific Coping Intervention (CommCope-I), which targets coping in the context of communication breakdown after traumatic brain injury. The study was carried out with two patients: Samantha,[1] a 30-year-old woman, and Thomas, a 34-year-old man, who both had sustained severe traumatic brain injury several years before, and still experienced substantial communication problems. The primary outcome variable in this study was the Discourse Coping Scale – Clinician Rating, which uses a 10-cm visual analogue scale to measure the appropriate use of communication-specific coping strategies.

The raw data of the first A phase and the B phase are shown in Tables 1 and 2 for Samantha and Thomas respectively. As can be seen in the tables all scores in the B phase are larger than the scores in the A phase and the authors conclude that: 'This study provides sound phase-1 evidence for the effectiveness of CommCope-I'. (p. 190).

[1] Pseudonyms are used to protect privacy.

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press

TABLE 1
Weekly Scores on the Discourse Coping Scale – Clinician Rating for Samantha During a 9-Week SCE with a 3-Week Baseline (A phase) and a 6-Week Communication-Specific Coping Intervention (B phase)

Week	1	2	3	4	5	6	7	8	9
Phase	A	A	A	B	B	B	B	B	B
Score	4.5	5.0	4.8	6.3	8.2	7.1	5.8	8.0	9.7

Note. There was one missing week and data point in the A phase.

TABLE 2
Weekly Scores on the Discourse Coping Scale – Clinician Rating for Thomas During a 9-Week SCE with a 4-Week Baseline (A phase) and a 6-Week Communication-Specific Coping Intervention (B phase)

Week	1	2	3	4	5	6	7	8	9	10
Phase	A	A	A	A	B	B	B	B	B	B
Score	5.4	4.1	5.7	5.3	9.0	7.2	8.3	9.3	9.8	9.8

In the remainder of this article we will use the data in Tables 1 and 2 for a simple numerical demonstration of the statistical analysis and meta-analysis of SCE data. In most applications, the data are more numerous, complicated, diverse or fragmented and we will reflect on a few of these complicating issues (and their solutions) in the Discussion section. Furthermore, Table 1 and Table 2 contain two simplifications compared to the original data in Douglas et al. (2015):

1. We ignored the missing data point in the A phase of Samantha because dealing with missing data is a separate and complicated issue that would distract from the main ideas in the demonstration.
2. Douglas et al. (2015) referred to their design as 'an ABA design' (p. 194), but as the authors acknowledged: 'The withdrawal design is somewhat limited in the case of skill-based interventions like CommCope-I which are designed to facilitate lasting change in behaviour and are thus not strictly amenable to evaluation using return to baseline single-case designs'. (p. 199) If the data in the follow-up phase were modelled as '*a return to baseline phase*' then the results would be opposite to what is expected. Consequently, for both Samantha and Thomas, we did not include the follow-up data to avoid this interpretational pitfall.

We will also return to the implications of these simplifications and refer to some statistical options to deal with missing data and withdrawal designs in the Discussion section. The Appendix presents R code for all computations that are used in this manuscript.

A Simple Statistical Model for the AB Design: Model 1

Taking into account the essential and implied features of the SCE definition, a simple statistical model for data from a single patient considers the patient as the population and the repeated measurements as a random sample (i.e., realisations of a random outcome variable). In an AB design, the manipulated variable takes two values: one value for the baseline phase and one value for the treatment phase.

Given these features, some authors have been analysing data from SCEs using the following simple statistical model (Gentile, Roden, & Klein, 1972; Shine & Bower, 1971):

$$\text{Score}_t = \beta_0 + \beta_1(\text{Treat})_t + \varepsilon_t \qquad (1)$$

with

- Score_t referring to the value of the outcome variable at time t; the outcome variable Score indicates the repeated measurements, with the index t going from 1 to N (for N repeated measurements available from the SCE),
- $(\text{Treat})_t$, referring to the manipulated variable, taking a value of 0 for each measurement occasion in the baseline phase and a value of 1 for each measurement occasion in the treatment phase,
- β_0 referring to the baseline mean (i.e., the value of Score_t when $(\text{Treat})_t$ and ε_t equal 0),

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press

TABLE 3
Summary Statistics for the Samantha Data in Table 1

Phase	*N*	*M*	*SD*
A	3	4.77	0.25
B	6	7.52	1.42

β_1 referring to the difference between the treatment mean and the baseline mean (i.e., the treatment effect),

ε_t referring to the residual at time t, which is the deviation of Score_t from the average value of the outcome variable in the corresponding phase, $\varepsilon_t = \text{Score}_t - [\beta_0 + \beta_1(\text{Treat})_t]$.

The values for Score_t are observed by conducting the study and the values of $(\text{Treat})_t$ are determined by the chosen design; the values for β_0 and β_1 have to be derived following a particular optimisation criterion. The classical optimisation criterion is the minimisation of the squared residuals ε_t^2, resulting in a so-called '*Ordinary Least Squares*' (OLS) regression (or a common Analysis of Variance, using slightly different notation).

For the data of Samantha in Table 1, we have N equal to 9, and the means and standard deviations as shown in Table 3.

For Model 1 we obtain $\beta_0 = 4.77$ and $\beta_1 = 2.75$ (see the first two columns of Table 4), so there is an average addition of 2.75 points for the scores in the B phase as compared to the A phase, which can also be verified in Table 3 ($4.77 + 2.75 = 7.52$). Notice that no distributional assumptions are needed for the derivation of the OLS estimates.

A distributional model is usually invoked if questions like: 'Is this value of $\beta_1 = 2.75$ statistically significantly different from zero?' or 'What is the 95% confidence interval for this treatment effect estimate?' need an answer. The most common model arises if the ε_t values are assumed to be independent and identically distributed Gaussian deviates with a mean of 0 and an unknown variance σ_ε^2 ($\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$). Under this assumption, the treatment effect for the Samantha data is considered statistically significant, $\beta_1 = 2.75$, $t(7) = 3.22$, $p = .0146$, with [0.73; 4.77] being the 95% confidence interval.[2] Notice that the same t- and p-values would be obtained if a classical two-sample Student t test was performed or an equivalent F ($= t^2$) and identical p-value for an Analysis of Variance with two independent groups.[3] Notice also that the statistical test for the β_0 parameter is usually not relevant because it is testing a hypothesis that is irrelevant for most applications, namely the null hypothesis that the baseline mean is zero, H_0: $\beta_0 = 0$.

The assumption with regard to the residuals can be relaxed by only assuming that the residuals are exchangeable, so without specifying any population distributional shape. In that case an exact regression permutation test can be performed (Ottoboni, Lewis, & Salmaso, 2017; Pesarin & Salmaso, 2010; Wheeler & Torchiano, 2016), with a result that is very similar to the Gaussian theory test, *perm* $t(7) = 4.037$, $p = .0238$. This value can easily be verified because there is no overlap between the scores in the two phases. In that case, the observed test statistic provides the most extreme value and the exact one-sided permutation test p-value can be obtained by inverting the binomial coefficient $\binom{N}{n}^{-1}$. The exact two-sided permutation test p-value can be obtained by doubling this value. For the Samantha data $2/\binom{9}{3} = .0238$. This procedure is equivalent to the procedure in a two-sample permutation test (Edgington & Onghena, 2007; Fay, 2010), so also for this test the same p-value of .0238 would be obtained.

Modelling Trend: Model 2

Equation (1) is overly simplistic because it only models a difference between the A phase average and the B phase average. By contrast, an SCE always involves a time component and usually there is an evolution in time within the phases, a so-called '*trend*' (Shadish & Sullivan, 2011; Solomon, 2014). Therefore, a more appropriate statistical model for time series data includes a parameter β_2 to capture this trend:

$$\text{Score}_t = \beta_0 + \beta_1(\text{Treat})_t + \beta_2(\text{Time})_t + \varepsilon_t \qquad (2)$$

[2] At this point in the text, the attentive reader might already wonder whether this assumption is plausible and whether it is allowed to apply such simple t-procedures, known from traditional linear regression analysis, to time series data. This reader is referred to the next sections in the text, which present more complex models, taking trend and autocorrelation into account.

[3] Notice that the model is not equivalent to using the A and B data as dependent samples. A model for the A and B data as dependent samples is not uniquely specified because an SCE does not uniquely link an A phase data point to a corresponding B phase data point, as is the case in a matched-pairs design or a randomized block design. An SCE contains time series data and therefore the link (or dependency) resides between consecutive measurements rather than between measurements from separated phases.

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press

TABLE 4

Parameter Estimates (and Standard Errors) of the Models Specified in Equations 1–3 for the Samantha Data in Table 1

Parameter	Model 1	Model 2	Model 3
β_0	4.77** (0.70)	3.96** (0.79)	3.96** (0.66)
β_1	2.75 (0.85)	0.94 (1.34)	0.94 (1.11)
β_2		0.40 (0.25)	0.40 (0.21)
φ			0.004 (0.37)

Note. *$p < .05$. **$p < .01$.

with

- Score_t and $(\text{Treat})_t$ identical as in Equation 1,
- $(\text{Time})_t$ referring to the value of the time variable at time t; for the models in this manuscript the time variable takes the same values as the index t (viz, from 1 to N for Week 1 to Week N), but for more sophisticated models other values can be used,[4]
- β_0 referring to the intercept, that is the expected value of the outcome variable Score when $(\text{Treat})_t$ and $(\text{Time})_t$ are zero,
- β_1 referring to the change in the expected value of the outcome variable Score when going from the baseline phase to the treatment phase (i.e., the additive treatment effect),
- β_2 referring to the change in the expected value of the outcome variable for each unit change in $(\text{Time})_t$ (i.e., the trend), and
- ε_t referring to the residual at time t, which is the deviation of Score_t from the best fitting equation at time t, $\varepsilon_t = \text{Score}_t - [\beta_0 + \beta_1(\text{Treat})_t + \beta_2(\text{Time})_t]$.

A trend can most conveniently be spotted by making a time series plot of the data. Figure 1 shows the time series plot of the Samantha data. There seems to be a steady increase in Samantha's communication from Week 1 to Week 9, but it is not clear whether that increase is statistically significant. The parameter estimates and their standard errors for this Equation (2) can be found in Table 4, in the column with '*Model 2*' as the header.

If we assume that $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$ then the additive treatment effect for the Samantha data is not considered statistically significant anymore, $\beta_1 = 0.94$, $t(6) = 0.70$, $p = .5109$, with $[-2.35; 4.22]$ being the 95% confidence interval after taking the linear trend into account. Also the trend itself is not statistically significant, $\beta_2 = 0.40$, $t(6) = 1.64$, $p = .1518$, with $[-0.20; 1.00]$ being the 95% confidence interval. The permutation test p-values are .5050 and .1550 (see Wheeler & Torchiano, 2016, and the Appendix for the calculation of the permutation test p-values).

Notice that the β_1 estimate in Equation (2), 0.94, is smaller than the β_1 estimate in Equation (1), 2.75. The reason for this difference is that Equation (2) also includes a trend parameter β_2 and that in the estimation all components of the equation are taken into account. The linear trend estimate in this example indicates that there is a score increase of $\beta_2 = 0.40$ for each unit of change in the time variable, even if there is no additive treatment effect. For the A phase, $(\text{Treat})_t = 0$, this means that the expected scores are: $3.96 + (1)(0.40) = 4.36$, $3.96 + (2)(0.40) = 4.76$, and $3.96 + (3)(0.40) = 5.16$. The average of these three numbers is 4.76, which is the A phase average (within rounding error) that can be found in Table 3. For the B phase, $(\text{Treat})_t = 1$, the expected scores from Equation (2) become: $3.96 + 0.94 + (4)(0.40) = 6.5$, $3.96 + 0.94 + (5)(0.40) = 6.9$, $3.96 + 0.94 + (6)(0.40) = 7.3$, $3.96 + 0.94 + (7)(0.40) = 7.7$, $3.96 + 0.94 + (8)(0.40) = 8.1$, and $3.96 + 0.94 + (9)(0.40) = 8.5$. The average of these six numbers is 7.5, which is the B phase average (within rounding error) that can be found in Table 3. This demonstrates that the phase averages following Equation (2) are constructed by adding three components: an intercept, a time trend, and an additive treatment effect. Consequently, the β_1 estimate in Equation (2) should be interpreted differently than the β_1 estimate in Equation (1). In Equation (2) the β_1 estimate does not represent the raw difference between the phase averages but rather represents the score that is added in the B phase after taking the linear trend into account.

