Probability and visual aids for assessing intervention effectiveness in single-case designs: A field test

Rumen Manolov[1 2], Matthew Jamieson[3 4], Jonathan J. Evans[3], & Vicenta Sierra[2]

[1] Department of Behavioural Sciences Methods, University of Barcelona, Spain

[2] ESADE Business School, Ramon Llull University, Spain

[3] Institute of Health and Wellbeing, University of Glasgow, Scotland, UK

[4] Glasgow Interactive Systems Group, School of Computing Science, University of Glasgow, Scotland, UK

Running head: Single-case data analysis: Probability and visual aids

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

Single-case data analysis still relies heavily on visual inspection and, at the same time, it is not clear to what extent the results of different quantitative procedures converge in identifying an intervention effect and its magnitude when applied to the same data; this is the type of evidence provided here for two procedures. One of the procedures, included due to the importance of providing objective criteria to visual analysts, is a visual aid fitting and projecting split-middle trend while taking into account data variability. The other procedure converts several different metrics into probabilities making their results comparable. In the present study we study to what extend these two procedures coincide in the magnitude of intervention effect taking place in a set of studies stemming from a recent meta-analysis. The procedures concur to a greater extent with the values of the indices computed and with each other and, to a lesser extent, with our own visual analysis. For the distinctions smaller and larger effects the probability-based approach seems somewhat better suited. Moreover, the results of the field test suggest that the latter is a reasonably good mechanism for translating different metrics into similar labels. User friendly R code is provided for promoting the use of the visual aid, together with a quantification based on nonoverlap and the label provided by the probability approach.

Key words: single-case designs, effect size, visual aids, split-middle, software

Single-case experimental designs (SCED) as a field of research has attracted a lot of attention in recent years, in relation to its strengths as a method for obtaining solid evidence (Howick et al., 2011) and due to the lack of consensus regarding an optimal analytical technique (Kratochwill et al., 2010; Smith, 2012; Tate et al., 2013). The former aspect has been promoted by methodological standards and quality scales (e.g., Horner et al., 2005; Reichow, Volkmar, & Cicchetti, 2008; Tate et al., 2013). The latter aspect has led to proposing several different indices based on data overlap (Parker, Vannest, & Davis, 2011), adapting standardized mean difference statistics (Shadish, Hedges, & Pustejovsky, 2014), advocating for the application of multilevel models for analysis and meta-analysis (Moeyaert, Ferron, Beretvas, & Van Den Noortgate, 2014), making clear that randomization tests can provide information about both effect size and statistical significance (Heyvaert & Onghena, 2014) and also proposing quantifications expressed in the same measurement units as the behaviour of interest being measured (Manolov & Solanas, 2013a; Solanas, Manolov, & Onghena, 2010). These alternatives all involve quantifying the difference between baseline and intervention phases, but visual analysis is still used for deciding whether an intervention has been effective or not and is still considered essential (Kratochwill et al., 2010; Parker, Cryer, & Byrns, 2006).

Therefore, one of the decisions that applied researchers have to make is how to complement visual inspection with some objective criterion and quantification that could help them to: (a) improve agreement among analysts; (b) communicate the results to other researchers and practitioners; (c) make their study eligible for meta-analysis (Jenson, Clark, Kircher, & Kristjansson, 2007); (d) assess whether a behavioural change has taken place; and (e) assess the magnitude of behavioural change. Regarding the latter two points, applied researchers have to deal with the fact that different

quantifications may lead to two different conclusions and that the benchmarks used in between-groups designs for assigning different labels (e.g., small, medium, large) may not be appropriate for SCED (Parker et al., 2005).

In the present paper we focus on two techniques that can be used to inform about the presence and the degree of effectiveness. In the first part of what follows, we present these techniques, always keeping in mind that the data analytical tools are but one piece of information when assessing intervention effectiveness and that the professional's knowledge of the client, behaviour of interest, and context, is crucial, as well the assessment of maintenance and generalizability of the intervention effect. In the second part of the paper, we describe a field test performed with these two alternatives as applied to a set of studies included in a recent meta-analysis (Jamieson, Cullen, McGee-Lennon, Brewster, & Evans, 2014). An application with real behavioural data was chosen, because the two methods have already been tested with simulated data in the papers presenting them and due to the fact that the comparison of procedures with real data (with unknown underlying parameters) has been scarce.

Some Alternatives for Assessing the Degree of Intervention Effectiveness

The importance of the intervention effect should be assessed considering substantive (clinical, educational, social) criteria. When evaluating the magnitude of effect, it is also possible to take into account other pieces of information. One option is to take into account the values of an effect size index that have been obtained for a set of studies included in a field (e.g., Parker and Vannest, 2009, offer a field test of the Nonoverlap of all pairs; NAP). This way of proceeding would imply an across-studies comparison, i.e., it enables stating whether the effect observed is among the larger or smaller ones as compared to previously published research. In the following, we deal with two

alternatives that complement the across-studies assessment of the magnitude of effect with a within-study assessment.

A probability-based approach

The first alternative tested here is called the Maximal reference approach (MRA) and it deals with assigning probabilities according to the likelihood of the results in case there was not intervention effect (Manolov & Solanas, 2012). MRA was proposed for (a) assessing the magnitude of effect according to whether the difference in conditions is likely to take place only by chance and (b) making the results expressed in different metrics (e.g., overlap, R-squared) comparable, for instance, in order to make possible integrating quantitatively such results. Monte Carlo sampling is used to estimate the probability of the actual outcome being obtained only by chance (i.e., in a data set with the same phase lengths, with no intervention effect, and considering that data might be autocorrelated and present different kinds of variability or error distributions). MRA is thus related to simulation modelling analysis (SMA; Borckardt et al., 2008), but it is based on constructing several sampling distributions (instead of one) for representing a variety of data conditions. Although SMA was proposed to be based on the point biserial correlation as measure and its associated probability, both SMA and MRA can be applied using a variety of indices (and require the use of at least one index) for quantifying the amount of difference between a baseline and an intervention phase.

Just as SMA, MRA deals with one of the issues that were recently stressed again in a review of SCED studies (Solomon, 2014): autocorrelation. We decided not to estimate autocorrelation (ρ_1), as is done in SMA, due to low precision of the estimators for short series (Huitema & McKean, 1991; Matyas & Greenwood, 1991; Solanas, Manolov, & Sierra, 2010) which implies that there is a great deal of uncertainty around the true value

of autocorrelation. We rather chose to use the bias-corrected averages from the Shadish & Sullivan (2011) review: for multiple-baseline designs ρ_1=.321, for reversal designs such as ABAB, or $(AB)^k$ in general, ρ_1=.191, and for AB (or ABCD) designs ρ_1= .752. This approach, based on using the mean of a set of studies, is similar to the considerably more complex solution proposed by Shadish, Rindskopf, Hedges, and Sullivan (2013), which, on the basis of Bayesian estimation, leads to shrinking individual autocorrelation estimates closer to the mean obtained in a set of studies. Regarding the assessment of degree of effectiveness, the probabilities obtained via MRA can be converted into labels, according to the likelihood of the outcome (e.g., if its likelihood is lower than .10 it could be referred to as a difference very unlikely to be due to chance). Actually, we suggest that the MRA indications are converted into different ordered categories of likelihood rather than into specific probability values in order to avoid giving an impression of high precision when the true process underlying the data is unknown and only approximated by the three models for the error term and the average autocorrelation found in previous studies.

A visual aid approach

The approach described in this section is closely related to visual analysis, visual aids, and exploratory data analysis (Tukey, 1977). The visual aid is related to the necessity of dealing with baseline trend, which is another issue relevant for SCED data (Parker et al., 2006). Moreover, it was recently shown that trend may be present in SCED studies (Solomon, 2014). Given the focus on trend, we do not discuss proposals related to standard deviation bands (i.e., statistical process control; Fisher, Kelley, & Lomas, 2003; Pfadt & Wheeler, 1995) which are more appropriate for stable data. The option tested here is actually better-aligned with recent proposals intended to make

dealing with trend a feasible endeavour for applied researchers with no statistical expertise (Parker, Vannest, & Davis, 2014).

The approach (Manolov, Sierra, Solanas, & Botella, 2014) is based on (1) fitting a split-middle trend line to the baseline (Miller, 1995), (2) projecting it into the intervention phase and (3) constructing a trend envelope according to the idea presented in Gast and Spriggs (2010), but taking into account the amount of data variability present. Note that trend estimation and projection should always be done with caution if the baseline phase in which the trend is estimated is excessively short and if the projection goes far away in time and/or leads to out-of-range values. (Parker, Vannest, Davis, & Sauber, 2011, comment on these aspects in relation to regression-based proposals, but they also need to be kept in mind here.) This envelope is defined by the variability present in the baseline data[1], quantified by means of the interquartile range (IQR), a measure of scatter excluding the 25% lowest and 25% highest measurements. When the limits of the envelope are obtained adding and subtracting 1.5 times the IQR from the projected split-middle trend line, this allows for detecting a moderate intervention effect in case there are values falling outside these limits (atypical values when using the boxplot rule, Tukey, 1977). When the limits are obtained using 3 times the IQR, this makes possible detecting a large intervention effect (extreme outliers when using the boxplot rule, Tukey, 1977). We refer subsequently to this approach as SMIQR.

