

**Further developments in summarising and meta-analysing single-case data:**

**An illustration with neurobehavioural interventions in acquired brain injury**

Rumen Manolov<sup>1,2</sup> and Lucien Rochat<sup>3,4</sup>

<sup>1</sup>Department of Behavioural Sciences Methods, University of Barcelona, Spain

<sup>2</sup>Institute for Brain, Cognition and Behavior (IR3C), University of Barcelona, Spain

<sup>3</sup>Cognitive Psychopathology and Neuropsychology Unit, University of Geneva, Switzerland

<sup>4</sup>Swiss Centre for Affective Sciences, University of Geneva, Switzerland

**Running head:** Summarising single-case data

**Word count:** 8856 (including references)

**Contact author**

Correspondence concerning this article should be addressed to Rumen Manolov, Departament de Metodologia de les Ciències del Comportament, Facultat de Psicologia, Universitat de Barcelona, Passeig de la Vall d'Hebron, 171, 08035-Barcelona, Spain. Phone number: +34934031137. Fax: +34934021359. E-mail: rrumenov13@ub.edu.

**Acknowledgement**

This work was partially supported by the Agència de Gestió d'Ajust Universitaris i de Recerca de la Generalitat de Catalunya grant 2014SGR71.

**Abstract**

Data analysis for single-case designs is an issue that has prompted many researchers to propose a variety of alternatives, including use of randomisation tests, regression-based procedures, and standardised mean difference. Another option consists in computing unstandardised or raw differences between conditions: the changes in slope and in level, or the difference between the projected baseline (including trend) and the actual treatment phase measurements. Apart from the strengths of these procedures (potentially easier interpretation clinically, separate estimations and an overall quantification of effects, reasonable performance), they require further development, such as (a) creating extensions for dealing with methodologically strong designs such as multiple baseline, (b) achieving comparability across studies and making possible meta-analytical integrations, and (c) implementing software for the extensions. The proposals are illustrated herein in the context of a meta-analysis of 28 studies on (neuro)behavioural interventions in adults who have challenging behaviours after acquired brain injury.

**Keywords:** single case, effect size, multiple-baseline designs, challenging behaviours, rehabilitation

**Further developments in summarising and meta-analysing single-case data: An illustration with neurobehavioural interventions in acquired brain injury**

Single-case experimental designs (SCEDs) are recognized as being useful for providing solid evidence for professional practices in several behavioural disciplines, including the treatment of people who have brain impairment (Perdices & Tate, 2010). This benefit of SCEDs comes from their methodological strengths (Horner et al., 2005; Kratochwill & Levin, 2010) and the possibility of computing effect size indices and performing meta-analyses of studies on the same intervention (Jenson, Clark, Kircher, & Kristjansson, 2007). Our aim in the current study is to propose an extension of unstandardised indices in order to (a) compute a single summary measure per study (e.g., from several tiers in a multiple-baseline design [MBD] or from several two-phase comparisons in a withdrawal designs) and (b) quantitatively integrate the outcomes of several studies. We illustrate these new developments with a meta-analysis of studies on neurobehavioural interventions to decrease problematic behaviours in adults with an acquired brain injury (ABI).

Regarding the SCED structures being well-suited for building evidence-based interventions, Kratochwill et al. (2010, 2013) and Tate et al. (2013) have emphasized the need for several transitions between phases with and without intervention and the importance of deciding the points of change in phase at random. Several possibilities exist for meeting the methodological criterion regarding the design structure: alternating treatments designs, reversal/withdrawal designs, MBDs, etc. Not all analytical alternatives to be applied to such design structures are equally straightforward.

## SCED analytical techniques

Among the current developments in the area of SCED data analysis, the most promising procedures include the SCED-specific *d*-statistic (Hedges, Pustejovsky, & Shadish, 2012, 2013), multilevel models (e.g., Ferron, Moeyaert, Van den Noortgate, & Beretvas, 2014; Moeyaert, Ferron, Beretvas, & Van den Noortgate, 2014), and randomisation tests (Levin, Ferron, & Kratochwill, 2012; Heyvaert & Onghena, 2014). All of these analytical options, in contrast to non-overlap indices (e.g., percentage of non-overlapping data: Scruggs & Mastropieri, 2013; Tau-U: Parker, Vannest, Davis, & Sauber, 2011), directly incorporate the option of summarising the results of an MBD by considering all comparisons between a baseline and a subsequent intervention phase. Nevertheless, we consider it necessary to propose a way to provide summary indices for other existing SCED analytical procedures that are applicable to MBD; specifically, we focus on the slope and level change (SLC; Solanas, Manolov, & Onghena, 2010) and the mean phase difference (MPD; Manolov & Solanas, 2013) procedures. The reason for this choice can be found in the desirable features of these indicators, as well as in the limitations of the above-mentioned procedures.

Regarding the MPD and the SLC, these procedures offer quantitative information in the same metric as the dependent variable or behaviour of interest being measured (e.g., number of cigarettes smoked, number of interactions initiated, number of words read). The joint use of these procedures answers (a) Beretvas and Chung's (2008) call for separate estimation of different effects, as the SLC quantifies change in slope and then the net change in level, something that is also possible with multilevel models (Van den Noortgate & Onghena, 2008) and (b) Swaminathan, Rogers, and Horner's (2014) emphasis on the need for quantification of the overall effect, as the MPD offers single quantification. Moreover, these procedures have

shown acceptable performance (Manolov & Solanas, 2013; Manolov, Solanas, Sierra, & Evans, 2011; Solanas et al., 2010) and are accompanied by easy-to-use code in the open-source software R, which makes their use straightforward.<sup>1</sup>

The *d*-statistic offers quantification in terms of a standardised mean difference, which takes autocorrelation into account and can be corrected against small-sample bias. Moreover, the effect size obtained is accompanied by its variance, which can be used in meta-analysis. The limitations of this indicator include the assumption of lack of baseline trend and the fact that it is not possible to obtain separate quantification for each of the tiers in the MBD.

Multilevel models are flexible in terms of the aspects being modelled (e.g., autocorrelation, trend, variation in trend and in intervention effectiveness across tiers). Nevertheless, their performance for estimating variances is less than optimal unless the series are long (Moeyaert, Ugille, Ferron, Beretvas, & Van Den Noortgate, 2013; Ugille, Moeyaert, Beretvas, Ferron, & Van den Noortgate, 2012). Additionally, conducting the analysis and interpreting the results (most frequently done in relation to statistical rather than to clinical significance) is not straightforward and requires a certain amount of statistical knowledge and training.

Randomisation tests are applicable to several design structures and allow one to define a test statistic according to the effect expected. Nevertheless, the importance of the results is judged in terms of statistical significance; in addition, it is not possible to focus on each tier separately. Finally, randomisation tests require the desirable (Kratochwill & Levin, 2010) but difficult-to-implement (Fisher & Lerman, 2014) random assignment of conditions to the measurement times.

---

<sup>1</sup> For more details about how the previously developed R code can be used, consult the original articles (Manolov & Solanas, 2013; Solanas et al., 2010) and the supplementary material of the article by Manolov, Gast, Perdices, and Evans (2014).

## Study aims and organization of the paper

We here continue developing the unstandardised indices SLC and MPD, as they allow interpretation of the results in potentially more meaningful terms (i.e., not in terms of standard deviations or  $p$  values, but in the measurement units of the variable of interest; Cumming, 2012). First, we present a modified version of the MPD in order to improve the way in which the baseline trend is fitted to the data. Second, we focus on the within-study level of analysis, proposing two different ways in which a single effect size can be obtained for comparisons performed according to the design structure. This step is specifically necessary to avoid dependencies between effect sizes and a greater influence of the results from a specific investigation in a meta-analysis.<sup>2</sup> Third, we focus on the between-studies level of analysis, proposing two ways in which the MPD and SLC values can be made comparable when different metrics are used in different studies. Finally, we provide user-friendly R code and a step-by-step manual on how to use it.

### A New Version of the MPD: Modification to Improve Fit to the Data

The original version of the MPD compares the obtained intervention measurements with the projection of the baseline trend, as extended from the first baseline phase data point (adding the estimate of the baseline trend times the order of the measurement). After several applications of the procedure, we decided to change the way in which the baseline trend is fitted, choosing as a pivotal point the middle point in the baseline on the abscissa (3 if there are 5 measurements; 3.5 if there are 6 measurements, etc.) and the median measurement on the ordinate. The slope of the

---

<sup>2</sup> For instance, a criticism by Baron and Derenne (2000) of an early quantitative integration of SCED results was directed toward using more than one outcome from an experiment.

line was defined from the estimated baseline trend and consisted in subtracting the trend value for the measurements prior to the middle point and adding it to the measures after the middle point. This procedure (finding middle points and medians) is similar to the split-middle method (Miller, 1985), but the trend is not estimated according to it. The modified procedure allowed us to fit the trend better to the baseline measurements before extending it as illustrated in Figure 1 which shows that baseline trend fitted by new version of the MPD matches better the data than the previous version. This is why in the following we use only this modified version.

### INSERT FIGURE 1 ABOUT HERE

The expression of this modified version can be written as  $MPD = \sum_{j=1}^{n_B} (y_j - \hat{y}_j) / n_B$ , the same as for the original version. Thus, it still reads that for all  $n_B$  measurements in the intervention phase we are comparing each actually observed measurement  $y_j$  with each measurement  $\hat{y}_j$  predicted by projecting the baseline trend. However, in this case, the way in which the predicted treatment phase data  $\hat{y}_i$  are obtained can be summarised as follows:

(1) Estimate the baseline trend through differencing as  $trend_A = \sum_{i=1}^{n_A-1} (y_{i+1} - y_i) / (n_A - 1)$ .

