Assigning and combining probabilities in single-case studies:

A second study

Rumen Manolov[1 2] and Antonio Solanas[1 2]

[1] Department of Behavioral Sciences Methods, Faculty of Psychology, University of Barcelona.

[2] Institute for Research in Brain, Cognition, and Behavior (IR3C).

Running head

Assigning and combining probabilities

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. Electronic mail may be sent to Rumen Manolov at rrumenov13@ub.edu.

Authors' note

This research was partially supported by the Spanish *Ministerio de Ciencia e Innovación* grant *PSI2009-07076* and by the *Agència de Gestió d'Ajust Universitaris i de Recerca de la Generalitat de Catalunya* grant *2009SGR1492*.

Abstract

The present study builds on a previous proposal for assigning probabilities to the outcomes computed using different primary indicators in single-case studies. These probabilities are obtained comparing the outcome to previously tabulated reference values and reflect the likelihood of the results in case there was no intervention effect. The current study explores how well different metrics are translated into *p* values in the context of simulation data. Furthermore, two published multiple baseline data sets are used to illustrate how well the probabilities could reflect the intervention effectiveness as assessed by the original authors. Finally, the importance of which primary indicator is used in each data set to be integrated is explored; two ways of combining probabilities are used: a weighted average and a binomial test. The results indicate that the translation into *p* values works well for the two nonoverlap procedures, with the results for the regression-based procedure diverging due to some undesirable features of its performance. These *p* values, both when taken individually and when combined, were well-aligned with the effectiveness for the real-life data. The results suggest that assigning probabilities can be useful for translating the primary measure into the same metric, using these probabilities as additional evidence on the importance of behavioral change, complementing visual analysis and professional's judgments.

Single-case designs (SCD) are becoming increasingly accepted as a means of obtaining solid evidence on the effectiveness of psychological interventions (Blampied, 2000; Gedo, 2000; Horner et al., 2005), with still on-going efforts on further methodological improvements (Kratochwill et al., 2010; Tate et al., 2008). Considering the need to support empirically the interventions (APA Presidential Task Force on Evidence-Based Practice, 2006), the replication of experimental effects has been deemed necessary for a practice to be called "evidence-based" (Horner & Kratochwill, 2012). In that sense, and in relation to external validity, the importance of the synthesis and comparison of effect size measures (Smith, 2012) and meta-analysis (Burns, 2012) has been highlighted. In the following we review briefly some alternatives for integrating several SCD studies.

Combining SCD studies' results

Meta-analysis

Meta-analysis is currently the predominant way of integrating results, as it provides information on the magnitude of effect (Hedges, Cooper, & Bushman, 1992) and on the variability of effects across studies (Becker, 1987). The latter aspect is related to the importance of identifying and controlling moderator variables (Burns, 2012). The simplest way of combining is to compute the mean or median value of the indicator across studies (e.g., Schlosser, Lee, & Wendt, 2008), whereas another option is to compute a weighted average (e.g., Schneider, Goldstein, & Parker, 2008) when a reasonable weight can be computed. Meta-analysis can also be carried out via hierarchical linear models (HLM; Van den Noortgate & Onghena, 2003) either using raw data or combining standardized effect size measures. HLM are based on parametric assumptions and it is still necessary to obtain evidence on the performance of the effect size estimators and variance estimators, and also on the confidence intervals coverage in typical SCD data (Ferron, Farmer, & Owens, 2010; Owens & Ferron, 2012). Further proposals have also focused on solid ways of deciding whether the evidence available from the studies conducted up to a certain point in time is sufficient (Kuppens, Heyvaert, Van den Noortgate, & Onghena, 2011).

Combining probabilities: Maximal reference approach (MRA)

The combination of *p* values has been discussed thoroughly (Becker, 1987; Edgington, 1972; Rosenthal, 1978) before the wide spread of meta-analysis via effect sizes. In the context of SCD, this possibility is potentially useful, given the lack of consensus on which the most appropriate summary measure is (Burns, 2012; Horner & Kratochwill, 2012; Smith, 2012). Combining *p* values has already been considered as an option for integrating single-case studies (Bulté, Onghena, Salmaso, & Solmi, 2010; Onghena, 1994; see Holden, Bearison, Rode, Rosenberg, & Fishman, 1999 for an applied study measuring pain and anxiety in hospitalized children). This possibility has been incorporated in an R package making more practical carrying out visual, statistical, and meta-analyses (Bulté & Onghena, 2008; 2012).