Differences between the approaches

Given that the first alternative presented here is based on probabilities and intensive computation, it is possible that not all applied researcher would be inclined to follow

[1] The scatter in intervention phase data is not taken into account as an improving trend (as a kind of instability) might be confounded with unexplained variability.

such an approach. In contrast, the second procedure is more accessible, but evidence is still needed on the performance of the both of them before judging their usefulness. Another difference stems from the fact that the MRA requires first computing an effect size (e.g., a nonoverlap index, a standardized mean difference) and then estimating the likelihood of the outcome. In contrast, SMIQR does not entail using any primary analytical technique. Moreover, for SMIQR there are only three labels assigned to the data sets: no effect, medium effect, and large effect, whereas the fact that probabilities represent a continuous variable makes possible establishing more labels.

Figure 1 presents an example using one of the studies included in the field test (Labelle & Mihailidis, 2006), illustrating the MRA and SMIQR (suggesting a medium effect) and how it can be used together with NAP, which in this case is equal to 87.64%. For this example, the probability of such a large NAP value being obtained in absence of intervention effect and for an autocorrelation ρ_1= .752 is between .10 and .05 according to MRA, denoting a difference very unlikely to be observed by chance only. Both approaches suggest the that a change in the behaviour has taken place, differing in terms of the degree of change, which is likely to be related to the fact that NAP does not take the improving trend into account and thus MRA does not do that either. Actually, according to the SMIQR criterion using 1.5 IQR the effect is on the limit, given that there is only one intervention phase measurement (slightly) out of the predicted range of values, which makes the decision about the presence of effect difficult. Therefore, the data presented in Figure 1 illustrate well the need to take into consideration the suggestions made by analytical techniques together with the essence of these techniques (i.e., what aspects of the data are taken into account and what is quantified). It is still up to the practitioner or applied researcher to decide whether, in cases similar to this one, baseline trend needs to be controlled for in order not to overestimate the behavioural

change. Thus, it is important to control for baseline trend when it is clear (Parker et al., 2014) and to be aware that such control has been applied to avoid omitting effects (Mercer & Sterling, 2012).

INSERT FIGURE 1 ABOUT HERE

R code for the two approaches

SMIQR can be applied by hand calculation, as it also the case for certain nonoverlap measures when the data series is relatively short. Nevertheless, we have also developed R code for that purpose. The SMIQR together with NAP can be applied using the R code freely available on https://www.dropbox.com/s/25vkk68xf60o26d/SMIQR_NAP.R (and as online supplementary material) so that a representation similar to Figure 1 can be obtained. This code only requires entering the data, specifying the number of baseline measurements, and the aim of the study (to increase or reduce target behaviour). We also developed R code for implementing the more computer-intensive probability approach, freely available at https://www.dropbox.com/s/56tqnhj4mng2wrq/Probabilities.R and as online supplementary material. It includes NAP, Percentage of nonoverlapping data (PND; Scruggs, Mastropieri, & Casto, 1987), two standardized mean differences, Slope and level change procedure (SLC; Solanas, Manolov, & Onghena, 2010); Mean phase difference (MPD; Manolov & Solanas, 2013a). We did not include in the code Tau (Parker, Vannest, Davis, & Sauber, 2011) and the SCED-specific *d*-statistics (Hedges, Pustejovsky, & Shadish, 2012; 2013), as their codes were not developed by us and their use is somewhat more complicated. The MRA code we developed requires the same

information as above (scores, baseline phase length, and aim of the intervention) plus choosing one of the abovementioned procedures which is then computed, a probability is obtained and translated into a label. (See the Data analysis subsection for more details.)

A Field Test

Aims

The aim of this study is twofold: (a) to compare the results of different procedures applied to the same set of studies dealing with the same topic, taking into account the fact that the probability-based approach is applicable to any type of quantification and there are several computed here; and (b) to test the performance of two procedures intended to aid SCED data analysis, as they are procedures which need more evidence before being recommended or discarded.

The issue of SCED data analysis has received a lot of attention, with a myriad of procedures already available. However, there has been more research dedicated on studying their performance individually and less work focusing on the degree to which similar results are obtained across several procedures when applying them to the same real data for which the truth about intervention effectiveness is unknown. One such recent effort (Shadish, 2014) focused on a single study and there have been some doubts expressed regarding the choice of the data (Fisher & Lerman, 2014). The current paper aims to start filling this research gap by applying a variety of procedures to the same data, with the particularity that these studies have been selected according substantive criteria and not due to being easily analyzed. The two approaches on which we focus represent two very distant ways of dealing with the same topic: as a visual aid and as a

probability. Thus we want to explore to what extent is there similarity in the results obtained using different procedures. This is important, given that practitioners and applied researchers would have to deal with situations with uncertainty about the magnitude of effect and the possibility to reach different conclusions with different analytical techniques.

In order to be able to recommend or discard any proposal intended to enhance data analysis, it is indispensable to provide evidence on its performance. Regarding the SMIQR, it is necessary to test to what extent the formal rule that the procedure offers agrees with (a) visual analysis performed without it (considering the recommendation made by Wolery, Busick, Reichow, and Barton, 2010) and (b) the judgements of the professionals and researcher who actually carried out the studies. Regarding the MRA, it was intended to help translating different metrics into a common one, but this translation has not been tested extensively with real data and with the amount of quantifications included here. Thus, the current study offers further evidence on that matter.

In summary, we compare the labels provided by SMIQR and MRA with: a) the labels according to the opinions expressed by the primary authors of the studies; b) the labels according to the visual analysis performed by two of the authors of the current paper; c) the values of the indices obtained. The context of application is a set of studies included in a meta-analysis on the effect of cognitive prosthetic technology (e.g., mobile phone, computer or television-based prompting device) for people with memory impairments, mostly result of acquired brain injury or degenerative disease (Jamieson et al., 2014).

Data analysis

Data Selection

In the current study we use the data of a published meta-analysis (Jamieson et al., 2014) and readers interested in the details about the bibliographic search, inclusion criteria, and access to and retrieval of the data are directed to it. Seventeen SCED studies published between 1987 and 2013 are included (and can be consulted from Jamieson et al., 2014), with a total of 38 participants. The studies used mostly reversal or AB designs replicated across participants and recorded as outcome the frequency or percentage of events missed or attended and success rates or number of errors in a task. The amount of studies included appear to be typical for a SCED meta-analysis, given that there is evidence suggesting that 60% of the meta-analyses include fewer than 30 studies (Moeyaert, Ugille, Ferron, Beretvas, & Van Den Noortgate, 2013).

Data access and coding

The following information was obtained for each of the 38 participants in the 17 studies: 1) explicitly stated efficacy category by the primary authors; 2) the rating of two of the authors of the current paper (MJ and JJE); 3) quantification by the effect size indices included here. (These three pieces of information described are part of a larger study [Manolov, Jamieson, Evans, & Sierra, 2015].) Regarding the primary authors' judgements on intervention effectiveness, these were coded initially by one of the authors (MJ) into five categories, as a result of reading thoroughly the articles and the primary authors' descriptions and evaluative comments. The categories with the following or equivalent descriptors used are: 0 = 'decline', negative impact', 'worse' after intervention; 1 = 'no effect', 'no improvement', 'no change'; 2 = 'somewhat', 'unclear', 'small improvement'; 3 = 'improvement' (with no qualifying adjective), 'good', 'moderate'; 4 = 'large', 'marked', 'substantial' improvement. The ratings of MJ for each

paper were then reviewed by a second researcher (JJE). If the second researcher agreed with the coding, this was then used. If the second researcher disagreed with the coding there was a discussion in order to reach consensus. In all cases it was possible to reach a consensus, which is the final aim of any such coding procedure, regardless of whether there are initially codes assigned independently by each researcher or not. This piece of information is subsequently referred to in the text, figures, as tables as "primary authors' judgement". (All the relevant information about the studies, including these codes and actually used terms by each of the primary researchers is included as online supplementary document.)

Our own assessment of intervention effectiveness was provided by two of the authors (MJ and JJE) reaching a consensus after inspecting the visual display of the data following the steps described by Kratochwill et al. (2010): (1) assess whether the pattern of baseline measurements is clearly defined and can be used for comparison with subsequent phases, (2) assess within-phase level, trend, and variability in order to identify whether this within-phase pattern is predictable; (3) compare adjacent phases in terms of level, trend, and variability, as well as overlap, immediacy of the effect, and consistency of patterns in similar phases; and (4) assess whether, when comparing observed and projected patterns, all the data in the design suggest that the expected changes take place at the expected points in time (i.e., whether the effects are replicated). The two visual analysts had no knowledge of the quantifications and SMIQR or MRA labels in the current paper, as the latter were obtained afterwards. This piece of information is subsequently referred to in the text, figures, as tables as "our own visual analysis".

It should be noted that only two individuals served as analysts of the graphed data and the visual judgements are not the product of the consensus of a large group of

experts. Therefore, these judgements should not be understood as representing the truth about the magnitude of effect, as such truth remains unknown, especially considering the evidence on the performance of visual analysts (e.g., Danov & Symons, 2008; Ninci, Vannest, Willson, & Zhang, 2015; Ottenbacher, 1993).