This expression reads that baseline trend is estimated as the average of the difference between each measurement  $y_i$  and each subsequent measurement  $y_{i+1}$ , focussing on all baseline  $n_A$  data points.

(2) Establish the pivotal point in the baseline at the crossing of  $Md(x) = Md(1, 2, \dots, n_A)$  on the abscissa and  $Md(y) = Md(y_1, y_2, \dots, y_{n_A})$  on the ordinate. This step implies that we are selecting as pivotal point the crossing between the median of the  $n_A$  baseline measurements and

the middle point in the series of measurement occasions  $Md(x)$ , which is equal to  $\frac{n_A+1}{2}$  if  $n_A$  is an odd number and to  $\frac{n_A+0.5}{2}$  if  $n_A$  is an even number.

(3) Establish a fitted value at an existing baseline measurement occasion by

$$\hat{y}_{[Md(x)]} = Md(y) - (Md(x) - [Md(x)]) \times trend_A .$$

This expression says that when the number of baseline measurements  $n_A$  is odd and  $Md(x)$  is a whole number, it is an actual measurement occasion and the pivotal point gets a fitted value equal to the median of the baseline measurements  $Md(y)$ . However, when  $n_A$  is an even number,  $Md(x)$  is a whole number and we use the immediately previous measurement occasion to fit a value, which is equal to the median  $Md(y)$  minus half the baseline trend.

(4) Fit the baseline trend to the whole baseline by

$$\hat{y}_i \begin{cases} \hat{y}_{[Md(x)]} - ([Md(x)] - i) \times trend_A \text{ for } i = 1, 2, \dots, [Md(x)] - 1 \\ \hat{y}_{[Md(x)]} + (i - [Md(x)]) \times trend_A \text{ for } i = [Md(x)] + 1, [Md(x)] + 2, \dots, n_A \end{cases} .$$

This expression states that for each measurement occasion after the first value fitted in the middle of the baseline phase, we add baseline trend as many times as each occasion is apart from the first value. Additionally, for each measurement occasion before the first value fitted in the middle of the baseline phase, we subtract baseline trend as many times as each occasion is apart from the first value.

(5) Project the baseline trend into the treatment phase as

$$\hat{y}_i = \hat{y}_{n_A} + i \times trend_A \text{ for } i = 1, 2, \dots, n_B .$$



The last expression states that, for each of the  $n_B$  intervention phase measurement occasions we fit a predicted value that entails adding the estimated baseline trend as as many times as each occasion is apart from the last fitted baseline value.

### **Within-Study Level of Analysis: A Single Effect Size per Study**

#### **Rationale**

A summary at the study level and a meta-analysis require a single effect size per study to be computed. We here follow one of the alternative ways of achieving this, namely, obtaining the average of the effect sizes reported in a study (Borenstein, Hedges, Higgins, & Rothstein, 2009). Such a practice has been deemed justified when the outcomes measure the same construct (Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013), which is the case when the same behaviour is measured across conditions. For obtaining a weighted average, the weight of the quantification for each tier can be based on the number of measurements in the tier (as suggested by Kratochwill et al., 2010; Shadish, Rindskopf, & Hedges, 2008), while also taking into account baseline stability around an increasing or decreasing trend (Hedges et al., 2012). This latter feature is especially important for the MPD and the SLC, as they both estimate the baseline linear trend as an initial step; the degree to which this initial estimate is a good representation of the data has considerable influence on the subsequent quantifications of the behavioural change. The current proposal is also well aligned with the observation that an inaccurately modelled trend (e.g., assumed linear when it is non-linear) can distort the results of SCED analytical techniques (Sullivan, Shadish, & Steiner, 2014).

The weight is defined as:

$$w_i = (n_{Ai} + n_{Bi}) + \frac{1}{MSE_i}, \text{ where } MSE_i = \frac{\sum_j^{n_{Ai}} (y_j - \hat{y}_j)^2}{n_{Ai}},$$

where  $i$  represents each of the tiers and  $n_{Ai}$  and  $n_{Bi}$  the number of measurements, and where MSE denotes mean square error, that is, the sum of squared differences between fitted ( $\hat{y}_j$ ) and actual baseline data points ( $y_j$ ), which is afterwards divided by the number of baseline measurements ( $n_{Ai}$ ). Finally, the weighted (within-studies) average is equal to  $ES_{MBD} = \frac{\sum_{i=1}^{tiers} ES_i \cdot w_i}{\sum_{i=1}^{tiers} w_i}$ .

It is also possible to deal with the weighting issue in a more classic way, namely, using only series length as a weight. Apart from Kratochwill et al. (2010) and Shadish et al. (2008), Beeson and Robey (2006) explicitly recommended this weight when obtaining a single effect size for several individuals in an MBD.

### **Example: Application to multiple-baseline data**

In order to illustrate the proposals made here, we use the data gathered by Alderman and Knight (1997), a data set that was chosen because it presents challenging issues such as improving baseline trends and marked differences in baseline data variability, which we consider useful for illustrating the analytical procedures to practitioners. The data refer to a 58-year-old man who had multiple injuries (including haemorrhage in the internal capsule and damage in the right occipital lobe) in a traffic accident. Physical and verbal aggressive behaviours are treated by using differential reinforcement of low rates of responding applied to various problematic behaviours such as throwing objects, shouting, making sexual comments, and swearing.

Figure 2 presents the data for each of the tiers, as well as the estimated baseline trend represented as a straight line. It is visually clear that the estimate of the baseline trend is closer to the actual baseline phase measurements for the first tier than it is for the remaining tiers, as the mean square error reflects. Other information available includes the quantifications of behavioural change in terms of MPD and the two effects of SLC: slope change and net level change.

#### INSERT FIGURE 2 ABOUT HERE

These data illustrate two aspects. First, correcting for improving baselines may lead to quantification, suggesting that the undesirable behaviour has increased after the intervention, as is the case for the sexual comments and swearing. This is contrary to the visual impression of the data, but they reflect the fact that the rate of improvement is no longer maintained after the intervention. However, there is no further improvement because problematic behaviour is reduced to a minimum and cannot decrease any further. Such data provide a dilemma: to control for the baseline trend and underestimate the intervention effect, or not to control and to overestimate it. The use of MPD and SLC offers the former, that is, the conservative solution. Second, regarding the influence of trend stability, the weight assigned to the tier related to throwing is very large, as the linear estimation of the baseline trend is a better representation of the actual data in this tier (i.e., its MSE is very low). The weighted average is strongly influenced by this outcome (e.g., the weighted average MPD =  $-0.91$ ), as can be seen in Figure 3. This strip chart, which is part of the user-friendly code developed, also offers information about the variability of outcomes at the within-study level, which would be used for assigning a weight to the study effect size when performing across-studies integration. For this data set, using only series length as a weight would have led to a weighted average, suggesting deterioration, and an

increase in the undesired behaviours ( $MPD = 1$ ), as Tier 1 is no longer influential on the results, given that Tiers 2 and 4 have as many as, or more, observations.

INSERT FIGURE 3 ABOUT HERE

### **Applicability of the proposals to different design structures**

Regarding the application to several design structures, it has been claimed that there are still no clear guidelines on how to obtain a single effect size for an individual or a study (Maggin et al., 2011). For instance, alternating treatment designs are not easily analysable with MPD and SLC, as the frequent change of conditions does not allow one to estimate baseline trends with sufficient precision. The analysis of ABAB designs is also not straightforward, given that the comparison between the first intervention ( $B_1$ ) and the second baseline (i.e., withdrawal;  $A_2$ ) phase may be problematic because of an incomplete return to initial baseline levels (Parker & Vannest, 2012). Regarding this issue, Strain, Kohler, and Gresham (1998) recommended using the quantification only of the initial AB comparison (which is what Parker et al., 2011, did when illustrating Tau-U), whereas Olive and Smith (2005) suggested comparing only the initial and final conditions, omitting both phases ( $B_1$  and  $A_2$ ), a practice followed in Heinicke and Carr's (2014) meta-analysis. These two proposals are related to Scruggs and Mastropieri's (1998) suggestion to perform only those comparisons that maintain the A-B sequence. As another option, a recent comparison between Tau-U and Allison and Gorman's (1993) regression model was performed only on MBDs (Ross & Begeny, 2014).

Regarding the current proposals, we suggest combining all two-phase (AB) comparisons in the same fashion for all design structures, as was illustrated for MBDs. In fact, the order of the phases could be the inverse (i.e., BA), given that the comparisons made via MPD and SLC

would still focus on the degree to which the existing trend in the data is continued after the change in conditions (in the BA case, after withdrawing the intervention). For designs involving more than one change in conditions (e.g., ABAB and extensions), we propose omitting the  $B_1A_2$  comparison (and subsequent comparisons), as the data from these phases would be overrepresented in the quantification. Specifically, the idea not to include the  $B_1A_2$  comparison can be related to the discussion on dependence between outcomes: If quantification of a two-phase comparison is considered an effect size, it is warranted to combine it with other effect sizes, as if independent, only if it comes from a different sample (Littell, Corcoran, & Pillai, 2008). However, in the case of the  $B_1A_2$  comparison, the data belong to the same sample of behaviour as in the  $A_1B_1$  and  $A_2B_2$  comparisons. In contrast, we consider the  $B_1A_2$  comparison to be crucial for assessing intervention effectiveness (Kratochwill et al., 2010) and it should be taken into account in the visual analysis informing the decision about whether there is a functional relation between the intervention and the target behaviour.

### **Across-Studies Level of Analysis: Comparability and Meta-Analytical Integration**

#### **Dealing with different operative definitions**

The drawback of having effect size measures expressed in terms of the measurement units of the behaviour of interest is that different studies are likely to use different operative definitions of the target behaviour. In contrast to the MPD and the SLC, the  $d$ -statistic is expressed in standard deviations, whereas multilevel models can be applied meta-analytically to both raw data and standardised data (Van den Noortgate & Onghena, 2008). Non-overlap measures are also expressed in the same metric across studies, allowing their use in meta-analyses when the same

indicator is applied to all data sets (e.g., Ganz et al., 2012; Jamieson, Cullen, McGee-Lennon, Brewster, & Evans, 2014).