We consider that combining probabilities may be a useful alternative to meta-analysis when raw data are not available and thus researchers cannot apply their indicator of choice. The main issues related to raw data access are the loss of precision when retrieving data from published graphs and a potential lack of response from the original authors, a problem reported by Shadish and Sullivan (2011).

In this context, the maximal reference approach (MRA) was proposed (Manolov & Solanas, 2012) with the aim to aid the interpretation, comparison, and combination of effect size measures expressed in different metrics. Nonetheless, it should be mentioned that labeling an index value as a “small” or “large” effect may be a result of joint visual and statistical analyses (Parker & Vannest, 2009; Petersen-Brown, Karich, & Symons, 2012).

The MRA is based on the idea present in simulation modeling analysis (SMA; Borckardt et al., 2008) in which the outcome (i.e., the value of the primary indicator for the actual data) is compared to the index values that could have been obtained in absence of intervention effect. A *p* value is obtained as a result of this comparison indicating the likelihood of such an extreme outcome under the null hypothesis. In that sense, a smaller *p* value may be used as an evidence of a larger effect.

The main difference between the MRA and SMA is that in the former it is not assumed that the random disturbance is normal; rather several models are compared, with different degrees of skewness and kurtosis. More importantly, in the context of MRA, autocorrelation estimation is avoided, as it has been shown to be problematic (Huitema & McKean, 1991). Instead a conservative solution is offered, constructing the sampling distribution for a moderately high degree of autocorrelation. Following the MRA, reference values for key *p* values (e.g., .01, .05, .10, .20, .30, and .50) are to be presented so that applied researchers would be able to compare their outcomes to these reference values and assign a *p* value accordingly. In that way, it is not required to carry out simulations in order to construct a sampling distribution in which to locate the outcome, as it would be necessary if SMA were followed.

The *p* values assigned to the primary indicators through the MRA can be integrated using the binomial test, as described by Darlington and Hayes (2000). A cut-off *p* value is established for distinguishing between “successes” (i.e., *p* values lower than or equal to the cut-off) and “failures”. Afterwards the probability of obtaining as many successes out of the amount of studies to integrate is computed. A small binomial probability would be indicative of a true positive effect does exist. Another option (Manolov &

Solanas, 2012) for integrating is just to compute the average of the *p* values, weighting each probability according to the standard error of the effect size measure originally used (i.e., using the Fisher information as a weight, as suggested by Whitlock, 2005).

Three aspects have to be highlighted regarding the possibility to assign *p* values via the MRA. First, the *p* values are only suggested as additional numerical evidence on the potential size of the results and are not intended to substitute visual criteria or the professional's judgment on which intervention effect is practically relevant. We should stress that *p* values are not sufficient for representing the type of effect present in SCD and that *p* values are not *per se* measures of magnitude of size, given that they are also affected by sample size and the amount of variability in the data. In that sense, effect size measures have been claimed more useful for quantifying treatment effectiveness (Hedges et al., 1992), especially considering the common inappropriate use *p* values in the context of hypothesis testing, the assessment of the practical significance or the replicability of results (Cohen, 1994; Nickerson, 2000). Specifically, separate quantifications of slope change and level change have been deemed necessary (Beretvas & Chung, 2008a) apart from the importance of considering data overlap (Parker, Vannest, & Davis, 2012).

Second, a distinction is necessary between the *p* values as a common metric for integrating studies (as in the MRA) and the *p* values as the main result of assessing intervention effectiveness. In the latter case, statistical significance can be obtained for SCD studies for specific situations in which visual judgments are made repeatedly (Ferron & Jones, 2006) or when randomization tests are employed (Edgington & Onghena, 2007) for quantifying the likelihood of the results obtained under the null hypothesis of no intervention effect. In contrast, in the context of the MRA, *p* values represent merely a transformation of the main quantification of the effect size rather than being the key outcome. Finally, the *p* values and their interpretation are subjected to the assessment of internal validity, as discussed in the following section.