Regarding quantifications, the statistical analysis included the following techniques: PND (Scruggs et al., 1987), NAP (Parker & Vannest, 2009), Tau-U (Parker, Vannest, Davis, & Sauber, 2011; using the R code described in Brossart, Vannest, Davis, & Patience, 2014, downloaded from https://dl.dropboxusercontent.com/u/2842869/Tau_U.R), Allison and Gorman's (1993) regression model; Standardized mean difference using the pooled estimation of the standard deviation (Cohen's, 1992, *d*) and using only the baseline phase standard deviation (paralleling Glass' delta, *Δ*; Glass, McGaw, & Smith, 1981); the *d*-statistics by Hedges et al. (2012 [for $(AB)^k$ designs]; 2013 [for multiple-baseline designs], using the R code from James Pustejovsky's page http://blogs.edb.utexas.edu/pusto/software/); SLC (Solanas, Manolov, & Onghena, 2010); MPD (Manolov & Solanas, 2013a) and SMA (Borckardt et al., 2008). We chose to compute that many (although not all possible) single-case indices in order to test in a broader context how well the MRA functions as a mechanism translating different metrics into similar labels.

Maximal reference approach

The probabilities provided by the MRA were categorized using the criterion used by Parker, Vannest, Davis, and Sauber (2011) in the field test of the Tau-U nonoverlap index – percentiles 10, 25, and 50. Therefore, a probability lower than or equal to .10 (but greater than .05) reflected a difference "very unlikely" to be due to chance, a probability value between .10 and .25 refers to an "unlikely" null difference, and a

probability value between .25 and .50 a "somewhat unlikely" null difference. Finally, a probability as small as or smaller than the conventional .05 level was labelled as a difference that is "extremely unlikely" in absence of effect. For obtaining the probabilities we used three different random disturbance distributions[2] (normal, exponential, and uniform) incorporated in the u_t term of the first-order autoregressive model: $\varepsilon_t = \rho_1 \cdot \varepsilon_{t-1} + u_t$. For each analytical procedure 1,000 samples were generated in R (R Core Team, 2013) in order to obtain the location of the actual quantification in the distribution of values expected in absence of intervention effect and in presence of the pre-specified autocorrelation ρ_1, according to the evidence provided by Shadish and Sullivan (2011). Therefore, for each value of each of the analytical procedures used we obtained the label, after translating the probabilities yielded by MRA.

Split-middle trend and envelope based on interquartile range

For implementing SMIQR, an effect was labelled as "medium" when there is at least one intervention phase measurement beyond the 1.5 IQR limits of the split-middle projected trend. Analogously, an effect was labelled as "large" when there was at least one treatment measurement beyond the 3 IQR limits. In both cases only deviations in the direction of improvement were considered.

Quantification of the agreement in the labels assigned

The degree of agreement between the labels assigned via the MRA and SMIQR was assessed through Goodman and Kruskal's (1954) gamma coefficient. The same index

[2] Random variables are drawn from three different distributions in order to relax the assumption made in Simulation modeling analysis (Borckardt et al., 2008) that the underlying distribution is normal. Normality is also assumed regarding the residuals, data, or effects in multilevel models (Moeyaert et al., 2014) and in the SCED-specific *d*-statistic (Hedges et al., 2012; 2013). The alternative distributions were chosen due to difference in skewness (the normal distribution is symmetric, whereas as the exponential is positively skewed) and in kurtosis (the uniform distribution is more platykurtic than the normal) in order to represent a broader set of conditions.

was used to explore the degree to which the MRA served for translating the outcome of the different techniques into the similar labels. For the association between SMIQR (or MRA) labels and the values of the effect size indices computed we provide grouped boxplots, appropriate for representing the relationship between a qualitative and quantitative variables. The association between the labels stemming from the different sources (SMIQR, MRA, proponents' benchmarks for the indices, our own visual analysis) graphical representations are provided as well.

Results

Relationship between the approaches and the quantifications provided by the effect size indices

The results represented on Figure 2 show the relationship between the labels provided by SMIQR and the nonoverlap and standardized difference indices. There are some differences between the nonoverlap indices themselves (NAP yielding greater values than PND, which yields greater values than Tau), but in all cases SMIQR labels suggesting larger effect are associated with higher values of the indices. This association is clearer for standardized difference indices and NAP, but for Tau there is no clear difference between "medium" and "large" effects, and for PND there is excessive variability.

INSERT FIGURE 2 ABOUT HERE

For MRA the distinction is clearer than for SMIQR, as illustrated on Figure 3 (for nonoverlap indices) and Figure 4 (for standardized difference indices), given that the labels provided by the probability based approach are based on the quantifications obtained (and also on the number of measurements).

INSERT FIGURES 3 AND 4 ABOUT HERE

Figure 5 represents the relationship between the labels provided by SMIQR (upper panel) and MRA applied to the Tau value (lower panel) and the probabilities provided by SMA (represented on the Figure Y-axes). The SMA probabilities are also expected to reflect the degree of intervention effectiveness, with smaller *p* values expected to be related to stronger effects. For SMIQR the relationship is not optimal, but there is a clear coincidence in distinguishing studies with and without effect. For MRA applied to Tau, the results are clearer, with both very unlikely and extremely unlikely null differences related to statistically significant results ($p \leq .05$) according to SMA.

INSERT FIGURE 5 ABOUT HERE

Relationship between the approaches and the labels provided by primary authors or visual analysis

Regarding SMIQR (see Figure 6), the association is clearly imperfect and apparently low, especially with primary authors' judgements. Nevertheless, some association can be seen in: (a) "no effect" researchers' judgements being less frequent for SMIQR labels suggesting larger effects; (b) "medium effect" researchers' judgements being more frequent for SMIQR labels suggesting medium effect; and (c) "large effect" researchers' judgements being more frequent for SMIQR labels suggesting medium and large effects.

INSERT FIGURE 6 ABOUT HERE

Regarding MRA, Figures 7 and 8 focus on Tau and Delta, as we considered these two to be the potentially most useful representatives of nonoverlap and standardized mean difference indices, respectively. For MRA labels assigned to the value of Tau, the

agreement with researchers' judgements is seen in the "no effect" judgement being present only for labels suggesting a greater likelihood of a null difference, whereas medium and large effect judgements are more common for smaller probabilities. For MRA labels assigned to the value of Delta, the pattern is similar, but the association is even clearer, as judgements of large effect are even less common for data labelled by MRA as a likely or somewhat likely null difference. Despite that, once again, it has to be stated explicitly that the relationship between the different sources of labels is not close to being perfect, as it can be seen from Table 1 including Goodman-Kruskal gamma's values for MRA applied to all procedures included in the study. Note that the relation with the primary authors' judgments is low or even in the opposite direction, whereas the relation with our own (independent) visual analysis much stronger.

INSERT FIGURES 7 AND 8 ABOUT HERE

INSERT TABLE 1 ABOUT HERE

Relationship between the labels provided by the two approaches

Apart from comparing the results yielded by the approaches to the judgements of researchers, it is also relevant to explore how well the SMIQR and MRA agree. Figure 9 illustrates the findings for the labels assigned to the results of a nonoverlap index (Tau) and a standardized difference index (Delta): the same MRA labels which were presented in Figures 7 and 8. The association between the two sources of labels is made evident in the fact that a likely null difference according to the MRA only appears for a "no effect" label by SMIQR, whereas as MRA labels suggesting the smallest probabilities of a null difference are related to SMIQR labels also indicating larger effects. However, the strength of association is not constant across all effect size indices, as it can be seen from Table 1, with greater matching for MRA applied to NAP and to the standardized

mean difference using pooled standard deviation, lower agreement for the MRA applied to the Allison and Gorman model.

INSERT FIGURE 9 ABOUT HERE

Maximal reference approach as a translation mechanism

Table 2 contains the values of Goodman-Kruskal's gamma coefficient that quantify the relationship between the labels assigned to the different indices' values; the upper diagonal matrix is based on the data per participant and the lower diagonal matrix on the data per study (after averaging all within-study outcomes to produce a single effect or label per study). Note that the SCED-specific *d*-statistic can only be computed at the study level (i.e., it is no available as a column in the table) as several subjects are required in the computation. The correlation values suggest that in most cases the association between the labels is strong, especially when considering the nonoverlap indices (PND, NAP, and Tau), and the standardized differences (Delta, Pooled, *d*-statistic, MPD, level change estimate of the SLC). In contrast, the relationship with the Allison and Gorman model quantification is not that strong.

INSERT TABLE 2 ABOUT HERE

Discussion

The current studied aimed to explore, in the context of a set of studies selected on substantive (rather than analytical) basis, whether different procedures (visual analysis, a visual aids with a formal decision rule, and a probability-based approach applied to a

variety of techniques) agree in the magnitude of effect present. We also tested whether the probability-based approach works well as a translation mechanism.

Regarding the visual aid based on estimating and projecting split-middle trend (SMIQR), a study using generated data had suggested that it works well in most cases studied (Manolov et al., 2014), with the decisions made via SMIQR matching well simulation parameters. The current field test adds further evidence on the reasonably good correspondence between SMIQR and the values of SCED effect sizes, although some specific problems were detected for PND and Tau. In contrast, the matching between SMIQR and researchers' judgments and our own visual analysis was less than optimal. If we focus on this latter piece of evidence, apparently SMIQR is better for distinguishing between presence and absence of effect than between different magnitudes. However, larger indices' values are related to the "large effect" label.

Regarding the probability-based approach (MRA), previous evidence suggested that it works well as a translation mechanism between nonoverlap indices (Manolov & Solanas, 2013b). The evidence presented here shows that the similarities extend to standardized difference indices, given the positive and high correlations obtained. Moreover, the labels provided by MRA also match several other pieces of information: (a) to a greater degree the values of the SCED indices included here and the SMIQR criterion, especially for NAP, Tau, and the standard mean difference indices; (b) to a moderate degree our own visual analysis; and (c) practically not related to the judgements of primary researchers. Nevertheless, the correspondence between the different categories of likelihood and the labels amount of variability of effects is not perfect, according to the current data.