In order to achieve comparability, we here propose to transform the unstandardised indices into percentages. We have chosen percentages of change in the behaviour measure as a quantification in order to improve interpretability, which is well aligned with the search for measures that are more meaningful to applied researchers than are standardised mean differences (e.g., Pustejovsky, 2014, explicitly mentions percentage change in his recent proposals for SCED effect sizes). This transformation into a percentage change index is analogous to the calculation in the mean baseline reduction (e.g., as used by Herzinger & Campbell, 2007) and it is also related to the log response ratio measure (Pustejovsky, 2014).

Focusing first on the MPD, its percentage version quantifies the relative difference between actual ( $y_j$ ) and predicted intervention phase measurements ( $\hat{y}_j$ ) as a percentage of the predicted value. The expression for this indicator is as follows:  $MPD_{Percentage} = \left[ \sum_{j=1}^{n_B} \frac{100(y_j - \hat{y}_j)}{|\hat{y}_j|} \right] / n_B$ .

Regarding the SLC, this procedure includes two quantifications of effect size of two distinct types of effect. The slope change estimate quantifies the amount of progressive change during the intervention phase, after the baseline trend is eliminated. Therefore, the percentage version of this estimate represents the difference between the intervention phase trend ( $trend_B$ ) and the baseline trend ( $trend_A$ ) relative to the baseline trend:

$SC_{Percentage} = 100(trend_B - trend_A) / |trend_A|$ . The level change estimate of the SLC is

simply the mean difference between the baseline measurements after the baseline trend is

controlled for (average equal to  $\overline{\tilde{X}_A}$ ) and the intervention phase measurements after the baseline

trend and slope change are controlled for (average equal to  $\overline{\tilde{X}_B}$ ). The percentage version represents the average change relative to the baseline level:

$$LC_{Percentage} = 100 \left( \overline{\tilde{X}_B} - \overline{\tilde{X}_A} \right) / |\overline{\tilde{X}_A}|.$$

We consider this conversion of the original indices into percentages not to hamper their meaningfulness, given that (a) the unstandardised version is still available and (b) the percentage increase in behavioural level or in trend is also a useful way of summarising the data. We must mention that, among the limitations of the percentage versions of the indices, is the impossibility of obtaining numerical results when the denominator is equal to zero. For the MPD, this means that a comparison between an actual intervention data point and a predicted one is omitted if the latter is equal to zero. For the SLC, the undesirable case is when either the baseline trend is exactly equal to zero, or the baseline level after it has detrended is exactly equal to zero. Finally, it is possible to obtain very large values in some cases in which, for instance, the original metric is in percentages, rising from, for example, 2% to 100% (an increase of 5000%). This is why we would like to stress that the attainment of a comparable index across studies can also be achieved by standardising the MPD and the SLC. One manner of standardising is dividing the values by the standard deviation of the baseline data (as in Glass, McGaw, & Smith's, 1981,  $\Delta$  index), a procedure that does not consider the variability in the treatment phase, given that any improving trends might be confounded with unexplained variation in the data.

### **Weighting strategies**

When combining the effect sizes from different studies, we consider that the weight given to each effect should be once again based on the amount of data points available, following the

suggestions by the experts in the field (Kratochwill et al., 2010; Shadish et al., 2008), as well as an additional piece of information. In this case, when integrating (across-studies) summary values that have themselves been obtained after summarising (within-studies) individual outcomes, we consider it important to reflect in the weight how well these effect sizes represent the different outcomes within a study. This is why we propose using the inverse of the variability around the overall effect size as part of the weight for this effect size in the process of meta-analysis. This weighting strategy is well aligned with the attention paid to the importance of within-study variability of effects in the context of other techniques applicable to single-case data analysis (Van den Noortgate et al., 2013), in response to the observation that within-study heterogeneity is a relevant piece of information for meta-analysis (Cheung & Chan, 2004).

When defining the weight of the effect size per study, our initial intention was to mirror the way in which a random effects meta-analysis is performed (for more information, see Chapters 14 and 16 of Borenstein et al., 2009). In random effects models, a weight is assigned to an outcome according to the inverse variance of this outcome (closely related to the number of measurements available) and the variability of the outcomes across studies around their mean. In a similar fashion, we wanted to assign a weight according to the number of measurements in the study and the variability of the outcomes within (rather than across) the study, using the variance indicator called tau-squared, defined as shown below:

$$\tau^2 = \frac{\sum_{i=1}^{tiers} w_i (ES_i - ES_{MBD})^2 - (tiers - 1)}{\sum_{i=1}^{tiers} w_i - \frac{\sum_{i=1}^{tiers} w_i^2}{\sum_{i=1}^{tiers} w_i}}$$

According to this approach, the weight for the effect size of an individual study would have been defined as follows:  $w_k = \sum_{i=1}^{tiers} (n_{Ai} + n_{Bi}) + 1/\tau^2$ . However, borrowing this weighting strategy



from a group-designs random effects meta-analysis is not completely justified, for two reasons:

(a) The expression for tau-squared is used when the  $w_i$  weights are inverse variances of the effect size index ( $v_i$ ), which are not available for the MPD and SLC procedures, for which we used  $w_i = (n_{Ai} + n_{Bi}) + 1/MSE_i$  instead; and (b) the weight for the study effect size  $w_k$  would have been defined as  $w_k = 1/(v_i + \tau^2)$ . This is arguably the reason for obtaining (in a preliminary analysis not shown here) excessively high values for tau-squared; thus,  $w_k$  was reduced to the amount of measurements available. Instead of using  $\tau^2$ , we defined operatively the within-study variability of outcomes via an indicator analogous to the coefficient of variation, using the

weighted mean as a reference:  $CV'_k = \frac{\sqrt{\sum_{i=1}^{tiers} (ES_i - ES_{MBD})^2 / tiers}}{|ES_{MBD}|}$ . The expression for the weight of

an overall effect size for the  $k$ th study using an MBD with a certain amount of *tiers* (or a

withdrawal design with as many two-phase comparisons) is  $w_k = \sum_{i=1}^{tiers} (n_{Ai} + n_{Bi}) + \frac{1}{CV'_k}$ , and

the meta-analytical, weighted average across  $k$  studies is obtained as

$$\overline{ES} = \sum_{k=1}^{studies} ES_{MBD(k)} \cdot w_k / \sum_{k=1}^{studies} w_k.$$

Note that the meta-analytical weighting strategy is different from the weighting strategy for obtaining a single effect size per study, although both take variability into account—of data around the fitted baseline trend (at the within-study level) and of outcomes around the average per effect per study (at the across-studies level). For designs in which only one AB comparison is performed (designs that are methodologically weaker; Tate et al., 2013), CV is set to 1 in order to use only the number of measurements available as a weight. It is also possible to stick to the more classic number of measurements available in all data sets from the study as a weight for the study effect size, as we did for the standardised version of the MPD and SLC procedures applied here.

**Example: Meta-analysis of neurobehavioural interventions to decrease problematic behaviours in adults with an acquired brain injury (ABI)**

A search of the literature was performed to identify articles published in English in peer-reviewed journals in which psychological interventions based on a neurobehavioural approach were applied to decrease problem behaviour in adults diagnosed with an ABI. A behavioural management approach (based on operant learning theory) provided the main framework for treatment intervention in SCEDs considered in the current meta-analysis. Indeed, this approach has been largely used in the literature and encompasses various strategies (e.g., time out on the spot, various forms of differential reinforcement, token economy) to decrease severe problematic behaviours such as aggression, inappropriate sexual behaviours, perseverative and inappropriate comments, delusional outbursts, disorders of self-awareness, and so forth, in patients with severe cognitive impairments. We opted to gather a substantively meaningful set of studies in order including only studies with studies unambiguous or easy-to-handle results (Fisher & Lerman, 2014).

The PsycINFO, MEDLINE, and ScienceDirect databases were searched by using several keyword combinations such as brain injury, operant learning theory, token economy, neurobehavioural intervention, behaviour therapy, behaviour modification, differential reinforcement, and single-case. In addition, ancestral searches of recent literature reviews of psychological interventions, including neurobehavioural interventions in brain injury rehabilitation, were also conducted (i.e., Alderman & Wood, 2013; Cattelani, Zettin, & Zoccolotti, 2010; Heinicke & Carr, 2014; Wood & Alderman, 2011; Ylvisaker et al., 2007). The flow of information can be seen in Figure 4, which provides the reasons for excluding records at

the different stages; the information about the 28 studies included in the meta-analysis is available in a table in the supplemental online material. Although the number of studies included may seem rather limited, it should be noted that a review of single-case meta-analyses has shown that 60% of them included fewer than 30 studies (Moeyaert et al., 2013). *The data were extracted using Plot Digitizer 2.6.3 (<http://plotdigitizer.sourceforge.net>) - preliminary results suggest that there not much difference across data retrieval programs in terms of reliability (M. Moeyaert, personal communication, March 27, 2015).*

INSERT FIGURE 4 ABOUT HERE

**Results: Extensions of the unstandardised indices.** The results of the integration between studies via the proposals for MPD and SLC are summarised in Table 1, and the graphical representations for the percentage and standardised versions of the MPD are provided in Figures 5 and 6 via modified forest plots, in which the range of outcomes within a study are given instead of the (unavailable) confidence intervals.

INSERT TABLE 1 ABOUT HERE

INSERT FIGURES 5 AND 6 ABOUT HERE

In sum, all of the information suggests that, in general, the (neuro)behavioural interventions have been effective, and, especially for studies with clear reductions of problematic behaviour, all outcomes observed within the study indicate decrease. However, some of the study results had to be excluded because of excessively outlying values.