Internal validity when combining results

Replication has been mentioned here as being crucial for external validity, but it is also indispensable for internal validity, that is, for demonstrating the functional relationship between the intervention and the target behavior (Kratochwill et al., 2010). For the same aim, randomization (e.g., selecting at random when to intervene in each baseline in a multiple baseline design) has been highlighted as an option for increasing the confidence that the intervention, rather than external factors occurring simultaneously, is the cause of the change in the behavior and also for strengthening the scientific credibility of SCD (Kratochwill & Levin, 2010).

Internal validity has to be assessed not only when reporting the results of an individual study, but also in meta-analytic integrations (Burns, 2012). In order to support causal inference, visual analysis is necessary for guiding the calculation of an effect size (Parker & Vannest, 2012). In that sense, it has to be stressed that combining results, either through effect sizes or through *p* values, ought to take place only when experimental control has been demonstrated. Hence, the MRA is not recommended to be used alone, but always accompanied by the visual inspection of the data pattern in order to assess whether it matches the expected one. The magnitude of an effect is also not entirely reflected by the *p* value; rather the professional's judgment on the importance of the effect has to be taken into account. Therefore, visual, numerical, and professional experience criteria have to be used jointly in a SCD data analysis or integration.

Study aims

The present study has four aims. The first objective is to explore how well the translation of different indicators into a common metric (i.e., *p* values) works. This aim is related to the suggestion made by Horner and Kratochwill (2012) regarding the need to compare results over estimators to see whether they yield consistent results about the degree of intervention effectiveness. In the current study, three different primary indicators are applied to the same data generated via Monte Carlo methods. Thus, an adequate performance of the MRA as a translation method requires similar *p* values assigned across all primary indicators.

The second aim is to compare the *p* values assigned via the MRA to the judgments of the authors of real-life psychological studies. In this case, evidence for the positive functioning of the MRA would be, for instance, smaller *p* values assigned to studies in which the effect is judged to be larger. The study with real data will also focus on whether the combined probabilities reflect the general assessment of results reported in the articles. Note that this second aim is intended to be a pilot study into the relationship between *p* values and researcher's assessment rather than a large-scale field test. The process for assigning *p* values and a graphical summary of the first two aims are represented on Figure 1.

INSERT FIGURE 1 ABOUT HERE

The third aim is to compare the *p* values obtained through the MRA and the SMA-based approach in the context of the real-life data for which the "true" degree of autocorrelation is unknown. We explore what are the gains, if any, of constructing sampling distributions that match as closely as possible the autocorrelation estimated from the data at hand (as done in SMA) instead of assuming a conservative moderately high degree of autocorrelation (as done in the context of the MRA).

The fourth aim is to explore how the combined probabilities differ according to which primary indicator is applied to each data set. Thus, we will compare all ways in which the three indicators can be applied to the number of studies to be integrated. If the translation of an indicator into a *p* value functions properly, the results ought to show very similar combined probabilities, regardless of which indicator is used when. A representation of this process and the fourth aim regarding integration of *p* values is provided on Figure 2.

INSERT FIGURE 2 ABOUT HERE

To sum up, the aim of this study is to complement the presentation of the MRA with a test based on both generated and real data in order to obtain evidence on the performance of this proposal. That is, with the results presented here we would like methodologists working in the field of SCD to assess whether they consider the proposal to be statistically justified and sound. At the same time, we would like to make applied researchers acquainted with the MRA, with its strengths and limitations, so that they could decide whether it is useful as additional evidence when assessing intervention effectiveness in different psychological fields and settings.

Method

Data generation

To allow for a comparison in a variety of data patterns with known characteristics, a Monte Carlo study is carried out. The phase lengths were chosen to match the ones present in the real-life data sets detailed below. Short baseline phases are included, with n_A ranging from 2 to 4 measurements, in order to represent less favorable but potentially common situations (e.g., Shadish and Sullivan, 2011, found that almost 50% of the baselines contained less than 5 data points). Complementarily, treatment phase lengths (n_B) were 6, 8, 9, 11, 14, and 22, representing a wider range of situations. Thus, some series were shorter and others longer than the median series length of 20 reported by Shadish and Sullivan (2011).