Another important difference between the results using Equation (2) as opposed to the results using Equation (1) is that the treatment effect has become statistically non-significant. This means that the treatment effect cannot be distinguished from the mere increment of the scores as a

[4] See Huitema and McKean (2000) and Anumendem, De Fraine, Onghena, and Van Damme (2013) for the options and implications.

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press

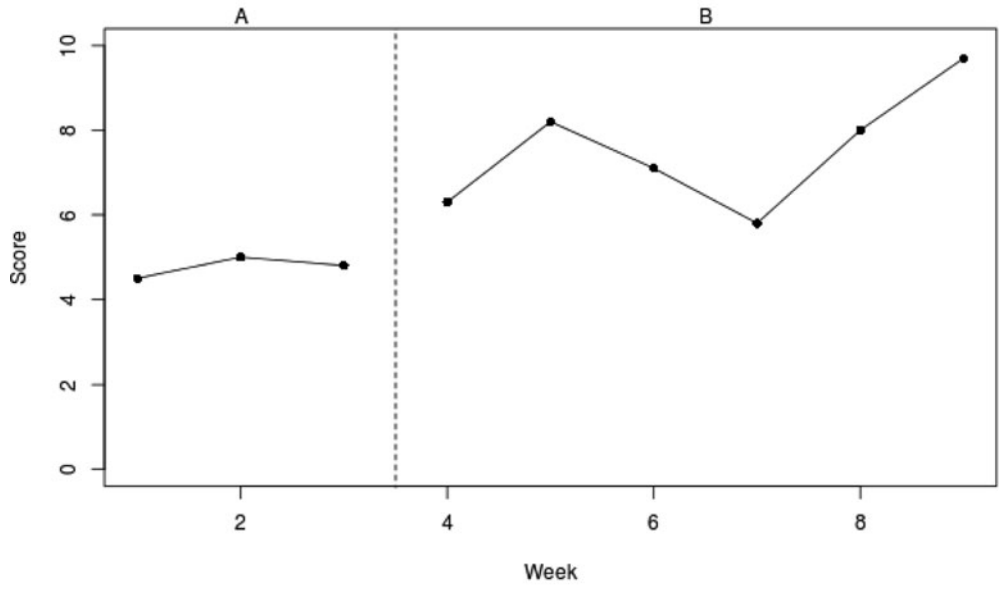


FIGURE 1

Weekly scores on the Discourse Coping Scale – Clinician Rating for Samantha during a nine-week SCE with a three-week baseline (A phase) and a 6-week Communication-specific Coping Intervention (B phase).

function of time. Although Figure 1 may give the impression that there is an additive treatment effect in addition to a small trend, none of these effects is large enough to be statistically significant. The observed data using this design and this (small) number of repeated measures are not convincing enough to attribute the larger average in the B phase to either an additive treatment effect or to a trend.

Modelling Serial Dependency: Model 3

A simplifying assumption that we made in the previous analyses is that the residuals are independent or exchangeable, given that all relevant parameters are included in the model. This assumption is not obvious for time series data (Harrington & Velicer, 2015; Jones, Weinrot, & Vaught, 1978) and therefore Equation (2) is sometimes extended using an autoregressive parameter for the residuals:

$$\begin{aligned} &\text{Score}_t = \beta_0 + \beta_1(\text{Treat})_t + \beta_2(\text{Time})_t + \varepsilon_t, \\ &\text{with the residuals } \varepsilon_t = \varphi\varepsilon_{t-1} + \omega_t \text{ and} \\ &\omega_t \sim N\left(0, \sigma_\omega^2\right), \end{aligned} \tag{3}$$

and with all the other parameters as defined in Equation (2).

The results are shown in Table 4, in the column with 'Model 3' as the header. The autoregressive parameter appears to be very small, and the results for Model 3 are essentially identical to the results for Model 2.

More Complex Statistical Models for the AB Design

It is possible to extend Equations (2) and (3) in several ways. An elaboration of these extensions is beyond the scope of the present article, but we give some references for the reader who is interested to explore these extensions.

A first possible extension is the inclusion of an interaction term $(\text{Treat})_t \times (\text{Time})_t$ to assess whether the trend changes from the baseline to the treatment phase. Issues regarding specification of the linear model, coding of the variables and interpretation are presented in a very accessible way by Huitema and McKean (2000).

A second possible extension is the inclusion of non-linear terms with respect to time, such as $(\text{Time}^2)_t$ to model quadratic trends. The interested reader is referred to Aiken, West, and Pitts (2003).

A third possible extension is the inclusion of additional terms to model the serial dependency. A very general methodology to accomplish this involves the use of so-called 'Autoregressive

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press

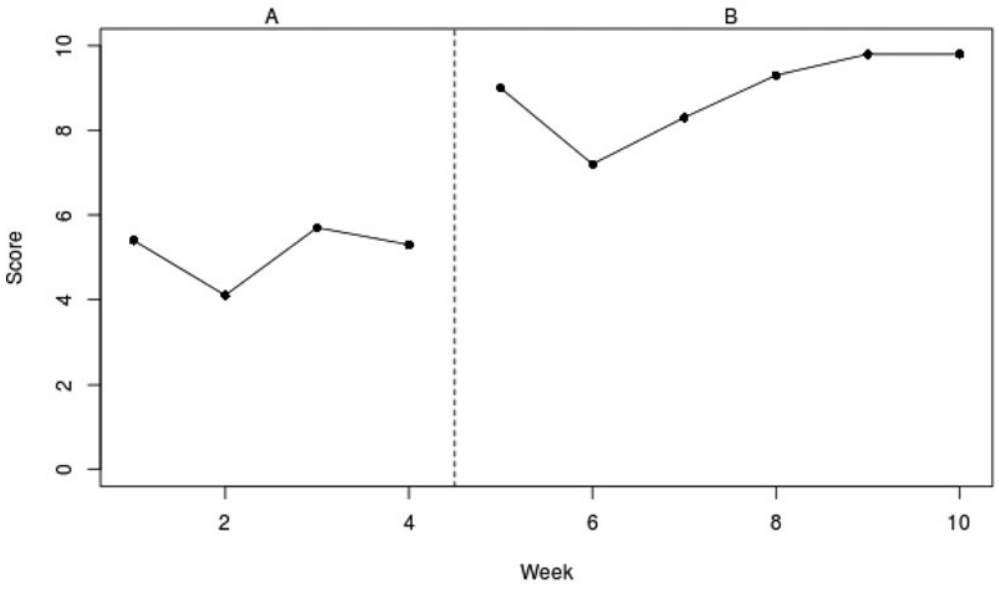


FIGURE 2

Weekly scores on the Discourse Coping Scale – Clinician Rating for Thomas during a 10-week SCE with a 4-week baseline (A phase) and a 6-week Communication-specific Coping Intervention (B phase).

Integrated Moving Average' models (see Box, Jenkins, Reinsel, & Ljung, 2016, for a modern presentation of this methodology). Those models make provision for higher-order autoregressive parameters and moving average parameters, based on repeated cycles of model identification, parameter estimation, and diagnostic checking. Furthermore, lagged predictor variables can be added to examine the relationship between the outcome variable and other manipulated or non-manipulated predictor variables with possible time delays. For example, if data about weekly mood were collected, a third predictor variable $(\text{Mood})_{t-1}$, with the corresponding parameter β_3, could be added to Equation (3) to examine the relationship between communication at week t and Mood the week before $(t-1)$.

For some applications, those extensions may be relevant, but essentially any modelling attempt will boil down to the three basics: modelling the treatment effect, modelling the trend and modelling the serial dependency, as demonstrated in the previous sections. Finally, notice that an important restriction in Equations (1), (2) and (3) is that a numerical outcome variable is considered. If the outcome variable is categorical, or even dichotomous, generalised models are needed (McCullagh & Nelder, 1989).

TABLE 5

Summary Statistics for the Thomas Data in Table 2

Phase	*n*	*M*	*SD*
A	4	5.12	0.70
B	6	8.90	1.00

Meta-Analysis: Modelling the Nested Structure for Replicated AB Designs in Models 4, 5 and 6

The same analyses can be performed on Thomas's data and the results can be found in Figure 2, Tables 5 and 6. For Thomas, Model 3 indicates a statistically significant treatment effect, a statistically significant linear trend, and a statistically significant negative autoregressive parameter.

In this example, there are two patients, but in general, there might be J patients and for each patient j ($j = 1, \ldots, J$) such a separate analysis can be performed. Moreover, if J is large enough and the SCEs are similar enough in design and research focus, then it might also be interesting to combine and compare the results of these multiple patients. Such a pooled analysis is called a 'meta-analysis' because it is an analysis at a higher level. Notice that the SCEs can be conducted simultaneously (as

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press

TABLE 6

Parameter Estimates (and Standard Errors) of the Models Specified in Equations 1–3 for the Thomas Data in Table 2

Parameter	Model 1	Model 2	Model 3
β_0	5.12** (0.45)	4.34** (0.57)	4.00** (0.28)
β_1	3.78** (0.58)	2.21 (0.97)	1.42** (0.50)
β_2		0.31 (0.17)	0.46** (0.09)
φ			−0.66** (0.24)

Note. *$p < .05$. **$p < .01$.

in multiple baseline designs across participants), at different points in time (as in replicated AB designs), at different research institutes, and the data may be extracted from different publications.

For the meta-analysis of these SCEs we will present the multilevel approach proposed by Van den Noortgate and Onghena (2003a, 2003c). For illustration purposes, we will use the data from Samantha and Thomas, but judicious application of this methodology should involve samples much larger than two (see Discussion section for the factors involved in determining the appropriate number of patients). In addition, we will assume that all raw data are available; an assumption that is not unrealistic for SCEs because in the SCE literature it is good practice to have the raw data published in a table or a graph.[5] Alternatively researchers can be asked to share the raw data upon request. If the raw data are not available and the meta-analysis has to be performed on summary statistics then the effect size model proposed by Van den Noortgate and Onghena (2003b) can be used.

For the meta-analysis, we have to adjust our notation. In order to make a distinction between the models for Samantha and the models for Thomas, an additional index has to be used to specify the particular patient that is involved. So for Samantha, Equation (1) becomes $\text{Score}_{t1} = \beta_{01} + \beta_{11}(\text{Treat})_{t1} + \varepsilon_{t1}$ and for Thomas $\text{Score}_{t2} = \beta_{02} + \beta_{12}(\text{Treat})_{t2} + \varepsilon_{t2}$. In general, for J patients, Equation (1) becomes

$$\text{Score}_{tj} = \beta_{0j} + \beta_{1j}(\text{Treat})_{tj} + \varepsilon_{tj} \qquad (4a)$$

Furthermore, if we want to include the data of all patients in one overall meta-analysis, we have to take the nested data structure into account: there are J patients and the repeated measures are nested within the patient. So for a meta-analysis, we are considering a sample from a population of patients and in each patient a random sample of repeated measurements. The additional index j is used to indicate that all values can differ between patients, so also the parameters can vary. The variation of the parameters can be modelled with a set of regression equations at the second level:

$$\begin{aligned} \beta_{0j} &= \gamma_0 + u_{0j} \\ \beta_{1j} &= \gamma_1 + u_{1j} \end{aligned} \qquad (4b)$$

Together Equations (4a) and (4b) constitute a simple meta-analytical model for SCEs. In this model γ_0 indicates the population average score for the baseline and γ_1 the difference between the population treatment average and the population baseline average. The residuals u_{0j} and u_{1j} indicate the deviations between the parameter values for each patient j and the value of the population parameter. By assuming that these residuals follow a multivariate normal distribution with zero means, variances σ_0^2 and σ_1^2, and covariance σ_{01}, in symbols this is:

$$\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{pmatrix} \qquad (4c)$$

we can perform statistical tests and construct confidence intervals.