**Results: Comparison with the *d*-statistic.** In order to offer applied researchers more information about the characteristics of the proposals made in the current article, we compared

the results obtained to those provided by the  $d$ -statistic (Hedges et al., 2012, 2013). Regarding its application, three aspects need to be mentioned: (a) Only those studies with more than one participant were included, in accordance with the way in which this indicator is computed, which led to the exclusion of 11 studies, as seen in Figure 4; (b) if both an increase in appropriate behaviour and a decrease in inappropriate behaviour was observed, the data set when an increase of the behaviour of interest was the effect desired was removed from the calculation of the  $d$ -statistic, as these cases were much less frequent; and (c) we further removed the results for Dixon et al. (2004) and Hegel and Ferguson (2000) from the summary shown in Figure 7 because of excessively high variances (23 and 251, respectively)—this did not, however, change the weighted average, because of the low weights assigned to these outcomes.

The forest plot for the  $d$ -statistic indicates that the intervention is effective for all studies except one. The difference between these results and those obtained after controlling for the baseline trend (i.e., the standardised versions of the MPD and the SLC) may reflect a potential overestimation of intervention effectiveness by the  $d$ -statistic. However, in some cases (e.g., for very long series or for measurements that can only range from 0 to 100%), the trend estimated and projected by MPD and SLC may not always be realistic.

INSERT FIGURE 7 ABOUT HERE

### **User-Friendly Software**

One of the main difficulties that practitioners and applied researchers face when analysing single-case design data may be the lack of software. We decided to implement the current developments in R, as most analytical procedures are available in a variety of R packages (see

Manolov et al., 2014). We have developed code for performing the within-study calculations for obtaining a single effect size per study, as well as for performing meta-analyses, as described here, apart from providing graphical representations such as those shown in the current paper. The software is explained in a step-by-step fashion in the 45-page supplemental material available at the web page of the journal and also at the following URL:  
[https://www.dropbox.com/s/z9x7nuy4vcmk7r4/Supplemental%20material\\_Manual.pdf?dl=0](https://www.dropbox.com/s/z9x7nuy4vcmk7r4/Supplemental%20material_Manual.pdf?dl=0).

## **Discussion**

### **Summary of the evidence on neurobehavioural interventions in adults with an ABI**

Although it has only been done in numerical terms, the summary of applied research in the current study provides support for the effectiveness of the psychological interventions derived from the operant learning theory. Indeed, intervention strategies such as token economy (e.g., response cost), various forms of differential reinforcement, or time out on the spot enable the reduction of a wide range of problem behaviours in persons with brain injury (even in the presence of severe cognitive impairments), such as aggression, inappropriate sexual behaviours, or perseverative and inappropriate comments, which represent a challenge to social and/or vocational reintegration.

Here we address, chronologically, the results of previous review papers on the topic in order to offer more information to applied researchers interested in the topic. Ylvisaker et al.'s (2007) review of behavioural interventions for children and adults with traumatic (rather than, more generally, acquired) brain injury reported that all 65 studies that were included showed some positive effects of the intervention. Cattalani et al. (2010) focused on a variety of designs and

studies on more than 1,000 adults, reporting greater effectiveness of comprehensive-holistic rehabilitation programs as compared with cognitive behaviour therapy. They also replicated the positive results reported by Ylvisaker et al. (2007) on approaches based on applied behaviour analysis. Wood and Alderman (2011) offer a narrative review of studies on traumatic brain injury that reported positive results of differential reinforcement (for low rates of responding, for other or incompatible behaviours), both in neurobehavioural units and in non-specialized settings, as well as the effectiveness of response-cost (negative punishment) for people who present cognitive impairment and challenging behaviours. In contrast to these three reviews, Heinicke and Carr (2014) performed a meta-analysis of 112 studies (on various aetiologies, including ABI), reporting higher standardised mean differences for skill acquisition (ranging between 14 and 20) than for behaviour reduction—the main object of the current paper—ranging between 4 and 6 *SDs*. The values obtained in our review (Table 2) are smaller, but the *d*-statistic controls for small sample bias and autocorrelation, and MPD and SLC control for baseline trend, whereas the standardised mean difference used by Heinicke and Carr (2014) does neither.

We reiterate Wood and Alderman's (2011) emphasis on considering environmental contingencies and any disorders of drive and motivation in the patients when planning an intervention. Furthermore, regarding SCED studies, internal validity is a strength, but even the accumulation of positive meta-analytical results should be interpreted with caution because of potential selection bias of the participants in each study (Ylvisaker et al., 2007). Accordingly, Alderman and Wood (2013) advise against over-reliance on specific interventions, as the same problematic behaviours may have different underlying causes. In contrast, Cattalani and colleagues (2010) stress that the limitations of group-design studies in terms of including participants with very different demographic characteristics, aetiology, site of brain damage, and

so forth, hinder the assessment of how large the effect of a treatment would be for a particular patient. Some of these issues of uncontrolled factors affecting the certainty of the causal effect of interventions can be addressed in SCEDs, especially if recommendations on study conduct (Horner et al., 2005; Kratochwill et al., 2010; Tate et al., 2013) are followed, including measures of maintenance and generalization (frequently missing according to Ylvisaker et al., 2007), social validity, and procedural fidelity (also rare according to Heinicke & Carr, 2014), among others. Thus, the combination of several methodological options and the integration of results across studies can prove to be very useful in the field.

### **Recommendations to researchers**

Although the results obtained here are necessarily restricted to the studies reviewed, several pieces of evidence need to be highlighted. First, the fact that changes in slope and in level were found suggests that both aspects need to be taken into account, as occurs in the SLC procedure (but not the *d*-statistic). Second, the MPD and SLC procedures indicate that for some studies, the improvement observed could be expected from the evolution of the behaviour prior to the intervention; thus, not considering baseline trends may lead to overestimating treatment effectiveness. Third, the MPD and SLC procedures allow the integration of studies that include one or more participants, whereas the *d*-statistic necessarily requires several subjects. Fourth, the percentage version of the MPD and SLC procedures led, in some cases, to extreme results whenever the baseline levels were very low and the increases due to the intervention were proportionately enormous (e.g., for the percentage version of the MPD we obtained once a result of 920% and once 750% in a total of 78 quantifications). Such results can affect the summary

measures obtained via meta-analysis **although large variability of the results within a study implies less weight in the meta-analytical summary (see Figure 5), as was the case for the Travis and Sturmev (2010) data.** In conclusion, a conservative recommendation would be to use the standardised version of these two procedures for quantitative integrations (i.e., at the across-studies level) if the **baseline measurement are very small and** extreme percentages are obtained, and to use the percentage versions as an additional indication of intervention effectiveness at the within-study level. Finally, if a meta-analysis includes a vast majority of studies with more than one participant and the visual inspection of the data suggests that baseline trends are rare, the *d*-statistic can be used, as it is based on solid statistical theory and enables the pooling together of single-case and group-design research.

### **Limitations and future research**

A limitation specific to the MPD and the SLC is that they account only for linear trends and their sampling distributions have not been derived, which makes it impossible to use the inverse of the index variance as a weight. In addition, we did not search for grey literature databases to deal with publication bias, which is why we also did not assess this issue with the funnel plot or the trim-and-fill method (Duval & Tweedie, 2000).

A question that is still open for discussion is whether simpler unstandardised indices are appropriate alternatives to more complex procedures such as the *d*-statistic and multilevel models. In order to help answer this question, simulation studies are called for, for instance, assessing the overestimation of effects when not controlling for baseline trend (*d*-statistic) and underestimating in case of projections of the baseline (MPD and SLC) that are outside the limits



of what is possible for the type of data gathered. The appropriate weights in SCED data analysis are still to be determined. Beyond discussions between statisticians, it will be especially important to present the analytical alternatives to applied researchers via presentations at professional conferences and through special issues of journals (e.g., *Journal of School Psychology*, Vol. 52, Issue 2, and *Neuropsychological Rehabilitation*, Vol. 24, Issues 3-4, both in 2014) to understand their perceptions on the usefulness and feasibility of these techniques and their willingness to use them in their everyday practice.

## References

- \* Articles included in the meta-analysis are indicated with an asterisk (\*).
- \*Alderman, N. (1991). The treatment of avoidance behaviour following severe brain injury by satiation through negative practice. *Brain Injury*, *5*, 77–86.
- \*Alderman, N. (2003). Contemporary approaches to the management of irritability and aggression following traumatic brain injury. *Neuropsychological Rehabilitation*, *13*, 211–240.
- \*Alderman, N., & Burgess, P. (1994). A comparison of treatment methods for behaviour disorder following herpes simplex encephalitis. *Neuropsychological Rehabilitation*, *4*, 31–48.
- \*Alderman, N., Fry, R. K., & Youngson, H. A. (1955). Improvement of self-monitoring skills, reduction of behaviour disturbance and the dysexecutive syndrome: Comparison of response cost and a new programme of self-monitoring training. *Neuropsychological Rehabilitation*, *5*, 193–221.
- \*Alderman, N., & Knight, C. (1997). The effectiveness of DRL in the management and treatment of severe behavior disorders following brain injury. *Brain Injury*, *11*, 79–101.
- \*Alderman, N., & Ward, A. (1991). Behavioural treatment of the dysexecutive syndrome: Reduction of repetitive speech using response cost and cognitive overlearning. *Neuropsychological Rehabilitation*, *1*, 65–80.
- Alderman, N., & Wood, R. L. (2013). Neurobehavioural approaches to the rehabilitation of challenging behaviour. *NeuroRehabilitation*, *32*, 761–70.

- Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research and Therapy*, *31*, 621-631.
- \*Andrewes, D. (1989). Management of disruptive behaviour in the brain-damaged patient using selective reinforcement. *Journal of Behavior Therapy and Experimental Psychiatry*, *20*, 261–264.
- Baron, A., & Derenne, A. (2000). Quantitative summaries of single-subject studies: What do group comparisons tell us about individual performances? *The Behavior Analyst*, *23*, 101–106.
- Beeson, P. M., & Robey, R. R. (2006). Evaluating single-subject treatment research: Lessons learned from the aphasia literature. *Neuropsychological Review*, *16*, 161–169.
- 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.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley & Sons.
- \*Burke, W. H., & Lewis, F. D. (1986). Management of maladaptive social behavior of a brain injured adult. *International Journal of Rehabilitation Research*, *9*, 335–342.
- Cattelani, R., Zettin, M., & Zocolotti, P. (2010). Rehabilitation treatments for adults with behavioral and psychosocial disorders following acquired brain injury: A systematic review. *Neuropsychology Review*, *20*, 52–85.

- Cheung, S. F., & Chan, D. K.-S. (2004). Dependent effect sizes in meta-analysis: Incorporating the degree of interdependence. *Journal of Applied Psychology, 89*, 780–791.
- \*Crane, A. A., & Joyce, B. G. (1991). Cool down: A procedure for decreasing aggression in adults with traumatic head injury. *Behavioral Residential Treatment, 6*, 65–75.
- Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. London, UK: Routledge.
- \*Davis, J. R., Turner, W., Rolider, A., & Cartwright, T. (1994). Natural and structured baselines in the treatment of aggression following brain injury. *Brain Injury, 8*, 589–597.
- \*Dayus, B., & van den Broek, M. D. (2000) Treatment of stable delusional confabulations using self-monitoring training. *Neuropsychological Rehabilitation, 10*, 415–427.
- \*Dixon, M.R., Guercio, J., Falcomata, T., Horner, M.J., Root, S., Newell, C., & Zlomke, K. (2004). Exploring the utility of functional analysis methodology to assess and treat problematic verbal behavior in persons with acquired brain injury. *Behavioral Interventions, 19*, 91–102.
- Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 455–463.
- Ferron, J. M., Moeyaert, M., Van den Noortgate, W., & Beretvas, S. N. (2014). Estimating causal effects from multiple-baseline studies: Implications for design and analysis. *Psychological Methods*. Advance online publication. [doi:10.1037/a0037038](https://doi.org/10.1037/a0037038)
- Fisher, W. W., & Lerman, D. C. (2014). It has been said that, “There are three degrees of falsehoods: Lies, damn lies, and statistics”. *Journal of School Psychology, 52*, 243–248.

- Gage, N. A., & Lewis, T. J. (2014). Hierarchical linear modeling meta-analysis of single-subject design research. *Journal of Special Education, 48*, 3–16.
- Ganz, J. B., Earles-Vollrath, T. L., Heath, A. K., Parker, R. I., Rispoli, M. J., & Duran, J. B. (2012). A meta-analysis of single case research studies on aided augmentative and alternative communication systems with individuals with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 42*, 60–74.
- \*Giles G. M., Fussey, I., & Burgess P. (1988). The behavioural treatment of verbal interaction skills following severe head injury: A single case study. *Brain Injury, 2*, 75–79.
- Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.
- \*Gouvier, W.D., Richards, J. S., Blanton, P. D., Janert, K., Rosen, L. A., & Drabman R. S. (1985). Behavior modification in physical therapy. *Archives of Physical Medicine and Rehabilitation, 66*, 113–116.
- \*Hanlon, R. E., Clontz, B., & Thomas, M. (1993). Management of severe behavioral dyscontrol following subarachnoid hemorrhage. *Neuropsychological Rehabilitation, 3*, 63–76.
- Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods, 3*, 224–239.
- Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2013). A standardized mean difference effect size for multiple baseline designs across individuals. *Research Synthesis Methods, 4*, 324–341.

- \*Hegel M. T. (1988). Application of a token economy with a non-compliant closed head-injured male. *Brain Injury*, 2, 333–338.
- \*Hegel, M. T., & Ferguson, R. J. (2000). Differential reinforcement of other behavior (DRO) to reduce aggressive behavior following traumatic brain injury. *Behavior Modification*, 24, 94–101.
- Heinicke, M. R., & Carr, J. E. (2014). Applied behavior analysis in acquired brain injury rehabilitation: A meta-analysis of single-case design intervention research. *Behavioral Interventions*, 29, 77–105.
- Herzinger, C. V., & Campbell, J. M. (2007). Comparing functional assessment methodologies: A quantitative synthesis. *Journal of Autism and Developmental Disorders*, 37, 1430–1445.
- Heyvaert, M., & Onghena, P. (2014). Analysis of single-case data: Randomisation tests for measures of effect size. *Neuropsychological Rehabilitation*, 24, 507–527.
- 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.
- Jamieson, M., Cullen, B., McGee-Lennon, M., Brewster, S., & Evans, J. J. (2014). The efficacy of cognitive prosthetic technology for people with memory impairments: A systematic review and meta-analysis. *Neuropsychological Rehabilitation*, 24, 419–444.
- Jenson, W. R., Clark, E., Kircher, J. C., & Kristjansson, S. D. (2007). Statistical reform: Evidence-based practice, meta-analyses, and single subject designs. *Psychology in the Schools*, 44, 483–493.

- \*Knight, C., Rutterford, N. A., Alderman, N., & Swan, L. J. (2002). Is accurate self-monitoring necessary for people with acquired neurological problems to benefit from the use of differential reinforcement methods? *Brain Injury, 16*, 75–87.
- Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). Single case designs technical documentation. In *What Works Clearinghouse: Procedures and standards handbook (Version 2.0)*. Retrieved from [http://ies.ed.gov/ncee/wwc/pdf/wwc\\_scd.pdf](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.
- 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.
- Levin, J. R., Ferron, J. M., & Kratochwill, T. R. (2012). Nonparametric statistical tests for single-case systematic and randomized ABAB...AB and alternating treatment intervention designs: New developments, new directions. *Journal of School Psychology, 50*, 599–624.
- Littell, J. H., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis*. New York, NY: Oxford University Press.
- Maggin, D. M., Swaminathan, H., Rogers, H. J., O'Keefe, B. V., Sugai, G., & Horner, R. H. (2011). A generalized least squares regression approach for computing effect sizes in single-case research: Application examples. *Journal of School Psychology, 49*, 301–321.

- \*Manchester, D., Hodgkinson, A., & Casey, T. (1997). Prolonged, severe behavioral disturbance following traumatic brain injury: What can be done? *Brain Injury*, *8*, 605–617.
- Manolov, R., Gast, D. L., Perdices, M., & Evans, J. J. (2014). Single-case experimental designs: Reflections on conduct and analysis. *Neuropsychological Rehabilitation*, *24*, 634–660.
- Manolov, R., & Solanas, A. (2013). A comparison of mean phase difference and generalized least squares for analyzing single-case data. *Journal of School Psychology*, *51*, 201–215.
- Manolov, R., Solanas, A., Sierra, V., & Evans, J. J. (2011). Choosing among techniques for quantifying single-case intervention effectiveness. *Behavior Therapy*, *42*, 533–545.
- Miller, M. J. (1985). Analyzing client change graphically. *Journal of Counseling and Development*, *63*, 491–494.
- Moeyaert, M., Ferron, J., Beretvas, S., & Van Den Noortgate, W. (2014). From a single-level analysis to a multilevel analysis of single-case experimental designs. *Journal of School Psychology*, *52*, 191–211.
- Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Van Den Noortgate, W. (2013). The three-level synthesis of standardized single-subject experimental data: A Monte Carlo simulation study. *Multivariate Behavioral Research*, *48*, 719–748.
- Olive, M. L., & Smith, B. W. (2005). Effect size calculations and single subject designs. *Educational Psychology*, *25*, 313–324.
- Parker, R. I., & Vannest, K. J. (2012). Bottom-up analysis of single-case research designs. *Journal of Behavioral Education*, *21*, 254–265.



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

Perdices, M., & Tate, R. L. (2010). Single-subject designs as a tool for evidence-based clinical practice: Are they unrecognized and undervalued? *Neuropsychological Rehabilitation, 19*, 904–927.

\*Persel, C. S., Persel, C. H., Ashley, M. J., & Krych, D. K. (1997). The use of noncontingent reinforcement and contingent restraint to reduce physical aggression and self-injurious behaviour in a traumatically brain-injured adult. *Brain Injury, 11*, 751–760.

\*Peters, M. D., Gluck, M., & McCormick, M. (1992). Behaviour rehabilitation of the challenging client in less restrictive settings. *Brain Injury, 6*, 299–314.

Pustejovsky, J. E. (2014). Measurement-comparable effect sizes for single-case studies of free-operant behavior. *Psychological Methods*. Advance online publication.  
doi:10.1037/met0000019

Ross, S. G., & Begeny, J. C. (2014). Single-case effect size calculation: Comparing regression and non-parametric approaches across previously published reading intervention data sets. *Journal of School Psychology, 52*, 419–431.

\*Rothwell, N. A., LaVigna, G. W., & Willis, T. J. (1999). A non-aversive rehabilitation approach for people with severe behavioural problems resulting from brain injury. *Brain Injury, 13*, 521–533.

Scruggs, T. E., & Mastropieri, M. A. (1998). Summarizing single-subject research: Issues and applications. *Behavior Modification, 22*, 221–242.