For data generation we simplified the Huitema and McKean (2000) model to $y_t = \beta_1 \cdot LC_t + \beta_2 \cdot SC_t + \varepsilon_t$. This model allows simulating level and slope change, via the dummy variables LC (equal to 0 for phase A and to 1 for phase B) and SC (equal to 0 for phase A and to 0, 1, …, $n_B - 1$ for phase B). For the level change an abrupt and immediate increase of 1 was programmed (β_1 = 1), whereas for the slope change an increase of .1 per measurement occasion was simulated (β_2 = .1). The ε_t term was specified with a first-order autoregressive model $\varepsilon_t = \varphi_1 \cdot \varepsilon_{t-1} + u_t$ with the following degrees of autocorrelation (φ_1): −.3, 0, .3, and .6. For generating data according to the SMA-based approach, the phase-specific degrees of autocorrelation were used (whenever possible), according to the estimates of autocorrelation for each data series. The random disturbance was specified to follow either an exponential, normal, or a uniform distribution with a zero mean and unitary standard deviation. For each combination of series length and type of effect present 10,000 data sets were generated.

The model presented by Huitema and McKean (2000) has been frequently used in single-case methodological studies (e.g., Beretvas & Chung, 2008b; Ugille, Moeyaert, Beretvas, Ferron, & Van den Noortgate, 2012), given that it considers several aspects such as the serial relation of data and the possibility of different types of effect, as well as trend. However, this model represents continuous data, which may not always be available in single-case studies in which count data (e.g., number of behaviors) are used.

Real-life data selection

Two studies (Schlosser & Blischak, 2004; Taylor & Weems[1], 2011) using multiple baseline designs (MBD) across participants were included here. MBD across participants have three characteristics that make them suitable for quantitative integrations: there are several data series available, these series are independent (Jones & Fiske, 1953), and the same intervention and goals is present for each of them (Rosnow & Rosenthal, 2009). However, note that, in contrast with the Taylor & Weems data, not all the data sets are independent in the Schlosser and Blischak (2004) study as four participants are studied in three different conditions each. The data sets were chosen to represent different psychological fields: the Schlosser and Blischak (2004) study deals with autism aiming to increase spelling performance, whereas the Taylor and Weems (2011) study is focused on reducing posttraumatic stress disorder symptoms. Apart from substantive differences, the studies are also distinct in terms of the baseline measurements, with the former presenting no data variability (all baseline measurements are equal zero). Another reason for including these data sets was the range of treatment phase lengths. Nonetheless, it should be stressed that as the current study includes few data sets, it can only serve as an initial approximation to degree of agreement between MRA's *p* values and the primary authors' conclusions.

Data analysis

[1] We would like to thank Dr. Leslie K. Taylor and Dr. Carl F. Weems for kindly providing their raw data for the analyses presented here.

Each simulated or real-life data set was analyzed via three primary indicators: the Percent of nonoverlaping data (PND; Scruggs, Mastropieri, & Casto, 1987), the Nonoverlap of all pairs (NAP; Parker & Vannest, 2009), and a regression-based procedure (hereinafter, AG) proposed by Allison and Gorman (1993). Hence, two relatively similar procedures and one more classical parametric model were included. The PND was chosen (for instance instead of TauU; Parker, Vannest, Davis, & Sauber, 2011) just for illustration purposes as an extreme case, an example of a nonoverlap procedure with deficiencies such as distortions due to outliers or improving trends in the baseline phase (with the latter being the case for participants Jennifer and Kelly from the Taylor and Weems study). The NAP, also based on data overlap, is not as affected by outliers. For obtaining the outcome of the PND, all treatment phase measurements are compared with the most extreme (in the desired direction) baseline measurement. In contrast with the PND, the NAP takes into account all baseline measurements as it compares each one of them with each treatment phase measurement in order to quantify the proportion of improved data points after the intervention.

The AG is a conceptually different primary indicator representing intervention effectiveness in terms of adjusted R^2. In this regression-based model, the baseline linear trend is initially controlled for. Afterwards, the detrended series are modeled using a dummy variable representing the change in phase and a variable representing slope change (i.e., the LC and SC, respectively, from the data generation model presented above). Therefore, the R^2 is the amount of variability in the detrended data accounted for by the level change and the slope change variables and, thus, it is a quantification of intervention effectiveness.