Implications and Recommendations for Applied Researchers

The results presented above suggest that both the visual aid and the probability based approach can be used for distinguishing between conditions with and without and effect, helping researchers to discern whether the effect observed is rather small or rather large. Specifically, the visual aid SMIQR can be used to compare the projected and the actual data pattern, as a fundamental part of visual analysis (Kratochwill et al., 2010), whereas the magnitude of effect can be further assessed using MRA as applied to NAP or Delta. However, the procedures tested here cannot be considered perfect means of assigning labels and further assessment using different pieces of information is required, for instance, substantive criteria and across-studies comparisons of effects (Parker & Vannest, 2009).

We have provided R code for the SMIQR approach and NAP together, given that they inform about different aspects of the data. The former deals with trend and enables comparing the projected-as-if-no-effect treatment data and the actual treatment phase measurements, whereas the latter deals with data overlap. We have also provided R code for obtaining the labels according to the probability-based Maximal reference approach, using several different indices. In this way, applied researchers have available a graphical representation of the data, a visual aid taking trend into account, the result of a quantification not dealing with trend (in case the researcher consider its estimation to be imprecise and /or its projection to be unrealistic), and a label about the degree of effectiveness according to the MRA. With these pieces of information and on the basis of their knowledge of the client, target behaviour, and context, a solid decision about the change in target behaviour can be made. In the current context of a diversity of SCED data analytical techniques focussing on different aspects of the data, the more procedures converge on the same conclusion, the greater the confidence of an applied researcher would be on the (degree of) effectiveness of the intervention.

We have here chosen to point out NAP as a quantification (as neither SMIQR nor MRA are stand-alone quantifications), but applied researchers are adverted that there are other possibilities, such as the ones mentioned in the Introduction. Criteria for choosing among these procedures have been proposed (Manolov, Gast, Perdices, & Evans, 2014; Wolery et al., 2010) and can be useful in the task of selecting how to analyse the data at hand. Moreover, given the imperfect agreement between the procedures tested here, applied researchers are adverted about the possibility

In any case, it has to be kept in mind that the presence of causal effects is assessed first and foremost on the basis of the design and following current standards and guidelines when collecting data (Horner et al., 2005; Kratochwill et al., 2010; Reichow et al., 2008; Tate et al., 2013). In that sense, quantitative analysis is only one piece of information when deciding whether a study and its results contribute to establishing the evidence base of the treatment (or lack thereof).

Limitations and Future Research

It should be noted that given that the current study is not based on simulation, the truth about the exact magnitude of effect is not known – a situation that practitioners have to deal with in real life. Given the unknown parameters of the underlying data generation process, the current paper does not help deciding which of the techniques is optimal or which techniques represent the features of the data more closely. It only intends to suggest tools and test how well they converge in suggesting the same or similar degree of intervention effectiveness.

Regarding the main aims of the study, until more evidence is available, it cannot be clearly recommended the use of SMIQR for distinguishing between medium and large effects, whereas the MRA offers more categories and is better equipped to distinguish between likely, unlikely, and very (or extremely) unlikely. There is still a real possibility that different analyses of the same lead to different conclusions, but the agreement that an effect has taken place on the basis of a variety of visual and quantitative criteria is a step forward. Research is still needed on establishing benchmarks for SCED effect size measures, as this topic has not been studied extensively.

The R code provided deals with two-phase comparisons only and thus applied researchers would have to use several times according to the design structure, for instance, three times for a multiple-baseline design across three participants. After the two-phase comparison results are obtained, the researcher still has to (decide how to) combine the information in order to have an estimate for the whole (e.g., multiple-baseline,ABAB) design. This decision is inherent to SCED data analysis, although procedures like randomization tests (Heyvaert & Onghena, 2014; using the software described in Bulté & Onghena, 2008; 2009), the *d*-statistics (Hedges et al., 2012, 2013) or multilevel models (Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2014), are more directly applicable to designs involving within- or across-subjects replication.

References

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.

Borckardt, J. J., Nash, M. R., Murphy, M. D., Moore, M., Shaw, D., & O'Neil, P. (2008). Clinical practice as natural laboratory for psychotherapy research: A guide to case-based time-series analysis. *American Psychologist, 63*, 77-95.

Brossart, D. F., Vannest, K., Davis, J., & Patience, M. (2014). Incorporating nonoverlap indices with visual analysis for quantifying intervention effectiveness in single-case experimental designs. *Neuropsychological Rehabilitation, 24*, 464-491.

Bulté, I., & Onghena, P. (2008). An R package for single-case randomization tests. *Behavior Research Methods, 40*, 467-478.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Danov, S. E., & Symons, F. J. (2008). A survey evaluation of the reliability of visual inspection and functional analysis graphs. *Behavior Modification, 32*, 828-839.

Fisher, W. W., Kelley, M. E., & Lomas, J. E. (2003). Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. *Journal of Applied Behavior Analysis, 36*, 387-406.

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.

Gast, D. L., & Spriggs, A. D. (2010). Visual analysis of graphic data. In D. L. Gast (Ed.), *Single subject research methodology in behavioral sciences* (pp. 199-233). London, UK: Routledge.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross-classification. *Journal of the American Statistical Association, 49*,732-764.

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.

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.

Howick, J., Chalmers, I., Glasziou, P., Greenhaigh, T., Heneghan, C., Liberati, A., et al. (2011). The 2011 Oxford CEBM Evidence Table (Introductory Document). *Oxford Centre for Evidence-Based Medicine*. http://www.cebm.net/index.aspx?o=5653

Huitema, B. E., & McKean, J. W. (1991). Autocorrelation estimation and inference with small samples. *Psychological Bulletin, 110,* 291-304.

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.

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).* Available at http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_scd.pdf

Labelle, K.-L., & Mihailidis, A. (2006). The use of automated prompting to facilitate handwashing in persons with dementia. *American Journal of Occupational Therapy, 60*, 442-450.

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., Jamieson, M., Evans, J. J. & Sierra, V. (2015). Establishing empirical benchmarks for interpreting single-case effect sizes: An illustration of a method. *Manuscript submitted for publication*.

Manolov, R., Sierra, V., Solanas, A., & Botella, J. (2014). Assessing functional relations in single-case designs: Quantitative proposals in the context of the evidence-based movement. *Behavior Modification, 38*, 878-913.

Manolov, R., & Solanas, A. (2012). Assigning and combining probabilities in single-case studies. *Psychological Methods, 17*, 495-509.

Manolov, R., & Solanas, A. (2013a). 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. (2013b). Assigning and combining probabilities in single-case studies: A second study. *Behavior Research Methods, 45*, 1024-1035.

Matyas, T. A., & Greenwood, K. M. (1991). Problems in the estimation of autocorrelation in brief time series and some implications for behavioral data. *Behavioral Assessment, 13*, 137-157.

Mercer, S. H., & Sterling, H. E. (2012). The impact of baseline trend control on visual analysis of single-case data. *Journal of School Psychology, 50*, 403-419.

Miller, M. J. (1985). Analyzing client change graphically. *Journal of Counseling and Development, 63*, 491-494.

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.

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S. N., & Van den Noortgate, W. (2014). The influence of the design matrix on treatment effect estimates in the quantitative

analyses of single-case experimental designs research. *Behavior Modification, 38,* 665-704.

Moeyaert, M., Ferron, J., Beretvas, S., & Van Den Noortgate, W. (2014). From a single-level analysis to a multilevel analysis of since-case experimental designs. *Journal of School Psychology, 52,* 191-211.

Ninci, J., Vannest, K. J., Willson, V., & Zhang, N. (2015, April 14). Interrater agreement between visual analysts of single-case data: A meta-analysis. *Behavior Modification*. Advance online publication. doi: 10.1177/0145445515581327

Ottenbacher, K. J. (1993). Interrater agreement of visual analysis in single-subject decisions: Quantitative review and analysis. *American Journal of Mental Retardation, 98,* 135-142.

Parker, R. I., Brossart, D. F., Vannest, K. J., Long, J. R., Garcia De-Alba, R., Baugh, F. G., & Sullivan, J. R. (2005). Effect sizes in single case research: How large is large? *School Psychology Review, 34*, 116-132.

Parker, R. I., Cryer, J., & Byrns, G. (2006). Controlling baseline trend in single-case research. *School Psychology Quarterly, 21,* 418-443.

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification, 35,* 303-322.

Parker, R. I., Vannest, K. J., & Davis, J. L. (2014). A simple method to control positive baseline trend within data nonoverlap. *Journal of Special Education, 48,* 79-91.

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.

Pfadt, A., & Wheeler, D. J. (1995). Using statistical process control to make data-based clinical decisions. *Journal of Applied Behavior Analysis, 28*, 349-370.

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Reichow, B., Volkmar, F., & Cicchetti, D. (2008). Development of the evaluative method for evaluating and determining evidence-based practices in autism. *Journal of Autism and Developmental Disorders, 38*, 1311-1319.

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education, 8*, 24-33.

Shadish, W. R. (2014). Analysis and meta-analysis of single-case designs: An introduction. *Journal of School Psychology, 52*, 109-122.

Shadish, W. R., Hedges, L. V., & Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology, 52*, 123-147.

Shadish, W. R., Rindskopf, D. M., Hedges, L. V., & Sullivan, K. J. (2013). Bayesian estimates of autocorrelations in single-case designs. *Behavior Research Methods, 45*, 813-821.

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.