The parameter estimates and standard errors for this Model 4 can be found in Table 7 [6]. In this meta-analytical model, the average treatment effect is considered statistically significant, $\gamma_1 = 3.24$, $t(16) = 4.09$, $p = .0009$, with [1.56; 4.92] being the approximate 95% confidence interval. The estimated standard deviations of the random effects indicate that there is little difference between the patients with respect to their average ($\sigma_0 = 0.04$),

[5] An SCE meta-analysis can therefore most frequently be characterised as a specific type of Individual Participant Data meta-analysis, which is considered the "gold standard" for systematic reviewing in medicine (Tierney et al., 2015).

[6] For this and subsequent analyses the covariance parameters were assumed to be zero to avoid convergence problems with these sparse data.

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press

TABLE 7

Model Estimates (and Standard Errors) for the Multilevel Meta-Analysis Containing both the Samantha and Thomas Data in Tables 1 and 2

Parameter	Model 4	Model 5	Model 6
γ_0	4.97** (0.39)	4.11** (0.45)	4.11** (0.44)
γ_1	3.24** (0.79)	1.46 (0.87)	1.44 (0.86)
γ_2		0.38 (0.13)	0.38 (0.13)
φ			−0.03 (0.61)
σ_0	0.04	0.01	0.01
σ_1	0.88	0.62	0.63
σ_2		<0.01	<0.01
σ_ε†	1.03	0.88	0.87

Note 1. $^*p < .05$. $^{**}p < .01$.
Note 2. Standard errors for the standard deviations of the random effects are not reported because the sampling distribution of these standard deviations is strongly asymmetrical (Pinheiro & Bates, 2000). Certainly for only two patients this standard error would be a very misleading measure of uncertainty. See Bates (2010) for an alternative approach.
†σ_ω for Model 6.

and that there are larger differences between the patients with respect to the treatment effect ($\sigma_1 = 0.88$) and between the repeated measurements after taking the treatment effect into account ($\sigma_\varepsilon = 1.03$).

Similarly, in a meta-analytical context, Equation (2) is extended to a model with three equations at the second level (one additional equation for the differences in time trend between the patients):

$$\begin{aligned}
&\text{Score}_{tj} = \beta_{0j} + \beta_{1j}(\text{Treat})_{tj} + \beta_{2j}(\text{Time})_{tj} \\
&\quad +\varepsilon_{tj}, \text{ with} \\
&\beta_{0j} = \gamma_0 + u_{0j}, \\
&\beta_{1j} = \gamma_1 + u_{1j}, \\
&\beta_{2j} = \gamma_2 + u_{2j}, \\
&\text{and } \begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{matrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} \\ \sigma_{10} & \sigma_1^2 & \sigma_{12} \\ \sigma_{20} & \sigma_{21} & \sigma_2^2 \end{matrix} \right),
\end{aligned} \tag{5}$$

and Equation (3) becomes

$$\begin{aligned}
&\text{Score}_{tj} = \beta_{0j} + \beta_{1j}(\text{Treat})_{tj} + \beta_{2j}(\text{Time})_{tj} \\
&\quad +\varepsilon_{tj}, \text{ with} \\
&\varepsilon_t = \varphi\varepsilon_{t-1} + \omega_t \quad \text{and} \quad \omega_t \sim N\left(0, \sigma_\omega^2\right), \\
&\beta_{0j} = \gamma_0 + u_{0j}, \\
&\beta_{1j} = \gamma_1 + u_{1j}, \\
&\beta_{2j} = \gamma_2 + u_{2j}, \\
&\text{and } \begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{matrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} \\ \sigma_{10} & \sigma_1^2 & \sigma_{12} \\ \sigma_{20} & \sigma_{21} & \sigma_2^2 \end{matrix} \right).
\end{aligned} \tag{6}$$

The results for Model 5 show that the treatment effect is again not statistically significant anymore, $\gamma_1 = 1.46$, $t(15) = 1.68$, $p = .1141$, once the trend is taken into account (Table 7). The approximate 95% confidence interval for the treatment effect is [−0.40; 3.32], which obviously contains zero. The time trend is statistically significant, $\gamma_2 = 0.38$, $t(15) = 2.82$, $p = .0129$, with an approximate 95% confidence interval of [0.09; 0.66], which means that there is a general upward trend. The standard deviations of the random effects show the same pattern as in Model 5, with the addition that there is little difference between the patients with respect to this general time trend ($\sigma_2 < 0.01$). As can be seen in Table 7, the results for Model 6 are essentially identical to the results for Model 5, due to the small (and non-significant) value for the autoregressive parameter ($\varphi = -0.03$).

According to Douglas et al. (2015, p. 190), these SCEs provide 'sound phase-1 evidence for the effectiveness of CommCope-I' (p. 190), but our analysis demonstrates that with this simple design, with this statistical model, only two SCEs, and these data, it is impossible to convincingly distinguish between the treatment effect and the general time trend. Instead of calling it 'sound phase-1 evidence' it would be more cautious to label these results as 'promising'. Of course, we have to acknowledge that our analysis was also tentative, based on questionable assumptions, and for demonstration purposes only. When readers would look at the complete Douglas et al. (2015) study, including the follow-up data and more qualitative information, and they would perform an analysis based on alternative statistical models (e.g., including a $(\text{Treat})_t \times (\text{Time})_t$ interaction), then it cannot be ruled out that other conclusions are drawn.

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press

Meta-Analysis: More Advanced Issues and Tools

The multilevel meta-analytic model can be extended in several ways. Without going into the technical details, we mention nine additions and variations, and provide the references for the interested reader:

1. Usually a multilevel meta-analysis is performed not only to combine the patient results, as in the example, but also to compare the patients. Formally, this comparison is carried out by adding a moderator variable at the higher-level regression equation. For example, a comparison between men and women with respect to the treatment effect in Equation (4b) is accomplished by dummy coding men and women in the data set and extending the Equation as $\beta_{1j} = \gamma_{10} + \gamma_{11}(\text{Sex})_j + u_{1j}$. Notice that there is no t index for this additional variable because sex only varies between patients (i.e., in the second-level equation); this variable is a characteristic of the patients, not a characteristic of the measurement occasions. Notice also that it is possible to include numerical moderator variables. For example, the patient's age can be added and also interactions between sex and age examined: $\beta_{1j} = \gamma_{10} + \gamma_{11}(\text{Sex})_j + \gamma_{12}(\text{Age})_j + \gamma_{13}[(\text{Sex})_j \times (\text{Age})_j] + u_{1j}$. If there is substantial variation in the lower-level parameter between patients (e.g., the β_1's), then the addition of moderators at the higher level will allow a better understanding of how this variation comes about (Baek et al., 2014; Van den Noortgate & Onghena, 2003a; Van den Noortgate & Onghena, 2008).
2. Multilevel meta-analysis of SCEs need not be restricted to two levels (repeated measurements within patients) as in the Samantha-Thomas example. The model can also be applied to three or more levels, for example, when a literature review involves multiple SCEs nested within each publication (three levels: repeated measurements within patients within publications) (Moeyaert, Ferron, Beretvas, & Van den Noortgate, 2014; Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2013a, 2013b).
3. Multilevel meta-analysis of SCEs need also not be restricted to one outcome variable. In many applications, there are multiple outcome variables or multiple effect size measures and the model has to be adapted accordingly. This can be accomplished in a multilevel meta-analysis by including the outcome variable as a separate level in the model (Moeyaert et al., 2016; Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013, 2015).
4. Multilevel meta-analysis of SCEs is also applicable to other designs than replicated two-phase AB designs (Moeyaert et al., 2015). Moeyaert et al. (2014) showed how design matrices can be specified for reversal designs, alternating treatments designs, and multiple-baseline designs. Brosnan et al. (2017), for example, used such an extended multilevel model for multiple-baseline data in a study evaluating precision teaching to improve the fluency in foundational reading skills for at risk readers.
5. In the example, we performed the analyses at the individual patient level first, and afterwards combined the two analyses in a two-level meta-analysis. The other way around, the two-level meta-analysis also enables a more accurate estimation of the patient-level treatments effects afterwards. The advantage of these 'empirical Bayes estimates', using a weighted average of the estimate based on the individual patient data and an estimate based on the average of all patients' data, is that for short time series data at the patient level it is still possible to obtain reliable estimates by 'borrowing strength' from the average estimate (Ferron, Farmer, & Owens, 2010; Van den Noortgate & Onghena, 2003a).
6. As mentioned before, multilevel meta-analysis of SCEs can also handle data sets with only descriptive statistics or effect sizes available. Without having access to the raw data, a multilevel meta-analysis can still be performed if sufficient information is available to estimate the sampling variance of the statistics prior to performing the meta-analysis (Ugille, Moeyaert, Beretvas, Ferron, & Van den Noortgate, 2012; Van den Noortgate & Onghena, 2003b).
7. Because multilevel meta-analysis of SCEs usually involves consecutive patients or consecutive publications, techniques from sequential analysis can be applied. Kuppens and Onghena (2010) and Kuppens, Heyvaert, Van den Noortgate, and Onghena (2011) proposed such a sequential meta-analysis of SCEs, which focuses on the question whether enough cumulative knowledge has been collected to draw convincing statistical conclusions and when data collection can be stopped. Determining the sufficiency of cumulative knowledge avoids the inefficient use of limited resources and prevents the dissemination of treatments with dubious benefits (Kuppens & Onghena, 2012).

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press

8. Multilevel meta-analysis of SCEs can be completed by performing sensitivity analyses and publication bias analyses in a similar way as with standard meta-analysis (Cooper, Hedges, & Valentine, 2009). Sensitivity analyses are used to examine whether certain researcher decisions (e.g., ways of handling missing data, adding a moderator variable or not) distort the conclusions and the inferences drawn from the analysis. Publication bias analyses are used to examine whether there are indications that the results in the published literature differ systematically from the results that remain unpublished (e.g., by using a funnel plot-, calculating a fail-safe N, or performing Duval and Tweedie's trim and fill analysis). Examples of sensitivity analyses and publication bias analyses in multilevel meta-analyses of SCEs can be found in the applied studies by Heyvaert, Maes, and Onghena (2010), Heyvaert, Maes, Van den Noortgate, Kuppens, and Onghena (2012), and Vanderkerken, Heyvaert, Maes, and Onghena (2013).
9. Finally, besides deriving the patient-level treatments effects using 'empirical Bayes estimates', it is also possible to perform a fully Bayesian analysis of multilevel data if prior distributions for the parameters can be specified. The fully Bayesian approach is promising for small data sets, offers an elegant solution for imputing missing values, and also lends itself very nicely to a sequential or cumulative meta-analysis (de Vries & Morey, 2013; Kravitz & Duan, 2014; Moeyaert, Rindskopf, Onghena, & Van den Noortgate, 2017).

Meta-Analysis: Alternative Methods

Applications of multilevel meta-analysis of SCEs are still scarce. Jamshidi et al. (2017) reviewed 178 SCE meta-analyses and found only 22 studies (12%) that applied a multilevel analysis. This may be partly due to the fact that single-case researchers are reluctant to use advanced statistical techniques on their data. Most single-case researchers prefer visual techniques and calculate simple non-overlap measures to quantify the treatment effects in a separate SCE (Manolov & Moeyaert, 2017a, 2017b; Manolov this issue). For the analysis of multiple SCEs this preference translates into the use of simple methods to aggregate studies. Most SCE meta-analyses report a simple average or a range of effect sizes (Jamshidi et al., 2017).