After applying the primary indicators, a *p* value is assigned to each of them following the SMA or the MRA[2] reference values specific for the phase lengths and for the indicator. Three informal comparisons are carried out: a) for the simulated data, the distribution of MRA's *p* values was compared across primary indicators and considering whether an effect is present or not; b) for the real-life data, SMA's and MRA's *p* values assigned to the different primary indicators were compared; and c) the *p* values were related to the original authors' effectiveness assessment. These steps are represented on Figure 1.

For combining probabilities (i.e., integrating the data series included in each MBD) the weighted average of the *p* values and the binomial test are used. The aim was to answer the following two questions: a) are the combined probabilities similar if all studies are analyzed using PND vs. the NAP vs. the AG? and b) are the combined probabilities similar regardless of which indicator is used for analyzing each study? For answering the second question, we studied all possible distinct ways in which the indicators could be applied to the different data sets (i.e., all possible variations with repetition). In order to illustrate the concept of variations with repetition as used here, consider the following example. Suppose we were studying only two primary indicators (PND and NAP) and there were only two data sets whose results are intended to be integrated (data set 1 and data set 2). In such a situation there are 4 possible distinct ways in which the indicators could be applied to the different data sets: 1) PND applied to both data sets 1 and 2; 2) NAP applied to both data sets 1 and 2; 3) PND applied to data set 1 and NAP applied to data set 2; and 4) NAP applied to data set 1 and PND applied to data set 2. Thus, there are 4 variations with repetition, obtained after

[2] Note that for the current methodological study identifying the reference values for the MRA required constructing the sampling distributions for each primary indicator and phase length. Nonetheless, if the MRA proves to be a valuable approach, tables with reference values would be made available to applied researchers. Such tables are not provided here, given that they are not the primary focus of this study.

elevating the amount of indicators to a power equal to the amount of data sets, 2^2. Using the same logic, there are 3^6 ways in which the three indicators can be applied to the 6 Taylor and Weems data sets, that is, a total of 729 variations with repetition (2^6 = 64 including only the two nonoverlap procedures). For the Schlosser and Blischak data sets there are 3^{12} = 531,441 variations with repetition (2^{12} = 4,096 including only the nonoverlap procedures). For each of these variations with repetition we computed the combined probability and studied the variability in distribution of combined probability values. The desirable result is low heterogeneity indicating similarity regardless of which indicator is computed to each data set. Such a result, however, can be affected by any limitations associated with the performance of the procedures (e.g., typical values, sensitivity to outliers or baseline trends). These steps are represented on Figure 2.

Results

First aim: Probabilities assigned to generated data

When there is no intervention effect simulated the proportion of *p* values is equal to the *p* values themselves (e.g., approximately 1% of the *p* values are .01 or less, 5% are .05 or less) when the MRA for zero autocorrelation is used on independent data series (see Table 1). Note that for the nonoverlap procedures, in some cases the same reference values correspond to different *p* values due to the discreteness of the primary indicator values. However, the *p* values generally become conservative when the MRA value for φ = .6 is used for assigning probabilities to the outcomes.

INSERT TABLE 1 ABOUT HERE

Intervention effects were simulated in independent data series, but the conservative MRA *p* values (in this case based on $\varphi = .6$) were used, given that they are recommended when the researcher has no solid evidence regarding the expected autocorrelation in data. The *p* values assigned to the NAP and the PND (Table 2) are similar and they indicate that the effect simulated is detected, since the proportion of smaller probabilities is greater than expected by chance. For the AG, the proportions of small *p* values are lower (i.e., the procedure is less powerful) and they do not match the results for the two nonoverlap procedures.

INSERT TABLE 2 ABOUT HERE

First aim: Probabilities assigned to the real-life data

Assigning MRA *p* values to the primary indicators was done comparing the MRA reference values for *p* = .01, .05, .10, .20, .30, and .50 (*p* = 1 in case the outcome is smaller than the reference value for *p* = .50). The results of applying the primary indicators and the probabilities assigned are presented in Table 3. Comparing among indicators, the *p* values assigned to the NAP and the PND are practically identical, whereas for the AG the results diverge markedly. Also note that high indicator values (e.g., for the NAP) are not always related to small *p* values, given that such values are expected in some cases in absence of an intervention effect.