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

Solanas, A., Manolov, R., & Onghena, P. (2010). Estimating slope and level change in N=1 designs. *Behavior Modification, 34*, 195-218.

Solanas, A., Manolov, R., & Sierra, V. (2010). Lag-one autocorrelation in short series: Estimation and hypothesis testing. *Psicológica, 31*, 357-381.

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.

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.

Tukey, J. W. (1977). *Exploratory data analysis*. London, UK: Addison-Wesley.

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *Journal of Special Education, 44*, 18-29.

Table 1. Values of Goodman-Kruskal's gamma correlation coefficient, showing the relationship between the labels for the effect size indices as obtained according to the Maximal reference approach, on the one hand, and the judgements of the primary researchers, our own visual analysis, or the visual aid referred to as SMIQR.

Source of labels	AG	PND	NAP	Tau	LC	MPD	Delta	Pooled
Primary authors	−.08	.14	.30	.35	−.19	−.22	.03	.49
Visual analysis	.35	.46	.61	.61	.40	.23	.60	.40
SMIQR	.10	.49	.71	.40	.53	.50	.62	.73

Note. SMIQR - split-middle trend line and its projection with the limits based on 1.5 times the interquartile range. AG – Allison and Gorman regression model. PND – Percentage of nonoverlapping data. NAP – Nonoverlap of all pairs. Tau – indicator combining nonoverlap and treatment phase trend. LC – level change estimate of the Slope and level change procedure. MPD – Mean phase difference. Delta – standardized mean difference using the baseline data standard deviation in the denominator. Pooled – standardized mean difference using the pooled baseline and treatment data standard deviation in the denominator.

Table 2. Values of Goodman-Kruskal's gamma correlation coefficient, showing the relationship between the labels for the effect size indices as obtained according to the Maximal reference approach. The upper diagonal matrix is based on the data per participant and the lower diagonal matrix on the data per study.

Index	AG	PND	NAP	Tau	LC	MPD	Delta	Pooled
AG		.34	.46	.30	.26	.29	.02	.12
PND	.16		.80	.62	.66	.69	.71	.73
NAP	.21	.79		.86	.73	.58	.99	.87
Tau	.67	.44	.76		.85	.70	.73	.53
LC	.43	.78	.89	.69		.98	.87	.67
MPD	.16	.56	.52	.59	.88		.61	.45
Delta	.48	.69	1.00	.81	.87	.57		.64
Pooled	.15	.63	.85	.72	.49	.38	.85	
d	.00	.79	.89	.57	1.00	.78	.89	.71

Note. AG – Allison and Gorman regression model. PND – Percentage of nonoverlapping data. NAP – Nonoverlap of all pairs. Tau – indicator combining nonoverlap and treatment phase trend. LC – level change estimate of the Slope and level change procedure. MPD – Mean phase difference. Delta – standardized mean difference using the baseline data standard deviation in the denominator. Pooled – standardized mean difference using the pooled baseline and treatment data standard deviation in the denominator. *d* – standardized mean difference specifically created for single-case designs.

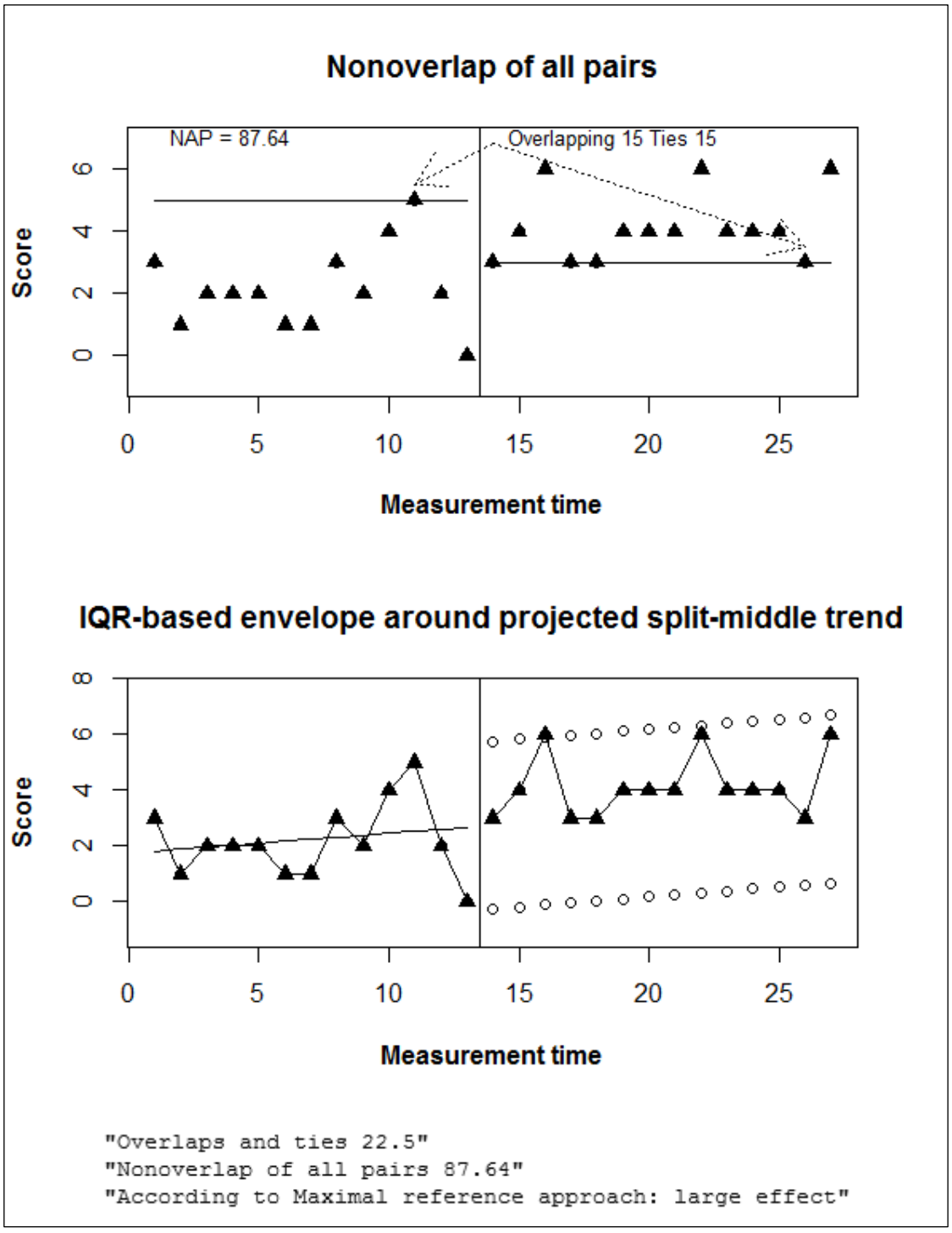


Figure 1. Data obtained by Labelle and Mihailidis (2006) with the aim being increase in the behaviour. The upper panel marks the maximal baseline measurement and the minimal baseline measurement and provides the value of the Nonoverlap of all pairs (arrows pointing at overlapping data). The lower panel provides the split-middle trend line and its projection with the limits based on 1.5 times the interquartile range of the baseline data. The two graphical representations are obtained using the code available at https://www.dropbox.com/s/25vkk68xf60o26d/SMIQR_NAP.R. The last three lines of results are obtained via https://www.dropbox.com/s/56tqnhj4mng2wrq/Probabilities.R.