In the original analysis of the Samantha-Thomas example, Douglas et al. (2015) also based their 'sound phase-1 evidence' on the calculation of a non-overlap measure for each patient (with a maximum score of 100% non-overlap for each patient). They used a measure corrected for baseline trend (Manolov & Solanas, 2009), but without impact for their data (because there was no noticeable baseline trend). Although the complementary use of visual techniques and the calculation of non-overlap remain strongly recommended (Lane & Gast, 2014; Lane & Ledford this issue; Manolov this issue), the sole use of these techniques may lead to unreliable or unreplicable conclusions (Chen, Peng, & Chen, 2015; Dart & Radley, 2017; Perdices & Tate, 2009). Particularly in short series, it appears to be very difficult to distinguish between treatment effects on the one hand and random fluctuations or time trends on the other hand, without a proper statistical analysis (Harrington & Velicer, 2015; Jones et al., 1978; Ottenbacher, 1993).

One concern single-case researchers might have regarding multilevel meta-analysis of SCEs, pertains to the parametric assumptions that are involved, and to the inability to check these assumptions without invoking new assumptions. Avoiding these parametric assumptions altogether, Edgington (1967, 1975, 1980) advocated the use of randomisation tests for the statistical analysis of SCEs. These randomisation tests can be used for a collection of randomised SCEs or, if the raw data are not available anymore, a *p*-value combining procedure (Edgington, 1972; Fisher, 1925) can be used on the randomisation test *p*-values resulting from the separate SCEs.

Instead of assuming random sampling, as in multilevel meta-analysis of SCEs, randomisation tests of SCEs assume random assignment of treatments to measurement occasions (Ferron & Levin, 2014; Heyvaert & Onghena, 2014). In single-case AB phase designs the random assignment refers to the random determination of the first measurement occasion in the B phase, given a minimum number of measurement occasions in each phase. For example, suppose that for the Samantha data the minimum number of weeks in each phase is three. Because there are nine weeks, the first week in the B phase could be the fourth, the fifth, the sixth, or the seventh week. One of these four possible designs is randomly picked for data collection. Suppose that the design with the fourth week is selected for the start of the B phase and that the data in Figure 1 are obtained. If there is no difference between A and B, what is the probability that we have selected the design that gives such a large difference between A and B, or an even larger difference than the difference that we observed? The answer to this question equals the randomisation test *p*-value for the Samantha data.

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press

In this simple example, the answer can easily be calculated by hand. First we have to specify what we mean by a 'difference between A and B'. Suppose that we were interested in the absolute difference between the arithmetic averages of the A and the B phase, $|\bar{A} - \bar{B}|$. This is our test statistic. For the observed data $|\bar{A} - \bar{B}|$ is equal to 2.75. This is the observed value of the test statistic.

Secondly we have to count the number of possible designs. In this example there are only four possibilities (four possible intervention weeks). Under the null hypothesis that there is no difference between A and B, the same data were observed regardless of the design that would have been applied. In technical terms: the data are fixed and the design is random. If we now calculate the test statistic for each of the possible designs, given the observed data, we obtain a reference distribution for the test statistic. Table 8 shows the test statistic values for each of the four designs and the designs applied to the observed data.

Finally, based on this reference distribution, we can calculate the *p*-value of the randomisation test. This *p*-value is the proportion of test statistic values in the reference distribution that is equal to or larger than the observed value of the test statistic. For the Samantha data the *p*-value is .25.

Suppose now that we followed the same procedure for the Thomas data, and that we want to perform a meta-analysis by combining both *p*-values. Assuming random determination of the start of the B-phase and a minimum of three weeks in each phase, the randomisation test *p*-value for the Thomas data turns out to be .20. Edgington's (1972) additive method for combining both *p*-values results in a combined *p*-value of $(.20 + .25)^2/2 = .1013$. This result is comparable to the treatment effect *p*-value of .1141 in Model 5 of the multilevel meta-analysis. Notice, however, that we did not make any assumption regarding random sampling or the shape of the population distribution to justify the calculation of this combined *p*-value. The simplicity, versatility and freedom from distributional assumptions makes *p*-value combining an attractive alternative (or addition) for SCE meta-analysis. More information about methods for *p*-value combining can be found in Rosenthal (1978), Strube and Miller (1986), Pesarin and Salmaso (2010) and Solmi and Onghena (2014). Software to perform randomisation tests for SCEs and to combine *p*-values is provided by Bulté and Onghena (2008, 2009, 2013).

Methods for multilevel meta-analysis on the one hand and methods for randomisation tests and *p*-value combining on the other hand may seem antithetical, with the first being firmly rooted in model-based inference and the latter being the prototype of design-based inference (Koch & Gillings, 1984; Sterba, 2009). However, both methods can also be integrated by realising that the randomisation test is not a 'test' in the strict sense (because the design and the test statistic are unspecified) but rather a very generic procedure to construct a reference distribution. Consequently, this generic procedure can also be put at work to calculate the *p*-value of the treatment effect parameter in a multilevel meta-analytic model for SCEs. This so-called 'randomisation-test wrapper' can turn any *p*-value based on parametric assumptions into a design-based *p*-value based only on random assignment (Cassell, 2002; Heyvaert et al., 2017).

Application of the randomisation test wrapper to the Samantha-Thomas data implies that we have to perform the multilevel meta-analysis 20 times, assuming that both SCEs have a randomly determined start of the B phase and that each phase lasts at least for three weeks. Under those assumptions there are five possible Thomas designs for each of four possible Samantha designs, and a multilevel meta-analysis can be carried out for each constellation of possible start weeks of the B phase. The reference distribution for this randomisation test wrapper is shown in Table 9. There are two test statistics with an absolute value larger than or equal to the observed statistic, and therefore the resulting *p*-value is $2/20 = .10$, which again is comparable to the treatment effect *p*-value of .1141 in the parametric analysis of Model 5 and the combined *p*-value of .1013.

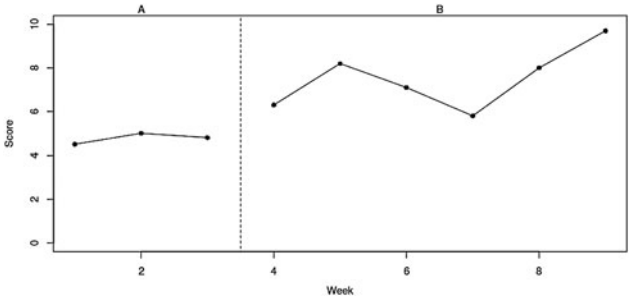
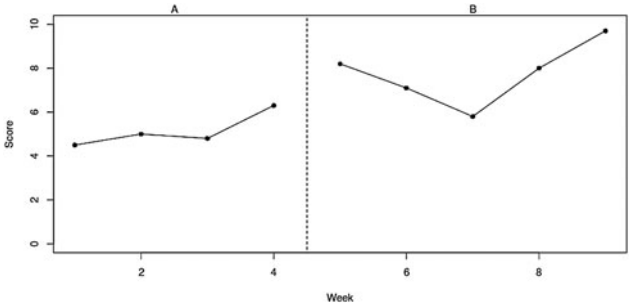
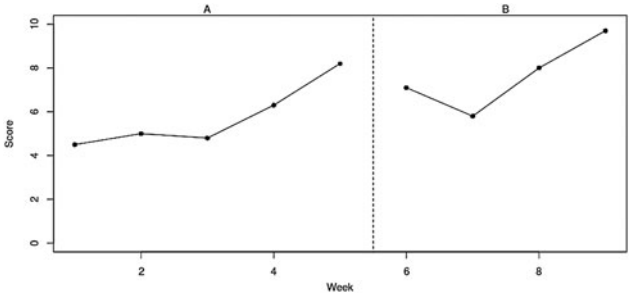
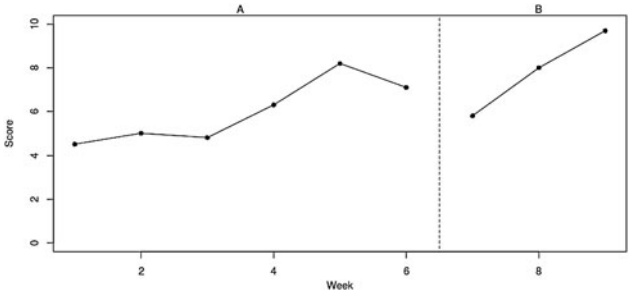
Although randomisation tests and the randomisation test wrapper are useful tools for design-based analysis and meta-analysis of SCEs, notice that they are only relevant for statistical inference regarding treatment effects. Because both procedures are based on random assignment, it is only for the manipulated variable that a reference distribution is generated. General time trends or variation between patients remain invariant under random assignment in replicated SCEs and therefore design-based inference is not possible for the corresponding parameters.

Finally, it should be mentioned that the foregoing presentation of randomisation procedures was framed in terms of null hypothesis testing, but that also confidence intervals can be constructed based on the random assignment rationale. Michiels, Heyvaert, Meulders, and Onghena (2017) showed how to derive confidence intervals for single-case effect size measures by inverting the randomisation test, assuming random assignment and a constant additive treatment effect. Michiels and Onghena (2017) extended this derivation to a collection of SCEs.

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press

TABLE 8

Reference Distribution for the Randomisation Test with $|\bar{A} - \bar{B}|$ as the Test Statistic Applied to the Data of Samantha in Table 1

Design	Design Applied to the Observed Data	Test Statistic $\lvert\bar{A} - \bar{B}\rvert$
AAABBBBBB*		2.75*
AAAABBBBB		2.61
AAAAABBBB		1.89
AAAAAABBB		1.85

*Actually used design and value of the test statistic for the observed data.

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press

TABLE 9

Reference Distribution for the Randomisation Test Wrapper with the *t*-value or the *p*-value of the treatment effect parameter γ_1 of Meta-Analytical Model 5 as the Test Statistic, Applied to the Combined Data of Samantha and Thomas in Tables 1 and 2

Design†	*t*-value	*p*-value
(4;5)*	1.68*	.1141*
(5;5)	1.61	.1286
(4;4)	0.70	.4935
(5;4)	0.60	.5571
(7;5)	0.19	.8499
(4;7)	0.17	.8650
(6;5)	0.16	.8714
(4;8)	0.11	.9173
(5;7)	0.02	.9869
(5;8)	− 0.08	.9380
(4;6)	− 0.23	.8216
(5;6)	− 0.36	.7263
(7;4)	− 0.51	.6168
(6;4)	− 0.68	.5099
(7;7)	− 0.97	.3470
(7;8)	− 1.03	.3213
(6;7)	− 1.26	.2284
(6;8)	− 1.30	.2138
(7;6)	− 1.33	.2035
(6;6)	− 1.73	.1050

†Designs are named according to the start week of the B phase for (Samantha; Thomas).
*Actually used design, test statistic, and *p*-value for the observed data.

Discussion

In this article we presented a unilevel approach for the analysis of SCE data and a multilevel approach for the meta-analysis of multiple SCEs, and we referred to a number of alternative and complementary approaches. Furthermore, we illustrated the unilevel approach, the multilevel approach, probability combining, and the randomisation test wrapper with data from a study by Douglas et al. (2015) on the effectiveness of a new treatment that is intended to strengthen communication-specific coping after traumatic brain injury. Although there are as many approaches to the meta-analysis of SCEs as there are approaches to meta-analysis, we believe that the multilevel approach provides a versatile, flexible, and comprehensive framework for the meta-analysis of SCEs, and that it enables the single-case researcher to connect to general statistical theory (Araujo, Julious, & Senn, 2016; Beretvas & Chung, 2008; Kratochwill & Levin, 2014; Shadish, 2014; Shadish, Kyse, & Rindskopf, 2013).

With our re-analysis of the data from the Douglas et al. (2015) study, we did not have the intention to be overly critical with respect to the main conclusions of their study. On the contrary, we selected this study because it tackled an important research question, and because its report was very clear, instructive, and transparent regarding data collection, data handling and data analysis. We acknowledge that, in drawing their conclusions, Douglas et al. (2015) linked their precious empirical results to substantive theory, qualitative information about the patients, their treatment, and their follow-up. At the same time, the Douglas et al. (2015) study enabled us to point out the hazards of basing conclusions on short time series and a limited number of patients.