- 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.
- Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention, 2*, 188–196.
- Solanas, A., Manolov, R., & Onghena, P. (2010). Estimating slope and level change in N=1 designs. *Behavior Modification, 34*, 195–218.
- Strain, P. S., Kohler, F. W., & Gresham, F. (1998). Problems in logic and interpretation with quantitative syntheses of single-case research: Mathur and colleagues (1998) as a case in point. *Behavioral Disorders, 24*, 74–85.
- Sullivan, K. J., Shadish, W. R., & Steiner, P. M. (2014). An introduction to modeling longitudinal data with generalized additive models: Applications to single-case designs. *Psychological Methods*. Advance online publication. doi:10.1037/met0000020
- Swaminathan, H., Rogers, H. J., & Horner, R. H. (2014). An effect size measure and Bayesian analysis of single-case designs. *Journal of School Psychology, 52*, 213–230.
- Tate, R. L., Perdices, M., Rosenkoetter, U., Wakima, 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.

- \*Ter Mors, B. J., van Heugten, C. M., & van Harten, P. N. (2012). Evaluation of electrical aversion therapy for inappropriate sexual behaviour after traumatic brain injury: A single case experimental design study. *BMJ Case Reports*. doi:10.1136/bcr-02-2012-5932.
- \*Travis, R., & Sturmey, P. (2010). Functional analysis and treatment of the delusional statements of a man with multiple disabilities: A four-year follow-up. *Journal of Applied Behavior Analysis, 43*, 745–749.
- \*Turner, J. M., Green, G., & Braunling-McMorrow, D. (1990). Differential reinforcement of low rates of responding (DRL) to reduce dysfunctional social behaviors of a head injured man. *Behavioral Residential Treatment, 5*, 15–27.
- Ugille, M., Moeyaert, M., Beretvas, S. N., Ferron, J., & Van den Noortgate, W. (2012). Multilevel meta-analysis of single-subject experimental designs: A simulation study. *Behavior Research Methods, 44*, 1244–1254.
- Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods, 45*, 576–594.
- 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.
- \*Wesolowski, M. D., Zencius, A., & Burke, W. H. (1993). Effects of feedback and behavior contracting on head trauma person's inappropriate sexual behavior. *Behavior Residential Treatment, 8*, 89–96.

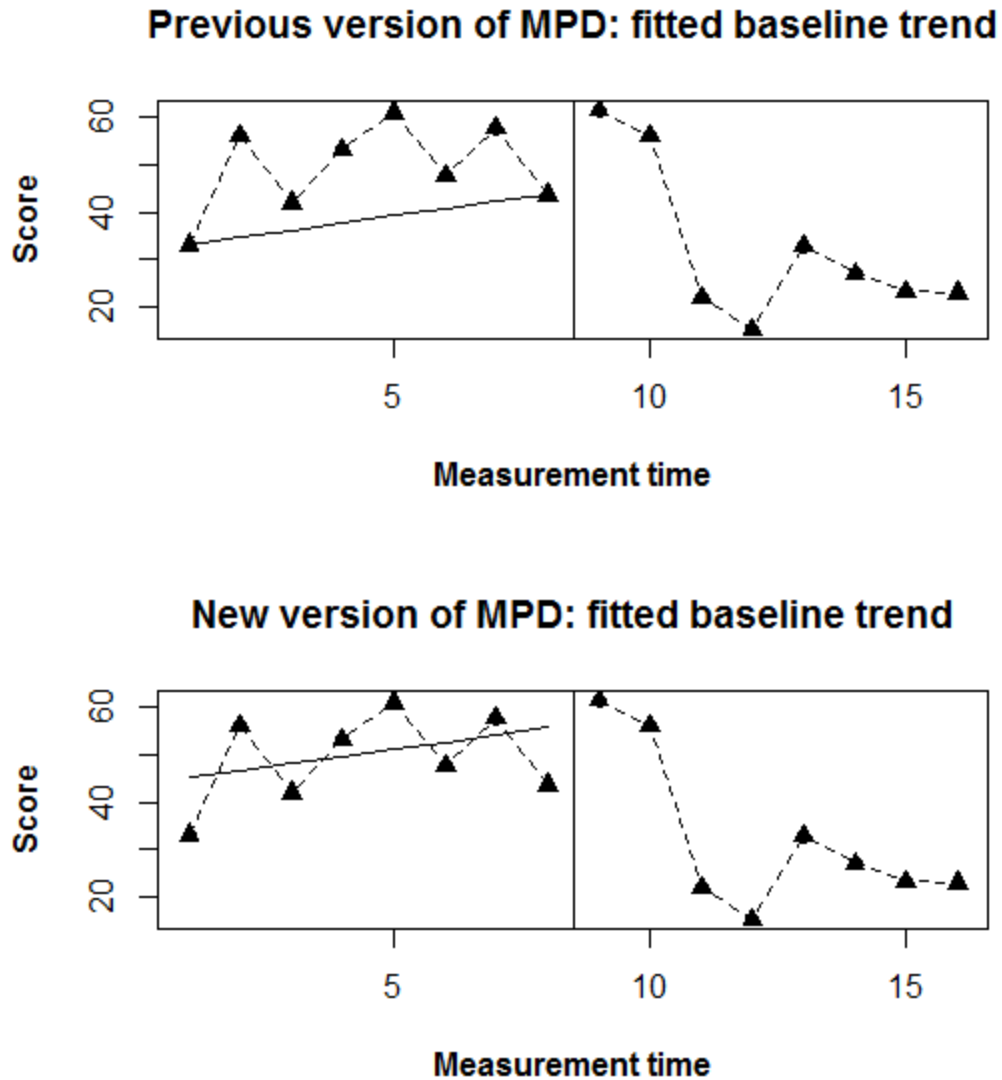
- Wood, R. L., & Alderman, N. (2011). Applications of operant learning theory to the management of challenging behavior after traumatic brain injury. *Journal of Head Trauma Rehabilitation, 26*, 202–211.
- Ylvisaker, M., Turkstra, L., Coehlo, C., Yorkston, K., Kennedy, M., Sohlberg, M. M., & Avery, J. (2007). Behavioral interventions for children and adults with behavior disorders after TBI: A systematic review of the evidence. *Brain Injury, 21*, 769–805.
- \*Yody, B. B., Schaub, C., Conway, J., Peters, S., Strauss, D., & Helsing S. (2000). Applied behavior management and acquired brain injury: Approaches and assessment. *The Journal of Head Trauma Rehabilitation, 15*, 1041–1060.
- \*Zencius, A. H., & Wesolowski, M. D. (1991). Reducing verbal aggression in adults with brain injury. *Behavioral Interventions, 6*, 155–164.

**Table 1.** Summary measures obtained in the quantitative integration of the studies included in the current review

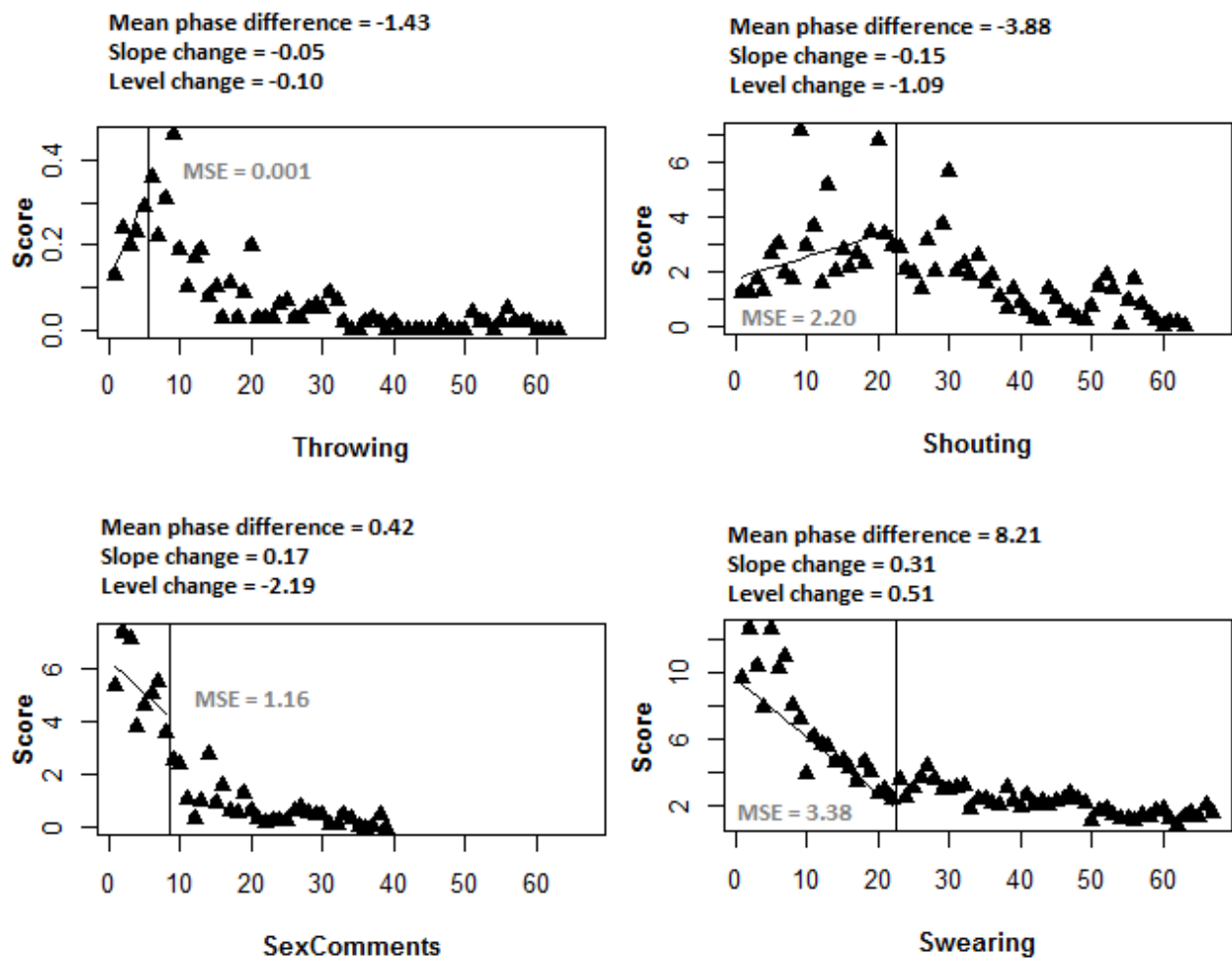
Index	Version	Weighted average	Range <sup>a</sup>
Mean phase difference	Standardised	-2.55	[-11.56, 5.53]
Slope change estimate	Standardised	-0.12	[-1.77, 0.44]
Level change estimate	Standardised	-1.36	[-6.57, 3.72]
<i>d</i> -statistic <sup>b</sup>	Standardised	-1.29	[-3.41, 0.00]
Mean phase difference	Percentage	-5.62	[-91.82, 206.22]
Slope change estimate	Percentage	-43.95	[-350.79, 137.40]
Level change estimate	Percentage	-120.84	[-445.31, 458.82]

<sup>a</sup>The range represents the minimal and maximal values obtained in the individual studies.