INSERT TABLE 3 ABOUT HERE

Second aim: Probabilities and primary authors' judgments

For the Schlosser and Blischak data the *p* values assigned to the NAP and the PND are ≤ .20, which is well-aligned with the authors' claim that the intervention was effective for all children regarding acquisition. For the data sets corresponding to participants referred to as Jennifer, Elizabeth, and Michael in the Taylor and Weems study, all *p* values assigned to the NAP and the PND are ≤ .10. These are the three children pointed out by Taylor and Weems (2011) as displaying more evident results.

Third aim: MRA vs. SMA comparison

For assigning probabilities via SMA, it was necessary to estimate the autocorrelation in each phase in each data set. We chose the autocorrelation estimators, r_1^+ (Huitema & McKean, 1991) for phases shorter than 10 data points (i.e., for all baselines) and the 1-recursive estimator for longer phases, considering bias and mean square error (Solanas, Manolov, & Sierra, 2010). For the Schlosser and Blischak data and for the participant called Jennifer by Taylor and Weems, it was not possible to estimate the autocorrelation for the baseline, due to the lack of data variability in that phase. Thus, we chose to sample both phases' data from a population with the same autocorrelation as estimated from the phase B data.

The MRA was based on autocorrelations equal to .60, which made this approach somewhat more conservative than the SMA-based approach for the Taylor and Weems data sets, given that autocorrelation estimates range from −.20 to .32 for baseline data and there are only two estimates greater than .60 for the treatment data. However, due to the high positive autocorrelation estimated for the Schlosser and Blischak treatment data (ranging from .12 to .90 with a median of .71), the differences between MRA and the SMA-based approach were minimized and, in some cases, inverted.

Fourth aim: Integrating the real-life data sets using the weighted mean

Before commenting the numerical results it has to be stressed that the twelve data sets from Schlosser and Blischak are not all of them independent. This is so given that the same four participants are subjected to three different intervention conditions with the same behavior (i.e., percentage of words spelled correctly) being registered in each of these conditions. This raises the question of whether they can be combined as being independent.

According to the internal validity discussion, certain data sets (e.g., participants John, Sarah, and Kelly in the Taylor and Weems study), probably should not be included in the quantitative integration given that it is not clear whether the expected data pattern is followed. Nonetheless, we used these data sets here to illustrate how the numerical values reflect, albeit insufficiently, the lack of clear effect and a clearly demonstrated functional relationship.

When the same indicator is applied to all data sets, the weighted average probabilities for the Taylor and Weems data are .30 for the PND, .32 for the NAP and .65 for the AG. For the Schlosser and Blischak data these values are .09, .10, and .86, respectively. The lower weighted means for the Schlosser and Blischak data agree with the fact for the Taylor and Weems data effectiveness was not equally evident for all participants. As expected, considering the *p* values assigned to the AG, the weighted averages suggest that in general the interventions do not seem to be successful. These results are related to the findings of low power for the AG (Table 2).

The second specific question related to the fourth aim was how the combined probabilities vary according to the indicator used for each of the data sets. The two boxplots on the left of Figure 3 represent the distribution of the weighted averages for all variations with repetition for the Taylor and Weems data. When only the two nonoverlap procedures are used (leftmost boxplot), the combined *p* values are very similar, ranging from .26 to .38, with a median value of .32. When the AG is also used, the variability is greatly increased, with a minimum of .12 and a maximum of .86. The results for the Schlosser and Blischak data (the two boxplots to the right of Figure 3) are alike and illustrate even more clearly how the use of the AG (and the usually greater probabilities assigned to it) leads to an increased variability in the weighted average probability.