In confronting statistical models with real data, statisticians often have to make concessions and seek ad hoc solutions. Data are frequently more 'messy' than the statistical models allows for. On the one hand, a general advice to single-case researchers would be to more carefully design SCEs and to more reliably collect the data. On the other hand, real life usually *is* messy, certainly when you are working in a hospital. When working with patients with brain injury, unplanned events happen. For example, the Samantha data in the Douglas et al. (2015) study unexpectedly contained one missing week (Table 1) and a thorough statistical analysis should take missing data points into account. We refrained from going into the problems associated with this missing data point because it was not the main purpose of the illustration. However, it is obvious that the increased uncertainty due to the missing data point makes the treatment effect in the Samantha data even less convincing than it already is. Statistical options to formally handle missing data include full maximum likelihood estimation (Hartley & Hocking, 1971), multiple imputation (Little & Rubin, 1987), or random assignment if it can be assumed that the missingness is unrelated to the treatment (Edgington & Onghena, 2007).

As mentioned before, it is recommended to perform a sensitivity analysis by comparing analyses without the missing data and analyses with techniques to deal with the missing data. If SCEs contain short time series with a lot of missing data, or if SCEs contain messy data in general, the researcher should be reminded of Fisher's (1938) warning: 'To consult the statistician after an experiment is finished is often merely to ask him to conduct a *post mortem* examination. He can perhaps say what the experiment died of' (Fisher, 1938, p. 17).

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press

Another simplification that we introduced in our analysis of the Douglas et al. (2015) data is that we ignored the results from the follow-up phase. It would have been easy to accommodate the multilevel model for a second A phase (see e.g., Moeyaert et al., 2015) and also randomisation tests for ABA designs are straightforward (see e.g., Onghena, 1992), but as we indicated before, the description of the Douglas et al. (2015) design as an ABA design may not have been the best choice. The design should have been labelled 'AB with follow-up'. Because the time intervals in the follow-up phase are completely different as compared to the A and B phase, it would have been difficult to include this phase in the multilevel meta-analysis without invoking additional implausible assumptions. Note also that the reference to ABA designs or to reversal or withdrawal designs in some publications might be a side effect of not allowing properly conducted AB designs with extensive follow-up as valid designs (see comment 6 of the SCE definition).

In the example, the multilevel meta-analysis, the *p*-value combining, and the randomisation test wrapper resulted in comparable *p*-values for the treatment effect. This in no way suggests that this will always happen. On the basis of this result, it might also be tempting to embark on simulation studies to compare the power and robustness of the different methods, but it is more realistic to consider the different methods being different perspectives for looking at the same data. A combined use of these methods and convergent results strengthen the inference.[7] For example, a researcher may start the analysis with a bare-bones approach of randomisation tests and *p*-value combining and may continue with more sophisticated statistical modelling, such as a multilevel meta-analysis, if more time and resources become available. If there are doubts about the parametric assumptions of the multilevel meta-analysis or a design-based inference is needed then the randomisation test wrapper may come into play. Notice also the 'as if' nature of the statistical inference in the example because neither random sampling nor random assignment were present. The multilevel meta-analysis is carried out 'as if' the two patients were a random sample of the target population and the repeated measurements were a random sample from the patients. The randomisation test and the randomisation tests wrapper are carried out 'as if' the start week for the B phase was randomly determined.

We want to end this discussion section with some general comments on the literature regarding the statistical analysis and meta-analysis of SCEs, on statistical modelling, on meta-analysis, and give some suggestions for future research.

Comment on the Literature Regarding the Analysis and Meta-Analysis of SCEs

When reviewing the literature regarding the analysis and meta-analysis of SCEs, one sometimes gets the impression that the wheel is reinvented. Sometimes it seems that a completely new statistical language is needed with specific overlap measures and inferential techniques. In other publications, statistical techniques seem to be borrowed, and at the same time divorced, from the mainstream statistical literature. These techniques might get nourished in particular niche journals with minor reference to the broader statistical literature, but this is not beneficial for the methodological development of the field.

By contrast, in this article we showed that standard multilevel analysis, common meta-analytical tools, and old school randomisation tests can be used for the analysis and meta-analysis of SCEs. The underlying tenet of our demonstration is an advocacy to connect to the general statistical literature and to consider an SCE data set as any other data set. The only special consideration refers to the specific features that all SCE data have in common: repeated observations or measurements and a manipulated variable with at least two levels. These features imply that particular attention has to be paid to time trends and problems of serial dependence, and that randomisation schedules should be compatible with the phasic or alternation structure of the design.

Connecting the analysis and meta-analysis of SCEs to the general statistical literature has the advantage that more than a century of statistical experience and research is made to bear on these data, and that the results of analyses and meta-analyses of SCEs might become more acceptable/

[7] Although multilevel meta-analysis, *p*-value combining, and the randomisation test wrapper can be used to test "treatments effects", technically speaking they are testing different null hypotheses. The null hypothesis for the multilevel test is that the average treatment effect is zero. The null hypothesis for *p*-value combining and the randomisation test wrapper is that the treatment effect is zero for every patient. This also means that the resulting *p*-values do not need to converge. When they do converge, it is additional evidence for the robustness of the treatment effect. When they diverge, the source of the divergence may be examined (e.g., violation of the assumptions or heterogeneity between patients). For more background and implications, see Heyvaert et al. (2017).

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press

digestible/convincing to the general scientific community. Furthermore, this connection avoids having to repeat all discussions and controversies regarding model-based versus design-based inference, population versus causal inference, frequentist versus Bayesian inference, parametric and nonparametric inference, the complementarity of visual and statistical techniques, the justified use of exploratory and confirmatory techniques, the combination of quantitative and qualitative data, ... within an isolated SCE community.

Comments on Statistical Modelling

We already indicated that statistical modelling involves making concessions. These concessions are a consequence of any scientific model being an attempt at providing an insightful reduction of a complex reality, guided by the epistemological principle of parsimony. Furthermore, if this principle is left aside, then statistical modelling might become problematic for trustworthy inferences. For example, the multilevel model is so flexible and has so many options that very complex models can be built, to the extent that the model is just a description of the data. This is exactly what happens if the number of parameters to be estimated equals the number of independent data points.

A corollary of this flexibility is that many decisions have to be made in the modelling enterprise and that some decisions might be guided by the researcher's biases pro or contra certain hypotheses. To counteract these biases, and to control for type I and type II errors in a formal statistical framework, it is necessary to specify the model and its test statistics before the data are inspected. Consequently, it might also be recommended for the SCE community to install formal preregistration of SCEs and their meta-analyses (see e.g., the PROSPERO registry; Booth et al., 2012) to steer clear of too many 'researcher degrees of freedom' and of too much opportunistic, misguided and self-deceiving post factum model building (Simmons, Nelson, & Simonsohn, 2011; Wicherts et al., 2016).

Comments on Meta-Analysis

According to Eysenck (1978) a meta-analysis is 'an exercise in mega-silliness' and he referred to a meta-analysis that was, in his opinion, a compilation of heterogeneous ('mixing apples and oranges') and poorly designed studies ('garbage in, garbage out'). Furthermore, meta-analysis is also sometimes contested as a valid scientific method because the results in the sampled or published studies are most likely unrepresentative of the phenomenon studied (the so-called 'file drawer problem'), meta-analysis conclusions can disagree with large-scale randomised controlled trials, meta-analyses are not reproducible, meta-analyses are performed poorly, and meta-analyses use inferior statistical tools (Borenstein, Hedges, Higgins, & Rothstein, 2009; Lakens, Hilgard, & Staaks, 2016; Scargle, 2000; Terrin, Schmid, Lau, & Olkin, 2003).

This controversy and skepticism regarding meta-analysis carries over to the meta-analysis of SCEs. While some of the criticisms can easily be debunked (see Borenstein et al., 2009, and Cooper et al., 2009, for excellent responses), single-case researchers should be encouraged to raise the bar with respect to the quality of their meta-analyses. The criticism of 'garbage in, garbage out' can be turned into a constructive approach to meta-analysis as 'waste management' by using methodological and reporting quality of the SCEs as inclusion criteria or moderator variables. Or more fundamentally, single-case researchers can engage in 'sustainable research' by producing less waste and improve the overall quality of the design and analysis of the primary studies. The instrument developed by Tate et al. (2013) for the assessment of the methodological quality of SCEs and the reporting guidelines developed by Tate et al. (2016a, 2016b) are very useful in this respect. By paying more attention to the methodological and reporting quality of the primary studies, the field of SCEs as a whole will benefit (Kratochwill et al., 2010, 2013; Lobo, Moeyaert, Cunha, & Babik, 2017). In addition, the methodological and reporting quality of the SCE meta-analyses, as such, can be improved by adhering to the general standards and guidelines for meta-analyses and systematic reviews (Moher, Liberati, Tetzlaff, Altman, & the PRISMA group, 2009; Shea et al., 2009).

Besides the general controversy and skepticism regarding meta-analysis, two concerns need special care when considering an SCE meta-analysis. The first concern is the relatively high risk for selectivity of cases submitted and published in the literature. If cases are published with a strong preference for spectacular, extreme, or successful cases, then any meta-analysis of these cases will be severely biased in the direction of success. A survey by Shadish, Zelinsky, Vevea, and Kratochwill (2016) found indications that the publication bias is larger for SCEs than for other research designs indeed. This raises important concerns with respect to the validity of SCE meta-analyses because no sophisticated statistical analysis can remedy the selectivity of cases.

Another concern is the kind of inference that is aimed at with an SCE meta-analysis. As

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press

mentioned in our reflection on the definition of an SCE, some SCEs are performed just for the benefit of a single patient (comment 5). Such SCEs are purely idiographic and do not have the intention to be generalisable. They aim at causal inference for one particular patient. When performing an SCE meta-analysis this should be taken into account. It should be clear to what extent the meta-analysis considers the patients as a random sample from a target population and to what extent the meta-analysis is an aggregation for causal inference, with generalisation based on non-statistical arguments.

Future Research

The analysis and meta-analysis of SCEs represents a thriving part of current methodological and statistical developments and applications. In the context of this article, we mention three topics that need more research than currently available.

The first topic is the number of patients and the number of studies needed to perform a reliable and efficient multilevel meta-analysis. Cools, Van den Noortgate, and Onghena (2008, 2009) developed general software to determine the efficiency of multilevel designs, but a validation in the context of SCE meta-analysis is still lacking. It is to be expected that there are numerous factors involved in determining the appropriate number of patients and studies (e.g., the number of repeated measurements and the size of the effects that are considered clinically relevant) but more guiding principles would be useful. Van den Noortgate and Onghena (2007) suggested that multilevel meta-analysis of SCEs should include at least about 20 entities at the highest level (patients or studies). Also Kuppens et al. (2011) recommended that the initial interim analysis of a sequential single-case meta-analysis should comprise at least about 20 single-case publications including single or multiple cases. By contrast, Heyvaert et al. (2017) performed an extensive simulation study and found acceptable statistical power values for replicated AB designs with seven patients, 20 repeated measurements each, for both multilevel meta-analysis and probability combining of randomisation test *p*-values.

A second topic is the combination of SCEs and group-comparison studies in one overall meta-analysis. Although, it would be important for evidence-based practice that all kinds of scientific evidence were given weight in comprehensive systematic reviews, it is not evident how one should proceed. One option is to aim at a quantitative integration of both kinds of evidence. For example, Pustejovsky, Hedges, and Shadish (2014) proposed a general framework for defining effect sizes in replicated SCEs that are directly comparable to the standardised mean difference in between-subjects randomised experiments (see Punja et al., 2016, for an application). Shadish, Rindskopf, and Boyajian (2016) found similar results between an SCE meta-analysis and the results of a randomised controlled trial but also mentioned that more work is needed to understand the conditions under which this holds. A second option is to use estimates of between- and within subject variances to convert SCE effect sizes in order to make them comparable to group-comparison study effect sizes, and combining them in a single meta-analysis, as described by Van den Noortgate and Onghena (2008). A third option is to separately perform the SCE meta-analysis and the group-comparison meta-analysis and to combine the results in a narrative way (looking for convergence, elaboration, or confrontation). The latter option has the advantage that also qualitative case studies can be involved (Heyvaert, Hannes, & Onghena, 2017; Heyvaert, Maes, & Onghena, 2013). Future research should examine the potential and pitfalls of these options, both in theory and application.