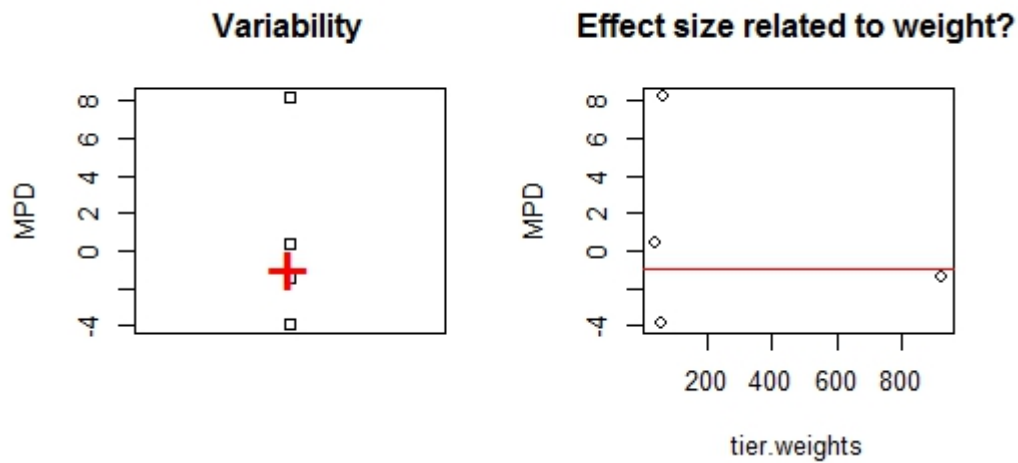
<sup>b</sup>The *d*-statistic was computed only in studies including more than one participant.



**Figure 1.** Illustration of how the baseline trend is fitted in the previous version of the Mean phase difference and in the modification proposed in the current paper, using data from Knight, Knight, Rutterford, Alderman and Swan (2002).

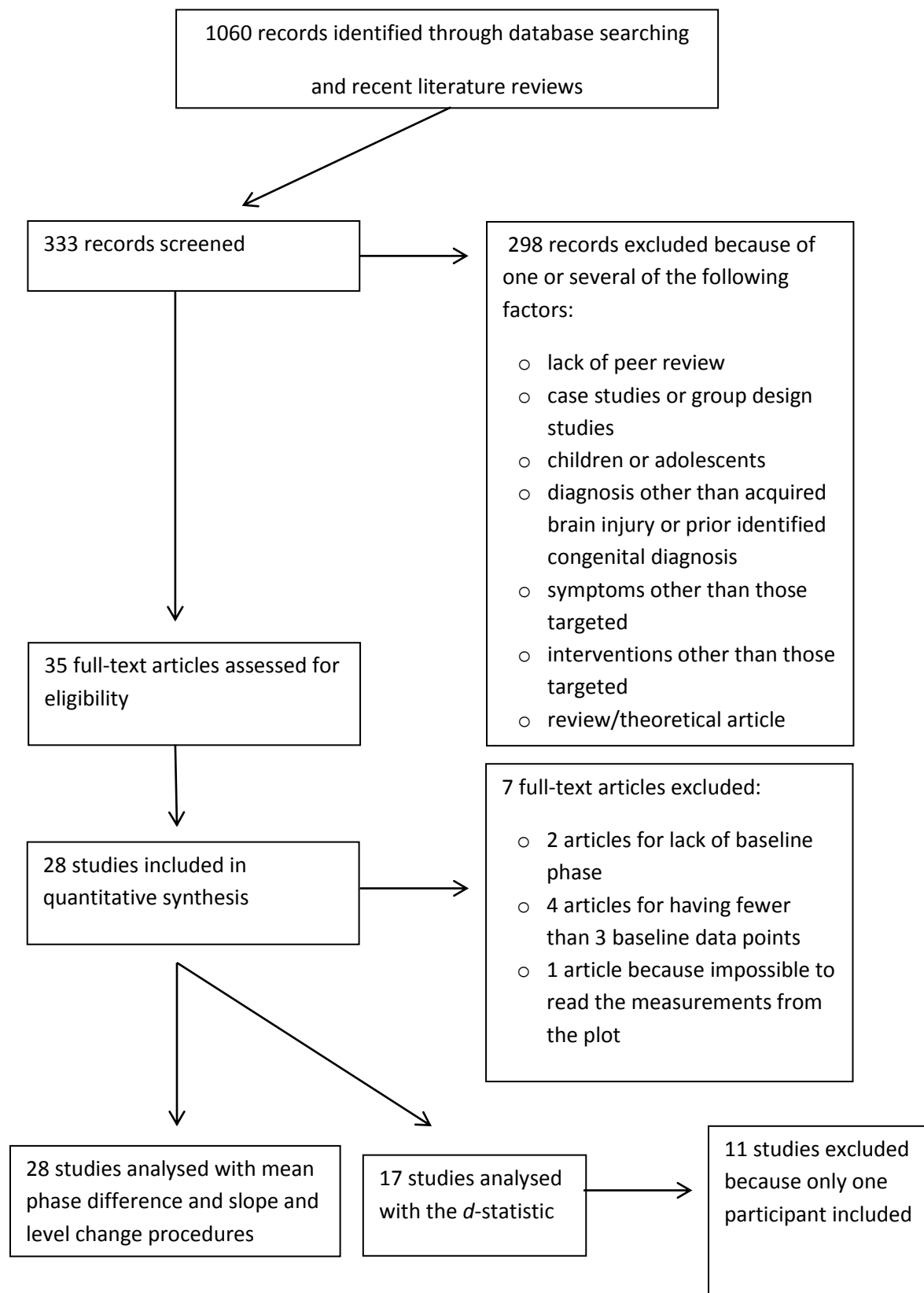


**Figure 2.** Data gathered by Alderman and Knight (1997), alongside the results of the application of the mean phase difference and the slope and level change procedure, as well as the mean square error around the fitted baseline trend.

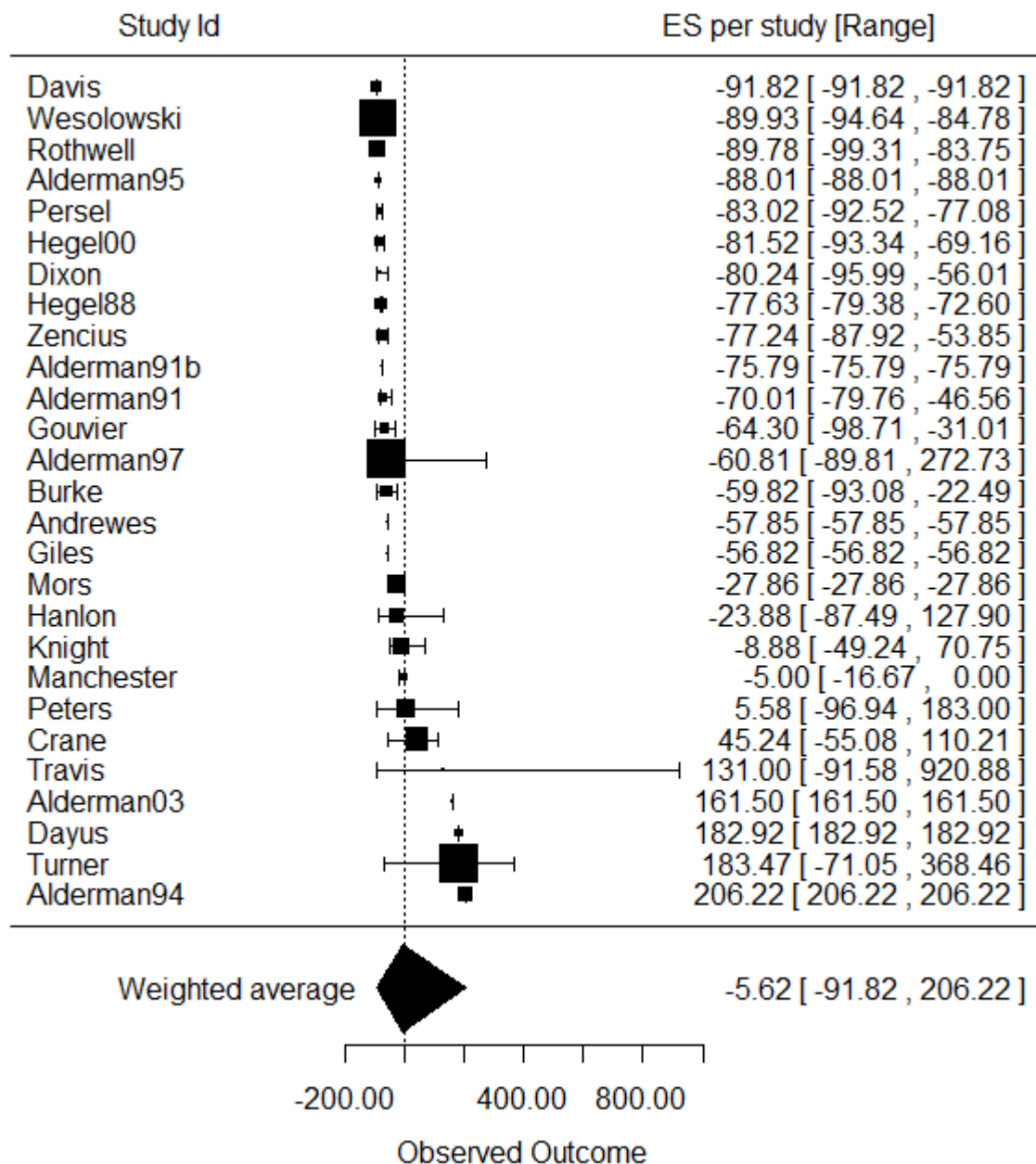


**Figure 3.** Graphical representation of the mean phase difference values obtained for each of the tiers in the multiple-baseline design (MPD) study by Alderman and Knight (1997), as well as a representation of these values in relation to the weights assigned to them. The plus sign (left panel) and the horizontal line (right panel) represent the weighted within-study average.