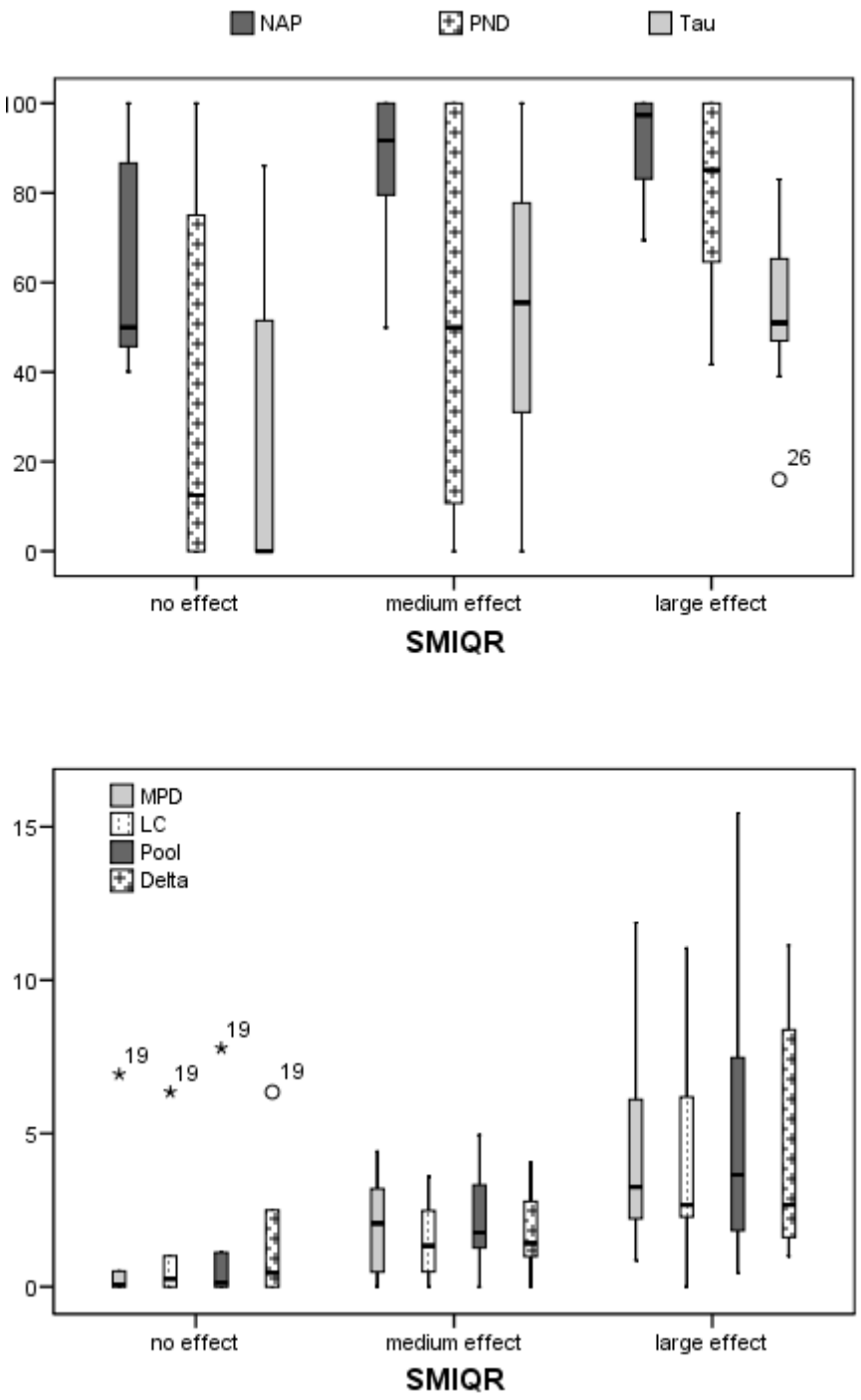


Figure 2. Graphical representation of the relationship between the SMIQR (split-middle trend line and its projection with the limits based on 1.5 times the interquartile range) labels and the values of nonoverlap indices (upper panel) and the standardized differences (lower panel).

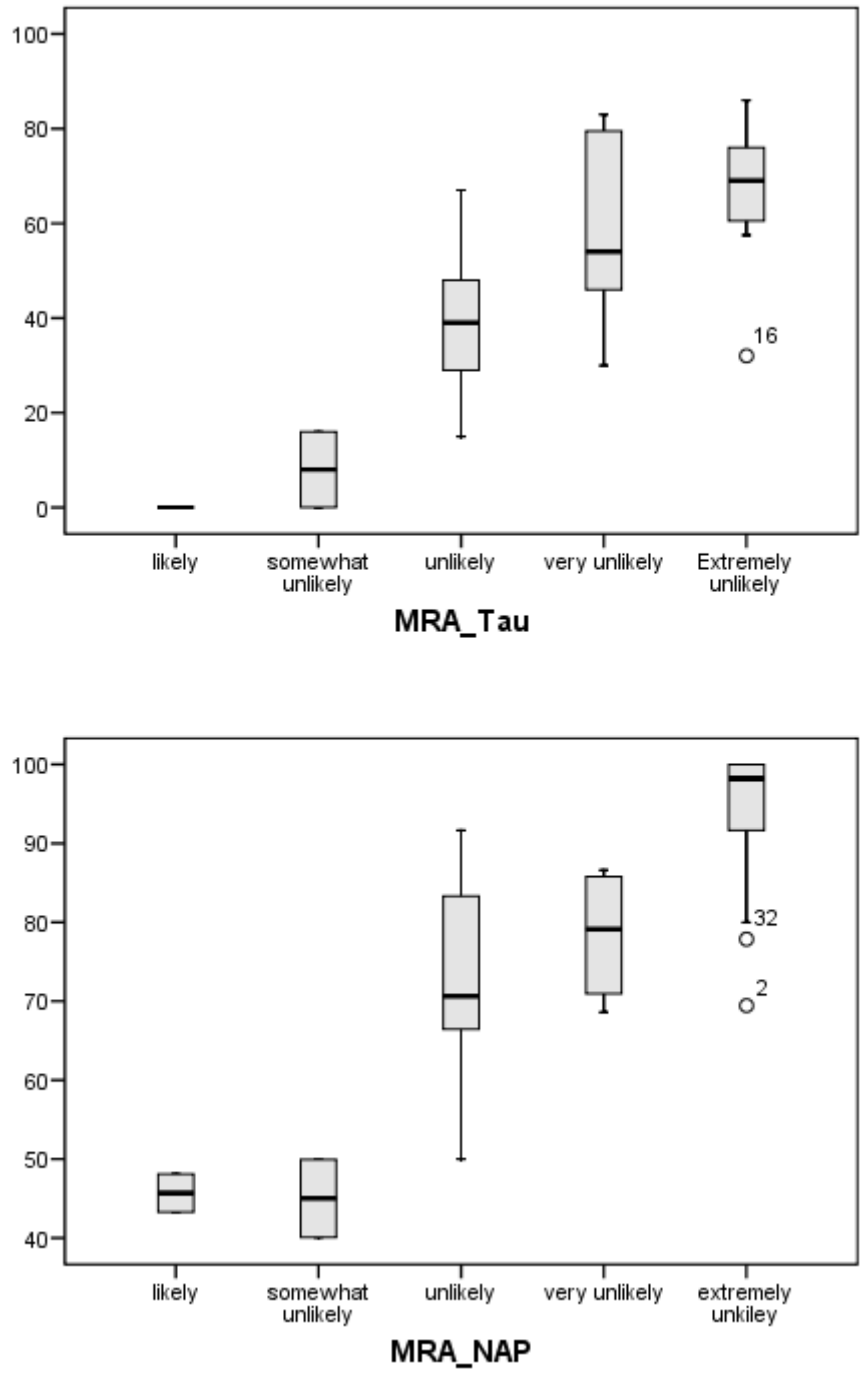
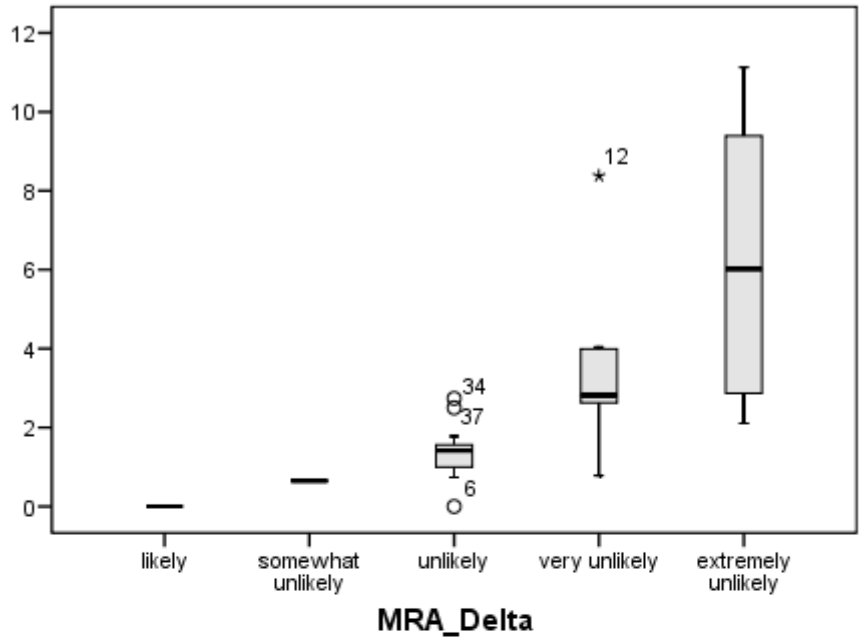


Figure 3. Graphical representation of the relationship between the Maximal reference approach labels for Tau (upper panel) and nonoverlap of all pairs (lower panel) and the values of these same indices (ordinate axes).