The third topic is the development of guidelines and standards for the statistical analysis and meta-analysis of SCEs. Many authors have referred to the lack of consensus regarding the statistical analysis and meta-analysis of SCEs (see e.g., Cohen, Feinstein, Masuda, & Vowles, 2014; Smith, 2012), which may even reinforce some researchers in just performing a visual analysis and calculating a simple Percentage of Non-overlapping Data (PND), without any further statistical analysis at all (Horner et al., 2005; Kratochwill et al., 2010, 2013; Schlosser, Lee, & Wendt, 2008; Scruggs & Mastropieri, 2013). Hence some guidelines and standards are more than welcome (see Manolov & Moeyaert, 2017b, for some provisional recommendations).

In developing these guidelines and standards we should not chase the chimera that there is only one optimal way of performing the statistical analysis and meta-analysis of SCEs. Optimality always refers to some criterion, and this criterion in its turn can be the subject of debate. We can emphasise robustness, power, small sample properties, asymptotic properties, unbiasedness, accuracy or some other criterion and depending on the criterion have another preference. As wittingly remarked by John Tukey: 'the collective noun for a group of statisticians is a quarrel' (Vallverdú, 2016), illustrating that this 'lack of consensus' is evident for any non-ritualistic statistical analysis (Tukey, 1969). As Stephen Senn (2017) put it: 'Consult two medics and you'll get two opinions but

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press

consult two statisticians and you could easily get three' (Senn, 2017).

Given the current state of the art, there is only one principle that we want to propose for the development of these guidelines and standards. This principle is the connection of any technique for the analysis and meta-analysis of SCEs to the general statistical literature, as we mentioned before. SCE data sets can be analysed as any other data set, with particular attention to time trends, possible serial dependence, and randomisation schedules that are consistent with specific single-case designs. This guiding principle implies a critical attitude towards an overreliance on visual techniques and PND to analyse and meta-analyse SCEs. For example, no one would seriously consider using PND to compare two groups of a randomised controlled trial. For some strange reason, though, we tend to accept this measure for the analysis and meta-analysis of SCEs, although it does not take into account any of the specific features of SCEs.

Conclusion

Patients enter a clinic or a controlled trial one by one and the analysis and meta-analysis of the patients' data can mimic this data collection procedure. This may be accomplished by testing treatment effects at the individual level first, and subsequently perform a meta-analysis on the collection of individual patient data. For testing treatment effects at the individual level, researchers are recommended to set up SCEs, using the most rigorous experimental procedures possible, given the research setting and resources. In this paper, we presented this bottom-up individualised experimental approach as a viable alternative to the more common group-comparison approach.

In order to select techniques for the analysis and meta-analysis of the SCEs' data, it is important to have an unambiguous SCE definition at our disposal and to identify the essential features of SCE data. We defined SCEs as 'true' experimental studies (i.e., we do not consider them as 'quasi-experiments'), in which one entity is observed repeatedly under different levels of at least one manipulated variable, with the single entity and the experimental nature as the essential features. For the analysis and meta-analysis of SCEs particular attention has to paid to trends, serial dependence, and the specific single-case design that has been used. The multilevel model, as elaborated in this paper, is particularly suited to deal with the analysis and meta-analysis of SCEs, but other techniques are possible. One other technique, the calculation of PNDs and averaging PNDs, is not recommended because this technique is not geared towards any of the specific features of SCE data.

Sometimes SCE data are messy and sparse and in those cases the appropriate use of multilevel models may be a difficult balancing act. Multilevel models can be very flexible, but they also entail statistical assumptions. So if the design and data are simple or contain all sorts of artefacts, the use of complex multilevel modelling might resemble the use of cannon to kill a fly (and miss it). Other techniques, such as probability combining and the randomisation test wrapper, might be more resilient to the hazards of data collection in clinical practice with brain injured patients, but also these techniques have their shortcomings. The randomisation test wrapper has a sound theoretical foundation, but hitherto applications and investigation of its operating characteristics are lacking. Probability combining is backed by a long history of statistical research, but is only subsidiary in the context of a meta-analysis because it does not provide an effect size estimate and may mask effects that go in opposite directions if the use of one-sided or two-sided *p*-values is not appropriately specified (Borenstein et al., 2009; Edgington & Onghena, 2007).

With respect to the number of patients for a proper use of multilevel meta-analysis of SCEs, single-case researchers should not settle for a small number of patients, each providing a small number of repeated measurements. The use of multilevel models requires the number of patients and the number of repeated measurements based on a corresponding statistical power analysis. In any case, single-case researchers are encouraged to justify their sample of patients and to be wary of indications of publication and reporting biases. If no other justification is available, it is safest to follow the tentative recommendations proposed by Kratochwill et al. (2010, 2013) and to refrain from meta-analysing published SCEs with the intention to derive a general causal effect unless the following criteria are met:

1. A minimum of five SCE research papers, using valid experimental designs,
2. A minimum of three different research teams at three different geographical locations who conducted the SCEs, and
3. A minimum of 20 SCEs across the papers.

Finally, one meta-conclusion after reviewing the literature on SCE meta-analysis: The single-case research community would benefit from more unified terminology and methodology. For example, the emergence of terms such as '*N*-of-1 trial' and 'Individual Participant Data meta-analysis'

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press

has created dissociated literatures with little or no cross-references to the original publications on SCEs and SCE meta-analysis. For the future of SCE analysis and meta-analysis, we hope that this can be remedied and that more cross-fertilisation will occur. It is not because another word is used, that another reality is represented.

Acknowledgements

We want to thank the guest editor, Robyn Tate, and two anonymous reviewers for their helpful suggestions regarding a previous version of this manuscript.

Financial Support

This work was supported by the Research Foundation – Flanders (FWO), Belgium (P.O., Grant number K8.017.15N), (B.M., Grant number G.0593.14); and by the Institute of Education Sciences, U.S. Department of Education (L.J., M.M. and W.V.d.N., Grant number R305D150007). The opinions expressed are those of the authors and do not represent views of the Research Foundation – Flanders or the U.S. Department of Education.

Conflict of Interest

Patrick Onghena has no conflicts of interest to disclose. Bart Michiels has no conflicts of interest to disclose. Laleh Jamshidi has no conflicts of interest to disclose. Mariola Moeyaert has no conflicts of interest to disclose. Wim Van den Noortgate has no conflicts of interest to disclose.

Ethical Standards

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

Supplementary Materials

To view supplementary material for this article, please visit https://doi.org/10.1017/BrImp.2017.25.

References

Aiken, L.S., West, S.G., & Pitts, S.C. (2003). Multiple linear regression. In J. Schinka & W. Velicer (Eds.), *Comprehensive handbook of psychology: Vol. 2. Research methods in psychology* (pp. 481–507. New York, NY: Wiley.

Anumendem, N.D., De Fraine, B., Onghena, P., & Van Damme, J. (2013). The impact of coding time on the estimation of school effects. *Quality & Quantity, 47*, 1021–1040. doi:10.1007/s11135-011-9581-3.

Araujo, A., Julious, S., & Senn, S. (2016). Understanding variation in sets of N-of-1 trials. *PLoS ONE, 11*(12), e0167167. doi:10.1371/journal.pone.0167167.

Baek, E., Moeyaert, M., Petit-Bois, M., Beretvas, S., Van den Noortgate, W., & Ferron, J. (2014). The use of multilevel analysis for integrating single-case experimental design results within a study and across studies. *Neuropsychological Rehabilitation, 24*, 590–606. doi:10.1080/09602011.2013.835740.

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

Bates, D.M. (2010). *Lme4: Mixed-effects modeling with R. Springer*. Retrieved from http://lme4.r-forge.r-project.org/book/.

Beretvas, S.N., & Chung, H. (2008). A review of meta-analyses of single-subject experimental designs: Methodological issues and practice. *Evidence-Based Communication Assessment and Intervention*, *2*, 129–141. doi:10.1080/17489530802446302.

Booth, A., Clarke, M., Dooley, G., Ghersi, D., Moher, D., Petticrew, M., & Stewart, L. (2012). The nuts and bolts of PROSPERO: An international prospective register of systematic reviews. *Systematic Reviews, 1*, 2. doi:10.1186/2046-4053-1-2.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., & Rothstein, H.R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley. doi:10.1002/9780470743386.ch43.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., & Ljung, G.M. (2016). *Time series analysis: Forecasting and control* (5th ed.). New York, NY: Wiley.

Brosnan, J., Moeyaert, M., Brooks Newsome, K., Healy, O., Heyvaert, M., Onghena, P., & Van den Noortgate, W. (2017). Multilevel analysis of multiple-baseline data evaluating precision teaching as an intervention for improving fluency in foundational reading skills for at risk readers. *Exceptionality*. doi:10.1080/09362835.2016.1238378.

Bulté, I., & Onghena, P. (2008). An R package for single-case randomization tests. *Behavior Research Methods, 40*, 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*, 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*, 450–478. doi:10.22237/jmasm/1383280020.

Carey, T.A., & Stiles, W.B. (2016). Some problems with randomized controlled trials and some viable alternatives. *Clinical Psychology & Psychotherapy, 23*, 87–95. doi: 10.1002/cpp.1942.

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press

Cassell, D.L. (2002). A randomization-test wrapper for SAS® PROCs. *SAS User's Group International Proceedings, 27*, 108–111. Retrieved from http://www.lexjansen.com/wuss/2002/WUSS02023.pdf.

Chen, L.-T., Peng, C.-Y.J., & Chen, M.-E. (2015). Computing tools for implementing standards for single-case designs. *Behavior Modification, 39*, 835–869. doi: 10.1177/0145445515603706.

Cohen, L.L., Feinstein, A., Masuda, A., & Vowles, K.E. (2014). Single-case research design in pediatric psychology: Considerations regarding data analysis. *Journal of Pediatric Psychology, 39*, 124–137. doi:10.1093/jpepsy/jst065.

Cools, W., Van den Noortgate, W., & Onghena, P. (2009). Design efficiency for imbalanced multilevel data. *Behavior Research Methods, 41*, 192–203. doi:10.3758/BRM.41.1.192.

Cools, W., Van den Noortgate, W., & Onghena, P. (2008). ML-DEs: A program for designing efficient multilevel studies. *Behavior Research Methods, 40*, 236–249. doi:10.3758/BRM.40.1.236.

Cooper, H., Hedges, L.V., & Valentine, J.C. (Eds.) (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage Foundation.

Dart, E.H., & Radley, K.C. (2017). The impact of ordinate scaling on the visual analysis of single-case data. *Journal of School Psychology, 63*, 105–118. doi:10.1016/j.jsp.2017.03.008.

de Vries, R.M., & Morey, R.D. (2013). Bayesian hypothesis testing for single-subject designs. *Psychological Methods, 18*, 165–185. doi: 10.1037/a0031037.

Douglas, J.M., Knox, L., De Maio, C., & Bridge, H. (2015). Improving communication-specific coping after traumatic brain injury: Evaluation of a new treatment using single-case experimental design. *Brain Impairment, 15*, 190–201. doi: 10.1017/BrImp.2014.25.

Dugard, P., File, P., & Todman, J. (2011). *Single-case and small-N experimental designs*. New York, NY: Routledge Academic.

Edgington, E.S. (1967). Statistical inference from N=1 experiments. *The Journal of Psychology, 65*, 195–199.

Edgington, E.S. (1972). An additive method for combining probability values from independent experiments. *The Journal of Psychology, 80*, 351–363. doi:10.1080/00223980.1972.9924813.

Edgington, E.S. (1975). Randomization tests for one-subject operant experiments. *Journal of Psychology, 90*, 57–68. doi:10.1080/00223980.1975.9923926.

Edgington, E.S. (1980). Validity of randomization tests for one-subject experiments. *Journal of Educational Statistics, 5*, 235–251. doi:10.2307/1164966.

Edgington, E.S. (1996). Randomized single-subject experimental designs. *Behaviour Research and Therapy, 34*, 567–574. doi:10.1016/0005-7967(96)00012-5.