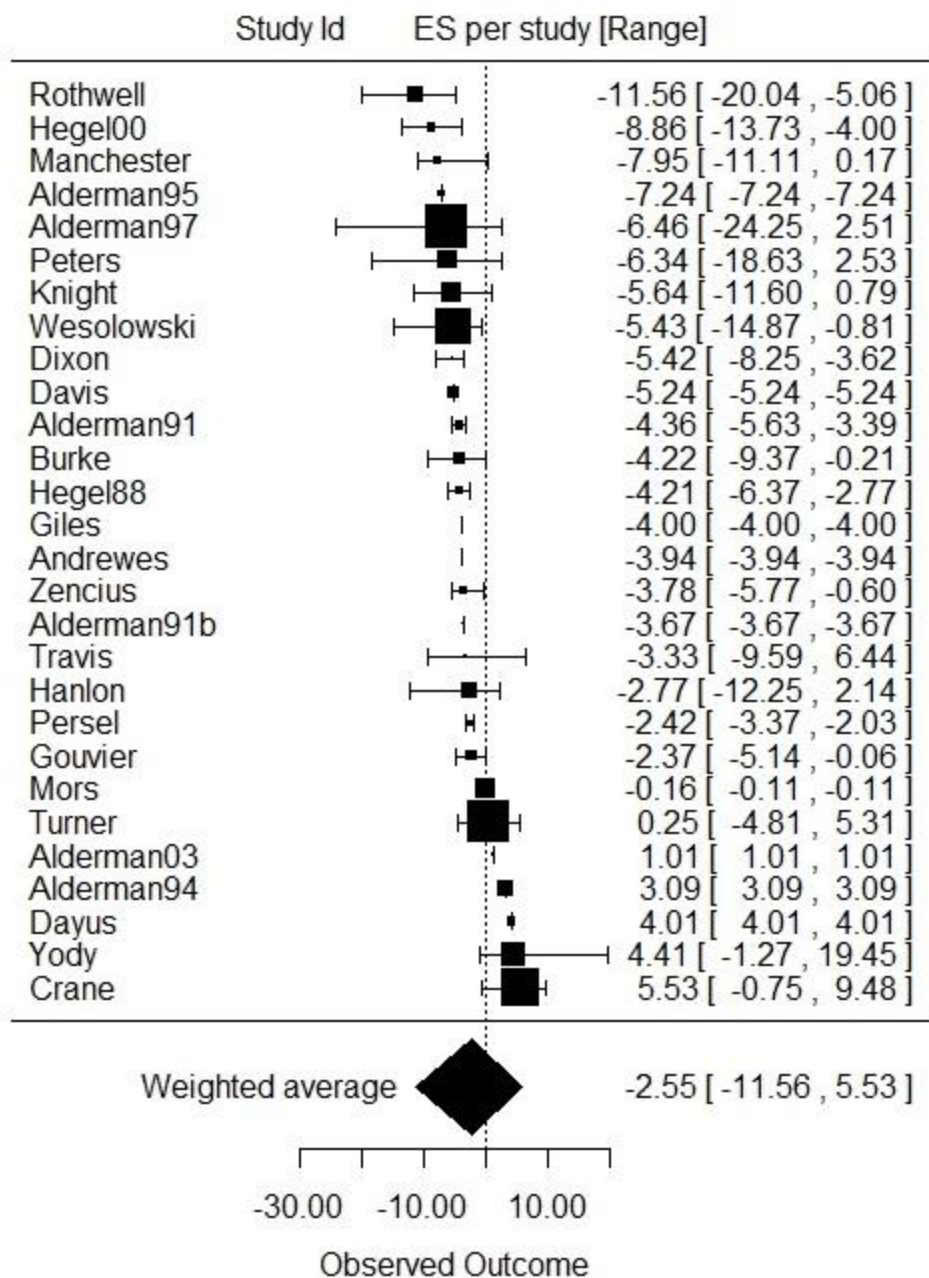




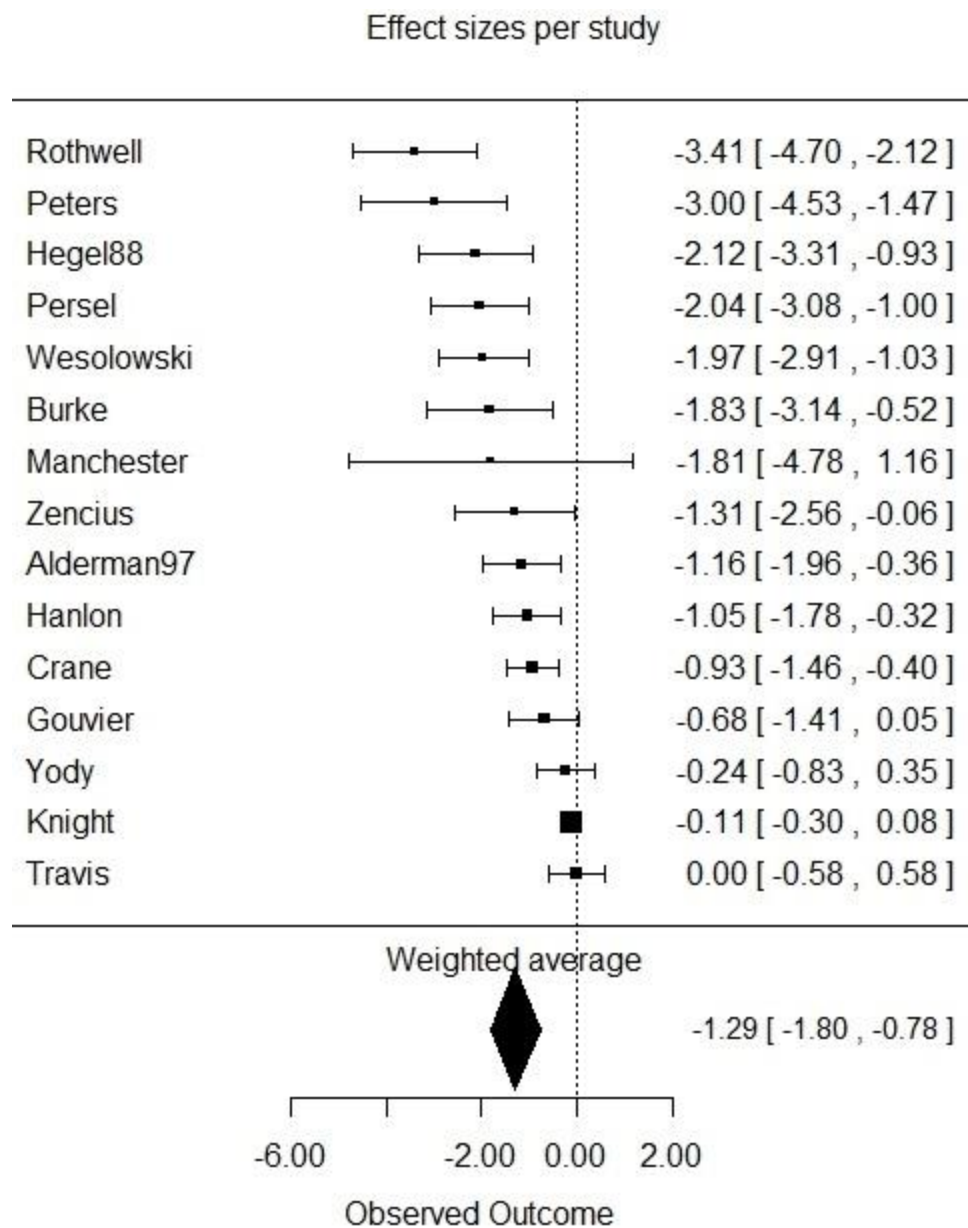
**Figure 4.** Flow of information through the different phases of the systematic review.



**Figure 5.** Graphical representation of the percentage version of the Mean phase difference as applied to the set of studies meta-analysed; the inverse of the residual variance around the baseline trend line and series length are used as a weighting strategy at the within-study level; the inverse of the within-study variability in outcomes and the number of measurements per study are used as a weighting strategy at the across-studies level. Id = identification; ES = effect size.



**Figure 6.** Graphical representation of the standardised version of the Mean phase difference as applied to the set of studies meta-analysed. Series length is used as a weighting strategy at the within-study and across-studies levels. ID = identification; ES = effect size.



**Figure 7.** Forest plot for the  $d$ -statistic as applied to the set of studies being meta-analytically integrated; the inverse of the index variance is used as a weight at the across-study level.