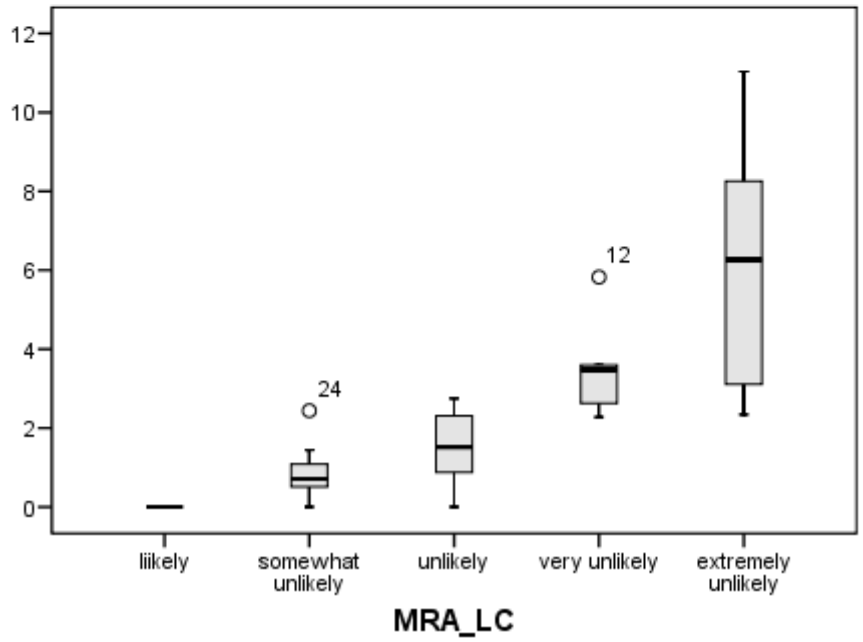
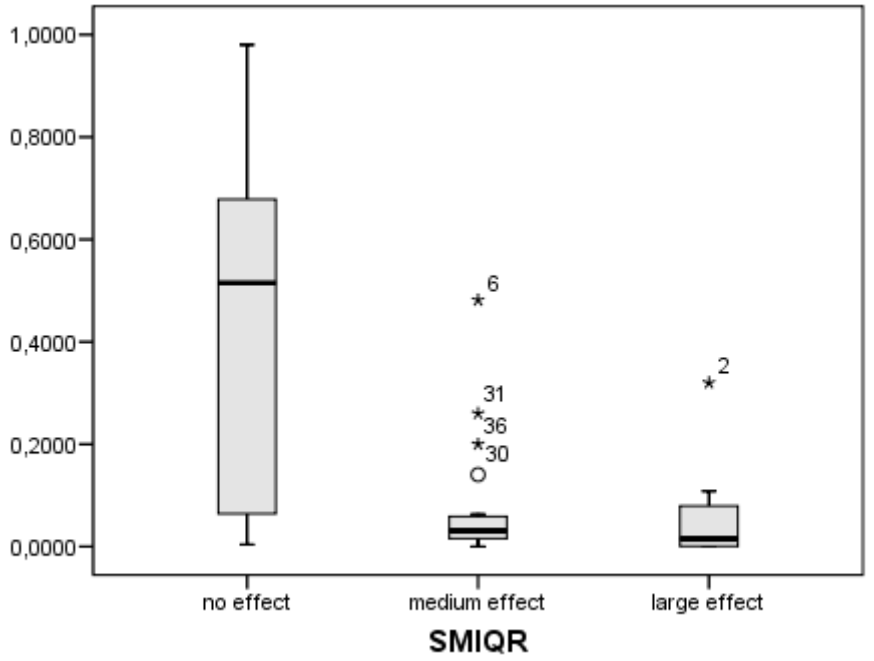


Figure 4. Graphical representation of the relationship between the Maximal reference approach labels for Delta (upper panel) and the level change estimate of Slope and level chance (lower panel) and the values of these same indices (ordinate axes).



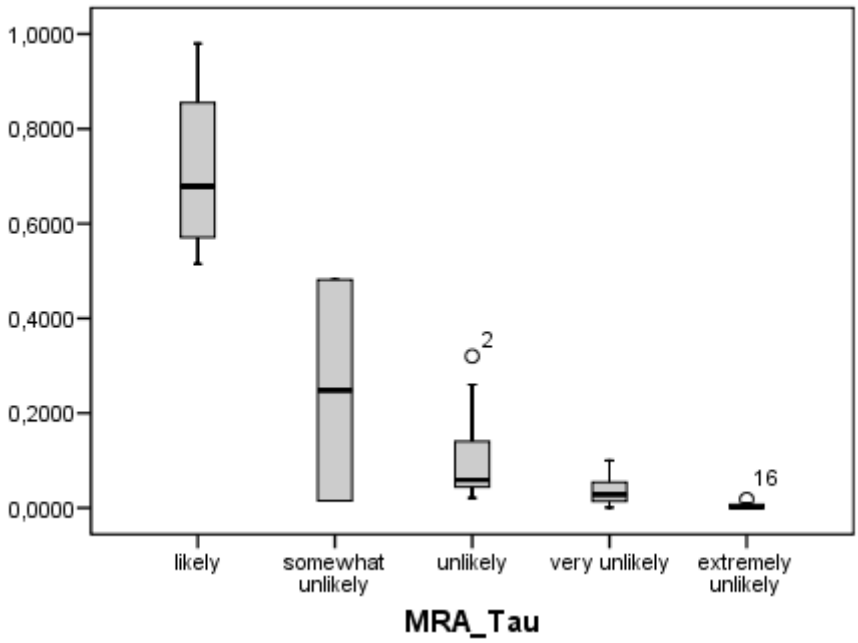


Figure 5. Graphical representation of the relationship between the probabilities provided by simulation modelling analysis (ordinate axes) and split-middle trend line and its projection with the limits based on 1.5 times the interquartile range labels (SMIQR, upper panel) and Maximal reference approach labels for Tau (MRA_Tau, lower panel).

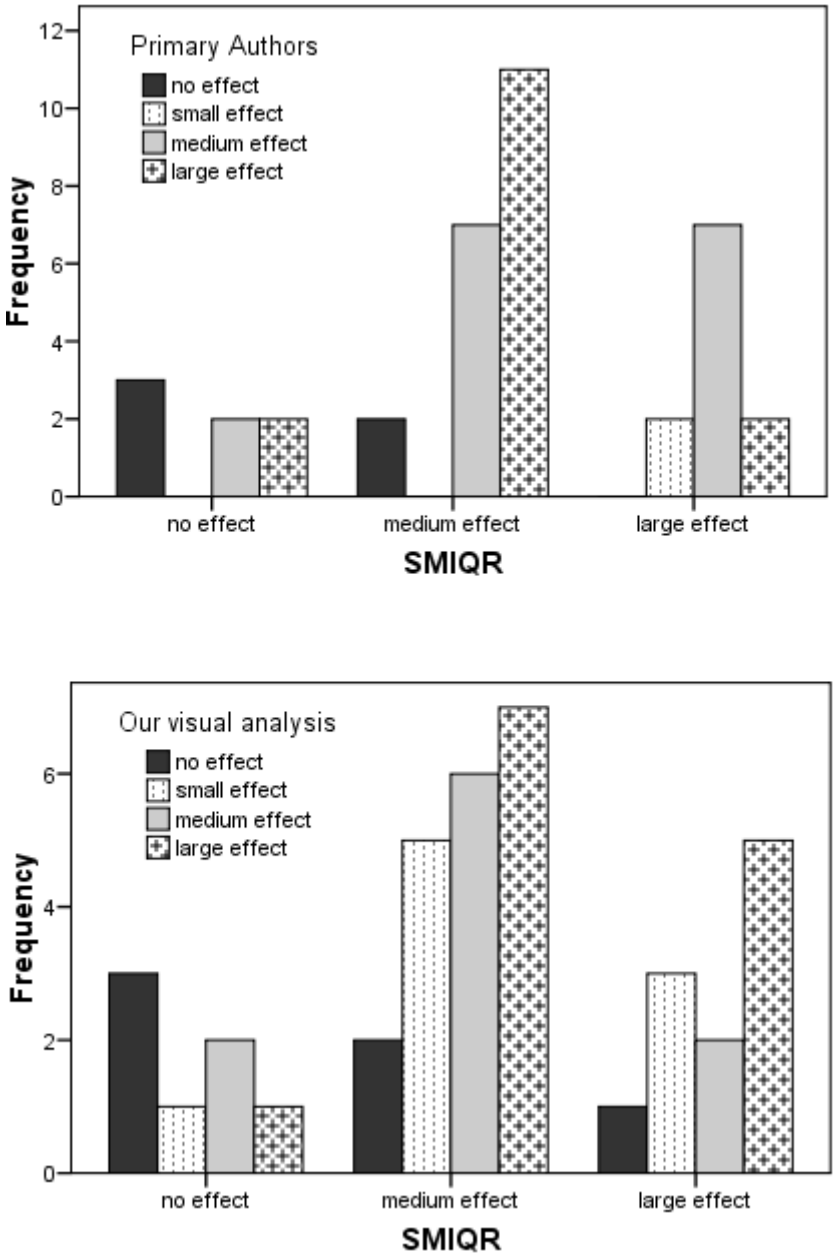


Figure 6. Graphical representation of the relationship between the SMIQR (split-middle trend line and its projection with the limits based on 1.5 times the interquartile range) labels and the judgements of the primary authors of the studies from the meta-analysis (upper panel) and our own visual analysis (lower panel).