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

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

Fay, M. (2010). *Exact or asymptotic permutation tests: The perm package (version 1.0-0.0) in R*. Retrieved from https://cran.r-project.org/web/packages/perm/perm.pdf.

Ferron, J.M., & Levin, J.R. (2014). Single-case permutation and randomization statistical tests: Present status, promising new developments. In T.R. Kratochwill & J.R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 153–183). Washington, DC: American Psychological Association.

Ferron, J.M., Farmer, J.L., & Owens, C.M. (2010). Estimating individual treatment effects from multiple-baseline data: A Monte Carlo study of multilevel-modeling approaches. *Behavior Research Methods, 42*, 930–943. doi:10.3758/BRM.42.4.930.

Fisher, R.A. (1925). *Statistical methods for research workers*. Edinburgh, UK: Oliver and Boyd.

Fisher, R.A. (1938). Presidential address. *Sankhyā: The Indian Journal of Statistics, 4*, 14–17.

Gast, D.L., & Ledford, J.R. (2014). *Single case research methodology: Applications in special education and behavioral sciences* (2nd ed.). New York, NY: Routledge.

Gentile, J.R., Roden, A.H., & Klein, R.D. (1972). An analysis of variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis, 5*, 193–198.

Guyatt, G., Jaeschke, R., & McGinn, T. (2002). PART 2B1: Therapy and validity. N-of-1 randomized controlled trials. In G. Guyatt, D. Rennie, M.O. Meade, & D.J. Cook (Eds.), *Users' guides to the medical literature* (pp. 275–290). New York, NY: McGraw-Hill.

Harrington, M., & Velicer, W.F. (2015). Comparing visual and statistical analysis in single-case studies using published studies. *Multivariate Behavioral Research, 50*, 162–183. doi:10.1080/00273171.2014.973989.

Hartley, H.O., & Hocking, R.R. (1971). The analysis of incomplete data. *Biometrics, 27*, 783–823.

Heyvaert, M., & Onghena, P. (2014). Randomization tests for single-case experiments: State of the art, state of the science, and state of the application. *Journal of Contextual Behavioral Science, 3*, 51–64. doi:10.1016/j.jcbs.2013.10.002.

Heyvaert, M., Hannes, K., & Onghena, P. (2017). *Using mixed methods research synthesis for literature reviews*. Thousand Oaks, CA: Sage.

Heyvaert, M., Maes, B., & Onghena, P. (2010). A meta-analysis of intervention effects on challenging behaviour among persons with intellectual disabilities. *Journal of Intellectual Disability Research, 54*, 634–649. doi:10.1111/j.1365-2788.2010.01291.x.

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press

Heyvaert, M., Maes, B., & Onghena, P. (2013). Mixed methods research synthesis: Definition, framework, and potential. *Quality & Quantity, 47*, 659–676. doi:10.1007/s11135-011-9538-6.

Heyvaert, M., Maes, B., Van den Noortgate, W., Kuppens, S., & Onghena, P. (2012). A multilevel meta-analysis of single-case and small-*n* research on interventions for reducing challenging behavior in persons with intellectual disabilities. *Research in Developmental Disabilities, 33*, 766–780. doi:10.1016/j.ridd.2011.10.010.

Heyvaert, M., Moeyaert, M., Verkempynck, P., Van den Noortgate, W., Vervloet, M., Ugille, M., & Onghena, P. (2017). Testing the intervention effect in single-case experiments: A Monte Carlo simulation study. *Journal of Experimental Education, 85*, 175–196. doi:10.1080/00220973.2015.1123667.

Heyvaert, M., Wendt, O., Van den Noortgate, W., & Onghena, P. (2015). Randomization and data-analysis items in quality standards for single-case experimental studies. *Journal of Special Education, 49*, 146–156. doi:10.1177/0022466914525239.

Horner, R.H., Carr, E.G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71*, 165–179. doi:10.1177/001440290507100203.

Huitema, B.E., & McKean, J.W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement, 60*, 38–58. doi:10.1177/00131640021970358.

Jacobson, K.H. (2017). *Introduction to health research methods* (2nd ed.). Burlington, MA: Jones & Bartlett.

Jamshidi, L., Heyvaert, M., Declercq, L., Fernández Castilla, B., Ferron, J. M., Moeyaert, M., ... Van den Noortgate, W. (2017). Review of single-subject experimental design meta-analyses and reviews: 1985–2015. Manuscript submitted.

Jones, R.R., Weinrot, R., & Vaught, R.S. (1978). Effects of serial dependence on the agreement between visual and statistical inferences. *Journal of Applied Behavioral Analysis, 11*, 277–283. doi:10.1901/jaba.1978.11-277.

Kaptchuk, T.J. (2001). The double-blind, randomized, placebo-controlled trial: Gold standard or golden calf? *Journal of Clinical Epidemiology, 54*, 541–549. doi:10.1016/S0895-4356(00)00347-4.

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

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

Kratochwill, T.R., & Levin, J.R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods, 15*, 124–144. doi:10.1037/a0017736.

Kratochwill, T.R., & Levin, J.R. (2014). Meta- and statistical analysis of single-case intervention research data: Quantitative gifts and a wish list. *Journal of School Psychology, 52*, 231–235. doi:10.1016/j.jsp.2014.01.003.

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

Kratochwill, T.R., Hitchcock, J.H., Horner, R.H., Levin, J.R., Odom, S.L., Rindskopf, D.M., & Shadish, W.R. (2013). Single-case intervention research design standards. *Remedial and Special Education, 34*, 26–38. doi:10.1177/0741932512452794.

Kravitz, R.L., & Duan, N. (2014). The DEcIDE Methods Center N-of-1 Guidance Panel. In N. Duan, I. Eslick, N.B. Gabler, H.C. Kaplan, R.L. Kravitz, E.B. Larson, W.D. Pace, C.H. Schmid, I. Sim, & S. Vohra (Eds.), *Design and implementation of N-of-1 trials: A user's guide*. AHRQ Publication No. 13(14)-EHC122-EF. Rockville, MD: Agency for Healthcare Research and Quality. Retrieved from https://effectivehealthcare.ahrq.gov/topics/n-1-trials/research-2014-5/.

Kuppens, S., & Onghena, P. (2010). Are there enough pieces to unravel the puzzle? A method to determine sufficiency in single-case research synthesis. *Annual meeting of the american educational research association (AERA)*. Denver, 30 April–4 May 2010.

Kuppens, S., & Onghena, P. (2012). Sequential meta-analysis to determine the sufficiency of cumulative knowledge: The case of early intensive behavioral intervention for children with autism spectrum disorders. *Research in Autism Spectrum Disorders, 6*, 168–176. doi:10.1016/j.rasd.2011.04.002.

Kuppens, S., Heyvaert, M., Van den Noortgate, W., & Onghena, P. (2011). Sequential meta-analysis of single-case experimental data. *Behavior Research Methods, 43*, 720–729. doi:10.3758/s13428-011-0080-1.

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology, 4*, 24. doi:10.1186/s40359-016-0126-3.

Lane, J.D., & Gast, D.L. (2014). Visual analysis in single case experimental design studies: Brief review and guidelines. *Neuropsychological Rehabilitation, 24*, 445–463. doi:10.1080/09602011.2013.815636.

Levin, J.R., Ferron, J. M., & Gafurov, B. S. (2014). Improved randomization tests for a class of single-case intervention designs. *Journal of Modern Applied Statistical Methods, 13*(2), 2–52. doi:10.22237/jmasm/1414814460.

Little, R.J.A., & Rubin, D.B. (1987). *Statistical analysis with missing data*. New York, NY: Wiley.

Lobo, M.A., Moeyaert, M., Cunha, A.B., & Babik, I. (2017). Single-case design, analysis, and quality assessment for intervention research. *Journal of Neurologic Physical Therapy, 41*, 187–197. doi:10.1097/NPT.0000000000000187.

Manolov, R., & Moeyaert, M. (2017a). How can single-case data be analyzed? Software resources, tutorial,

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press

and reflections on analysis. *Behavior Modification, 41*, 179–228. doi:10.1177/0145445516664307.

Manolov, R., & Moeyaert, M. (2017b). Recommendations for choosing single-case data analytical techniques. *Behavior Therapy, 48*, 97–114. doi:10.1016/j.beth.2016.04.008.

Manolov, R., & Solanas, A. (2009). Percentage of nonoverlapping corrected data. *Behavior Research Methods, 41*, 1262–1271. doi:10.3758/BRM.41.4.1262.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

McDonald, S., & Davidson, K.W. (2016). Using N-of-1 methodology to study or change health-related behaviour. *The European Health Psychologist, 18*, 38–42.

Michiels, B., & Onghena, P. (2017). Nonparametric meta-analysis for single-case research: Confidence intervals for combined effect sizes. Manuscript submitted for publication.

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*, 363–381. doi: 10.3758/s13428-016-0714-4.

Moeyaert, M., Ferron, J., 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*, 191–211. doi:10.1016/j.jsp.2013.11.003.

Moeyaert, M., Rindskopf, D., Onghena, P., & Van den Noortgate, W. (2017). Multilevel modeling of single-case data: A comparison of maximum likelihood and Bayesian estimation. *Psychological Methods*. doi:10.1037/met0000136.

Moeyaert, M., Ugille, M., Beretvas, S., Ferron, J., Bunuan, R., & Van den Noortgate, W. (2016). Methods for dealing with multiple outcomes in meta-analysis: A comparison between averaging effect sizes, robust variance estimation and multilevel meta-analysis. *International Journal of Social Research Methodology: Theory & Practice*. doi:10.1080/13645579.2016.1252189.

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S., & Van den Noortgate, W. (2013a). The three-level synthesis of standardized single-subject experimental data: A Monte Carlo simulation study. *Multivariate Behavioral Research, 48*, 719–748. doi:10.1080/00273171.2013.816621.

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S., & Van den Noortgate, W. (2013b). Three-level analysis of single-case experimental data: Empirical validation. *Journal of Experimental Education, 82*, 1–21. doi:10.1080/00220973.2012.745470.

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S., & Van den Noortgate, W. (2014). The influence of the design matrix on treatment effect estimates in the quantitative analyses of single-case experimental design research. *Behavior Modification, 38*, 665–704. doi:10.1177/0145445514535243.

Moeyaert, M., Ugille, M., Ferron, J., Onghena, P., Heyvaert, M, & Van den Noortgate, W. (2015). Estimating intervention effects across different types of single-subject experimental designs: Empirical illustration. *School Psychology Quarterly, 25*, 191–211. doi:10.1037/spq0000068.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., & the PRISMA group (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *PLoS Medicine, 6*(7), e1000097. doi:10.1371/journal.pmed1000097.

Molenaar, P.C.M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research and Perspectives, 2*, 201–211. doi:10.1207/s15366359mea0204_1.

Molenaar, P.C.M., & Campbell, C.G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychology, 18*, 112–117. doi:10.1111/j.1467-8721.2009.01619.x.

Moore, D.S., McCabe, G.P., & Craig, B.A. (2017). *Introduction to the practice of statistics* (9th ed.). New York, NY: D. H. Freeman.

Onghena, P. (1992). Randomization tests for extensions and variations of ABAB single-case experimental designs: A rejoinder. *Behavioral Assessment, 14*, 153–171.

Onghena, P. (2005). Single-case designs. In B. Everitt & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science, vol. 4* (pp. 1850–1854). Chichester, UK: Wiley.

Onghena, P. (2007). N-of-1 randomized clinical trials. In V. Berger (Ed.), *Design and analysis of randomized clinical trials: Design, analysis & theory*. London: Henry Stewart Talks Ltd. Retrieved from https://hstalks.com/t/555/.

Onghena, P. (2016). Randomization in N-of-1 clinical trials: Is it possible to draw causal inferences from single-patient data? In V. Berger (Ed.), *The risk of bias in randomized clinical trials*. London: Henry Stewart Talks Ltd. Retrieved from https://hstalks.com/bs/3311/.

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

Onghena, P., & Struyve, C. (2015). Case studies. In N. Balakrishnan, P. Brandimarte, B. Everitt, G. Molenberghs, W. Piegorsch, & F. Ruggeri (Eds.), *Wiley StatsRef: Statistics reference online* (pp. 1–5). Chichester, UK: Wiley. doi:10.1002/9781118445112.stat06656.pub2.

Onghena, P., Vlaeyen, J.W.S., & de Jong, J. (2007). Randomized replicated single-case experiments: Treatment of pain-related fear by graded exposure in vivo. In S. Sawilowsky (Ed.), *Real data analysis* (pp. 387–396). Charlotte, NC: Information Age Publishing.

Ottenbacher, K.J. (1993). Interrater agreement of visual analysis in single-subjects designs: Quantitative

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press

review and analysis. *American Journal of Mental Retardation, 98*, 135–142.

Ottoboni, K., Lewis, F., & Salmaso, L. (2017). A comparison of parametric and permutation tests for regression analysis of randomized experiments. arXiv:1702.04851v1.

Peat, J. (2002). *Health science research: A handbook of quantitative methods*. London, UK: Sage.

Perdices, M., & Tate, R.L. (2009). Single-subject designs as a tool for evidence-based clinical practice: Are they unrecognised and undervalued? *Neuropsychological Rehabilitation, 19*, 904–927. doi:10.1080/09602010903040691.

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

Pinheiro, J.C., & Bates, D.M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.

Piwek, L., Ellis, D. A., Andrews, S., & Joinson, A. (2016). The rise of consumer health wearables: Promises and barriers. *PLoS Medicine, 13*(2), e1001953. doi:10.1371/journal.pmed.1001953.

Punja, S., Schmid, C.H., Hartling, L., Urichuk, L., Nikles, C.J., & Vohra, S. (2016). To meta-analyze or not to meta-analyze? A combined meta-analysis of N-of-1 trial data with RCT data on amphetamines and methylphenidate for pediatric ADHD. *Journal of Clinical Epidemiology, 76*, 76–81. doi:10.1016/j.jclinepi.2016.03.021.

Pustejovsky, J.E., Hedges, L.V., & Shadish, W.R. (2014). Design-comparable effect sizes in multiple baseline designs: A general modeling framework. *Journal of Educational and Behavioral Statistics, 39*, 368–393. doi:10.3102/1076998614547577.

Rosenthal, R. (1978). Combining the results of independent studies. *Psychological Bulletin, 85*, 185–193. doi:10.1037/0033-2909.85.1.185.

Sackett, D.L., Rosenberg, W.M., Gray, J.A., Haynes, R.B., & Richardson, W.S. (1996). Evidence based medicine: What it is and what it isn't. *BMJ, 312*, 71–72. doi:10.1136/bmj.312.7023.71.

Scargle, J.D. (2000). Publication bias: The "file-drawer" problem in scientific inference. *Journal of Scientific Exploration, 14*, 91–106.

Schlosser, R.W., Lee, D.L., & Wendt, O. (2008). Application of the percentage of non-overlapping data (PND) in systematic reviews and meta-analyses: A systematic review of reporting characteristics. *Evidence-Based Communication Assessment and Intervention, 2*, 163–187. doi:10.1080/17489530802505412.

Schork, N.J. (2015). Personalized medicine: Time for one-person trials. *Nature, 520*(7549), 609–611. doi:10.1038/520609a.

Scruggs, T.E., & Mastropieri, M.A. (2013). PND at 25: Past, present, and future trends in summarizing single-subject research. *Remedial and Special Education, 34*, 9–19. doi:10.1177/0741932512440730.

Senn, S.J. (2017, 8 February 2017). *Consult two medics and you'll get two opinions but consult two statisticians and you could easily get three #thewonderofstats [Twitter moment]*. Retrieved from https://twitter.com/stephensenn/status/829593923123343363.

Shadish, W.R. (2014). Analysis and meta-analysis of single-case designs: An introduction. *Journal of School Psychology, 52*, 109–122. Retrieved from http://doi.org/10.1016/j.jsp.2013.11.009.

Shadish, W.R., & Sullivan, K.J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43*, 971–980. doi:10.3758/s13428-011-0111-y.

Shadish, W.R., Kyse, E.N., & Rindskopf, D.M. (2013). Analyzing data from single-case designs using multilevel models: New applications and some agenda items for future research. *Psychological Methods, 18*, 385–405. doi:10.1037/a0032964.

Shadish, W.R., Rindskopf, D.M., & Boyajian, J.G. (2016). Single-case experimental design yielded an effect estimate corresponding to a randomized controlled trial. *Journal of Clinical Epidemiology, 76*, 82–88. doi:10.1016/j.jclinepi.2016.01.035.

Shadish, W.R., Zelinsky, N.A., Vevea, J.L., & Kratochwill, T.R. (2016). A survey of publication practices of single-case design researchers when treatments have small or large effects. *Journal of Applied Behavioral Analysis, 49*, 656–673. doi:10.1002/jaba.308.

Shamseer, L., Sampson, M., Bukutu, C., Schmid, C.H., Nikles, J., Tate, R., … & the CENT group (2015). CONSORT extension for reporting N-of-1 trials (CENT) 2015: Explanation and elaboration. *British Medical Journal, 350*, h1793. doi:10.1136/bmj/h1793.

Shapiro, M.B. (1966). The single case in clinical-psychological research. *The Journal of General Psychology, 74*, 3–23. doi:10.1080/00221309.1966.9710306.

Shea, B.J., Hamel, C., Wells, G.A., Bouter, L.M., Kristjansson, E., Grimshaw, J., … Boers, M. (2009). AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *Journal of Clinical Epidemiology, 62*, 1013–1020. doi:10.1016/j.jclinepi.2008.10.009.

Shine, L.C., & Bower, S.M. (1971). A one-way analysis of variance for single-subject designs. *Educational and Psychological Measurement, 31*, 105–113. doi:10.1177/001316447103100108.

Sidman, M. (1952). A note on functional relations obtained from group data. *Psychological Bulletin, 49*, 263–269. doi:10.1037/h0063643.

Simmons, J.P., Nelson, L.D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366. doi:10.1177/0956797611417632.

Smith, J.D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods, 17*, 510–550. doi:10.1037/a0029312.

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press

Solmi, F., & Onghena, P. (2014). Combining P-values in replicated single-case experiments with multivariate outcome. *Neuropsychological Rehabilitation, 24*, 607–633. doi:10.1080/09602011.2014.881747.

Solomon, B.G. (2014). Violations of assumptions in school-based single-case data: Implications for the selection and interpretation of effect sizes. *Behavior Modification, 38*, 477–496. doi:10.1177/0145445513510931.

Sterba, S.K. (2009). Alternative model-based and design-based frameworks for inference from samples to populations: From polarization to integration. *Multivariate Behavioral Research, 44*, 711–740. doi:10.1080/00273170903333574.

Strube, M.J., & Miller, R.H. (1986). Comparison of power rates for combined probability procedures: A simulation study. *Psychological Bulletin, 99*, 407–415. doi:10.1037/0033-2909.99.3.407.

Tate, R. L., Aird, V., & Taylor, C. (2013). Bringing single-case methodology into the clinic to enhance evidence-based practices. *Brain Impairment, 13*, 347–359. doi:10.1017/BrImp.2012.32.

Tate, R.L., Perdices, M., Rosenkoetter, U., McDonald, S., Togher, L., Shadish, W., ... Vohra, S. (2016a). The single-case reporting guideline in BEhavioural Interventions (SCRIBE) 2016: Explanation and elaboration. *Archives of Scientific Psychology, 4*, 10–31. doi:10.1037/arc0000027.

Tate, R.L., Perdices, M., Rosenkoetter, U., Shadish, W., Vohra, S., Barlow, D. H., ... Wilson, B. (2016b). The single-case reporting guideline In BEhavioural interventions (SCRIBE) 2016 statement. *Archives of Scientific Psychology, 4*, 1–9. doi:10.1037/arc0000026.

Tate, R.L., Perdices, M., Rosenkoetter, U., Wakim, D., Godbee, K., Togher, L., & McDonald, S. (2013). Revision of a method quality rating scale for single-case experimental designs and *N*-of-1 trials: The 15-item Risk of Bias in *N*-of-1 Trials (RoBiNT) Scale. *Neuropsychological Rehabilitation, 23*, 619–638. Retrieved from https://doi.org/10.1080/09602011.2013.824383.

Terrin, N., Schmid, C.H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine, 22*, 2113–2126. doi:10.1002/sim.1461.

Tierney, J.F., Vale, C., Riley, R., Smith, C.T., Stewart, L., Clarke, M., & Rovers, M. (2015). Individual participant data (IPD) meta-analyses of randomised controlled trials: Guidance on their use. *PLoS Medicine, 12*(7), e1001855. doi:10.1371/journal.pmed.1001855.

Tukey, J.W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 24*, 83–91. doi:10.1037/h0027108.

Turner, L., Shamseer, L., Altman, D. G., Weeks, L., Peters, J., Kober, T., ... Moher, D. (2012). Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. *Cochrane Database of Systematic Reviews, 11*, MR000030. Retrieved from https://doi.org/10.1002/14651858.mr000030.pub2.

Ugille, M., Moeyaert, M., Beretvas, S., Ferron, J., & Van den Noortgate, W. (2012). Multilevel meta-analysis of single-subject experimental designs: A simulation study. *Behavior Research Methods, 44*, 1244–1254. doi:10.3758/s13428-012-0213-1.

Vallverdú, J. (2016). *Bayesians versus frequentists: A philosophical debate on statistical reasoning*. New York, NY: Springer.

Van den Noortgate, W., & Onghena, P. (2003a). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly, 18*, 325–346.

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

Van den Noortgate, W., & Onghena, P. (2003c). Multilevel meta-analysis: a comparison with traditional meta-analytical procedures. *Educational and Psychological Measurement, 63*, 765–790. doi:10.1177/0013164403251027.

Van den Noortgate, W., & Onghena, P. (2007). The aggregation of single-case results using hierarchical models. *Behavior Analyst Today, 8*, 196–208. doi:10.1037/h0100613.

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment and Intervention, 2*, 142–151. doi:10.1080/17489530802505362.

Van den Noortgate, W., López-López, J., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods, 45*, 576–594. doi:10.3758/s13428-012-0261-6.

Van den Noortgate, W., López-López, J., Marín-Martínez, F., & Sánchez-Meca, J. (2015). Meta-analysis of multiple outcomes: A multilevel approach. *Behavior Research Methods, 47*, 1274–1294. doi:10.3758/s13428-014-0527-2.

Vanderkerken, L., Heyvaert, M., Maes, B., & Onghena, P. (2013). Psychosocial interventions for reducing vocal challenging behaviour in persons with autistic disorder: A multilevel meta-analysis of single-case experiments. *Research in Developmental Disabilities, 34*, 4515–4533. doi:10.1016/j.ridd.2013.09.030.

Velicer, W.F., & Molenaar, P. (2013). Time series analysis. In J. Schinka & W.F. Velicer (Eds.), *Handbook of psychology: Research methods in psychology* (2nd ed., Vol. *50*, pp. 628–660). New York: Wiley.

Vohra, S., Shamseer, L., Sampson, M., Bukutu, C., Schmid, C.H., Tate, R., ... & the CENT group (2015). CONSORT extension for reporting N-of-1

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press

trials (CENT) 2015 Statement. *British Medical Journal, 350*, h1738. doi:10.1136/bmj/h1738.

Wheeler, B., & Torchiano, M. (2016). *Permutation tests for linear models: the lmPerm package (version 2.1.0) in R*. Retrieved from https://cran.r-project.org/web/packages/lmPerm/lmPerm.pdf.

Wicherts, J.M., Veldkamp, C.L.S., Augusteijn, H.E.M., Bakker, M., van Aert, R.C.M., & van Assen, M.A. L.M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology, 7*, 1832. doi:10.3389/fpsyg.2016.01832.

https://doi.org/10.1017/BrImp.2017.25 Published online by Cambridge University Press