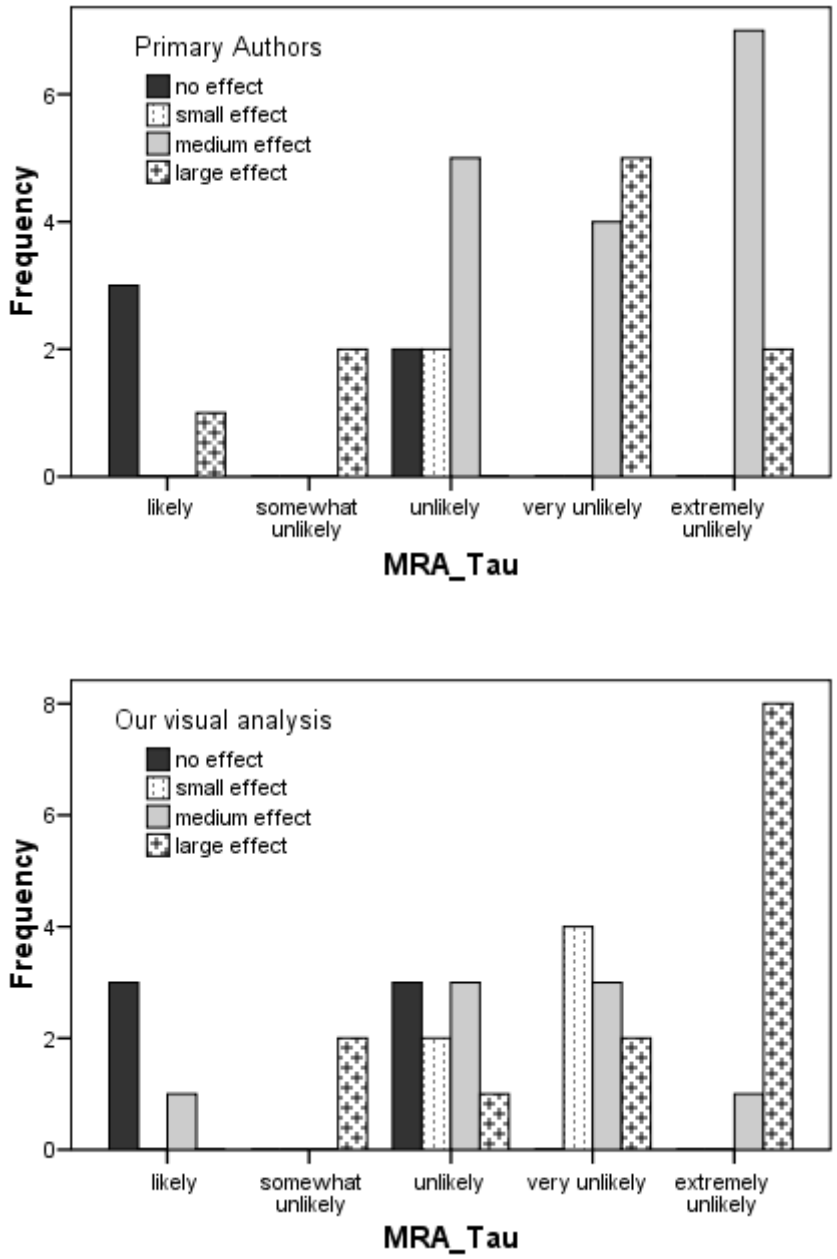


Figure 7. Graphical representation of the relationship between the MRA (Maximal reference approach) labels for Tau and the judgements of the primary authors of the studies from the meta-analysis (upper panel) and our own visual analysis (lower panel).

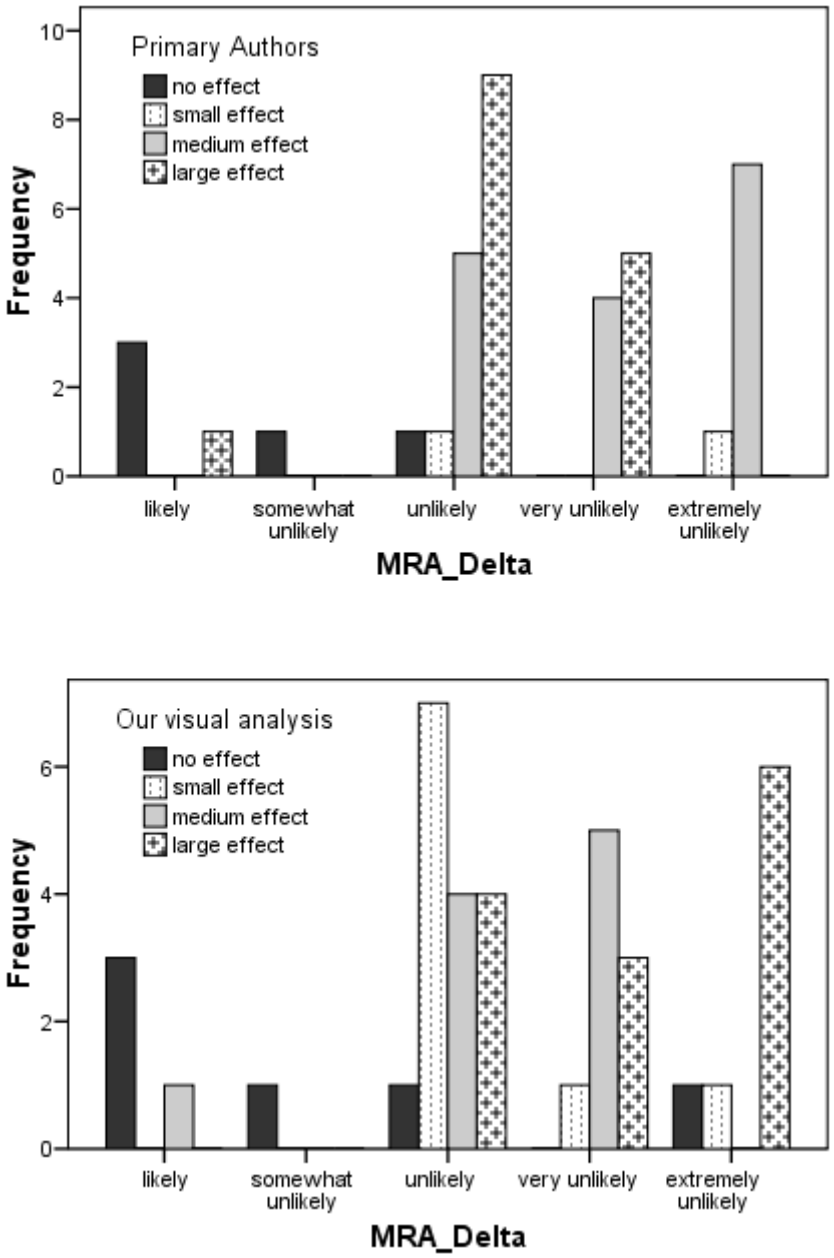


Figure 8. Graphical representation of the relationship between the MRA (Maximal reference approach) labels for Delta and the judgements of the primary authors of the studies from the meta-analysis (upper panel) and our own visual analysis (lower panel).

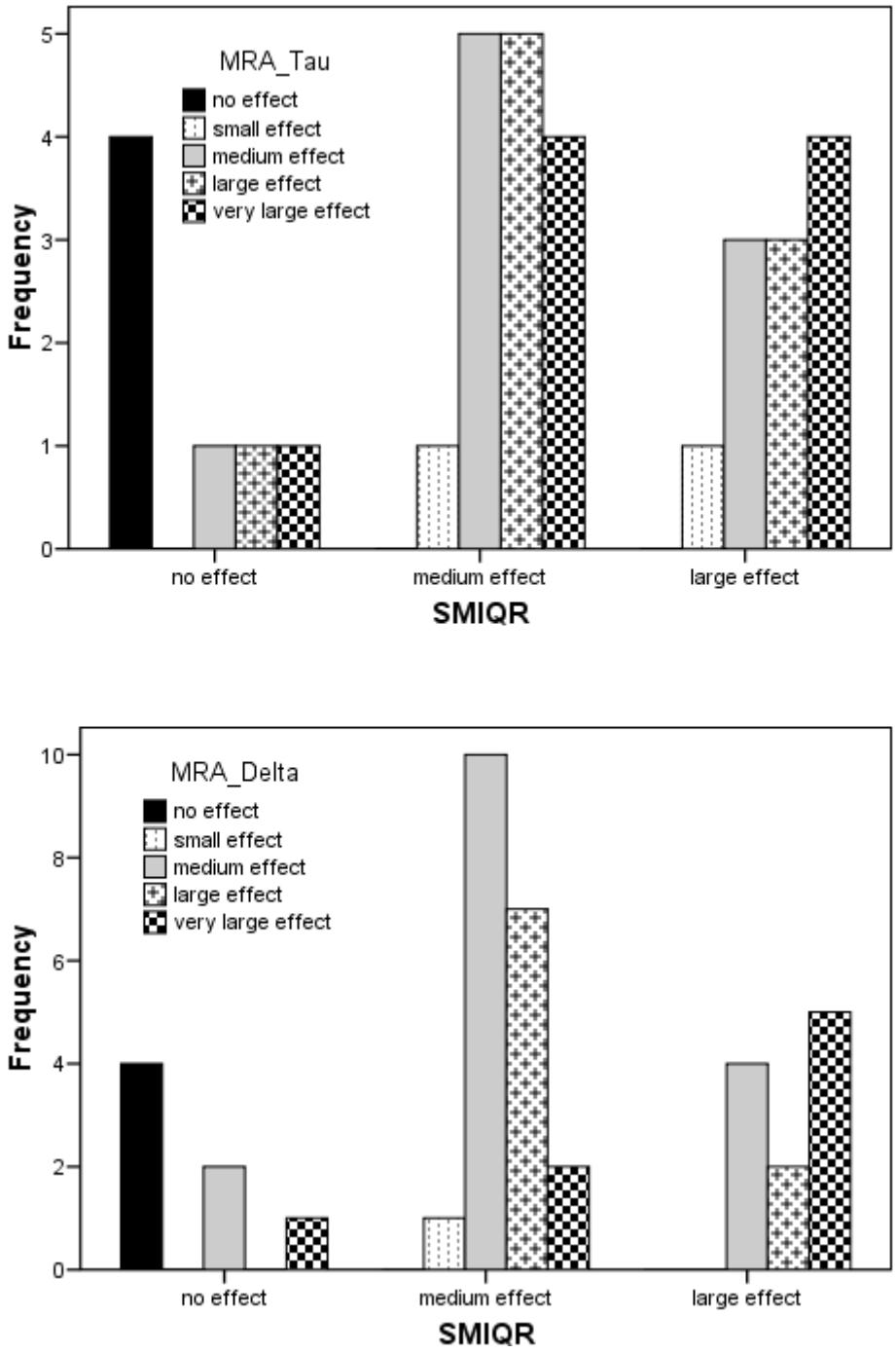


Figure 9. Graphical representation of the relationship between SMIQR (split-middle trend line and its projection with the limits based on 1.5 times the interquartile range) labels the MRA (Maximal reference approach) labels for the Tau (upper panel) and the Delta statistic (lower panel).