INSERT FIGURE 3 ABOUT HERE

Fourth aim: Integrating the real-life data sets using the binomial test

When the same indicator is applied to all data sets and the binomial test is used for integrating, with "successes" being defined as $p \leq .05$, the (binomial) probabilities of getting as many successes by change are a) for the Taylor and Weems 6 data sets: .0328 for the PND (2 successes) and .2649 for the NAP and the AG (1 success); and b) for the Schlosser and Blischak 12 data sets: $4.95 \cdot 10^{-7}$ for the PND and the NAP (7 successes) and 1 for the AG (0 successes). The lower probabilities assigned to the nonoverlap indices in the latter case once again concur with the idea that the intervention effect is clearer in the Schlosser and Blischak study. Note that a single difference between NAP and PND for the Taylor and Weems data is associated with a considerable change in the binomial probability given the few data sets being integrated.

When considering all possible variations with repetition (Figure 4), both for the Taylor and Weems and the Schlosser and Blischak data, the presence of AG as a primary indicator leads to a broader range of results. However, the proportion of binomial probabilities at or below .05 remains very similar: 40-50% in the former case and 95% in the latter case, in which the intervention has more evident effects.

INSERT FIGURE 4 ABOUT HERE

Discussion

Regarding the first aim of the study, when the three primary indicators were applied to generated data, the *p* values assigned to them via the MRA are close to the ones expected by chance when there is no intervention effect simulated. Therefore, the conversion of the primary indicators into probabilities represents properly the likelihood of obtaining extreme results under the null hypothesis of no intervention effects. Complementarily, although only for the two nonoverlap indices, the proportion of smaller *p* values increases when there is a level or a slope change simulated. The low probability of observing an extreme indicator value under the null hypothesis is well-aligned with the fact that the null hypothesis is actually false, given that an effect is present (simulated) in the data. Hence, the *p* values do reflect the features of the data, even after an intermediate step in which data are summarized by a primary measure (i.e., here a nonoverlap index), which is afterwards converted into probabilities.

Concerning the second aim, both the individual and the combined *p* values assigned via the MRA agree well with the original authors conclusions for the data included here That is, lower probabilities are obtained in the cases in which a greater effect is considered to be present. Note that these results are applicable only to the two nononverlap procedures, but not to the AG. Thus, this represents an initial, although very limited, indication of the validity of the *p* values as additional evidence for the degree of intervention of effectiveness.

As regards the third aim, the difference between the conservative MRA and the SMA-based approach was not found to be consistent in favor of either of the two. Specifically, the more data-specific SMA-based approach does not systematically lead to an increased detection of existing intervention effects. In fact, the SMA is more time-consuming given that it requires choosing an appropriate estimator according to series length and simulating serially related data in order to construct a suitable sampling distribution for obtaining a *p* value.

Regarding the fourth aim, for the two nonoverlap procedures, it was shown that the combined probabilities differ only slightly regardless of which of the two indicators is applied to each data set. Thus, when the PND and the NAP are translated into *p* values and afterwards combined, their similarity across all variations with repetition reflects well the fact that the same data sets have been analyzed and integrated. Once again, introducing the conservative AG makes the variability of the *p* values greater.

Methodological implications

With the current study it has been shown that *p* values can be used as an indication of whether the outcome obtained is likely to have arisen in absence of an intervention effect and also as evidence on the size of the effect observed. Thus probabilities may constitute an aid in labeling an effect as "large" or "small", although such an interpretation is necessarily subjected to the practical importance of the results and can also be guided via visual analysis (e.g., Petersen-Brown et al., 2012).

The translation of different indicators (i.e., different metrics) into *p* values can be carried out following the MRA, an approach that provides reference values for deciding whether an outcome should be assigned, say, a *p* value of .20 or .30. This approach has been shown (although more evidence in this direction is needed) not to differ systematically from the SMA which supposedly matches the data at hand more closely. Furthermore, the advantage of the MRA is that researchers would have to compare their outcomes to tabulated reference values instead of carrying out simulations in order to construct an appropriate sampling distribution. Moreover, with the MRA the decisions on which autocorrelation estimator to use and what to do when it is not possible to estimate the autocorrelation are avoided.

Following the MRA the issue of serial dependence is explicitly tackled assuming that moderately high positive autocorrelation may be present in the data. Another issue that needs consideration when integrating SCD data is baseline trend (Van de Noortgate and Onghena, 2003). The reference values obtained via the MRA can also be applied to data with trend, if for such data trend is removed before computing the primary indicator either via ordinary least squares regression or through first-order or second-order differencing (for linear and quadratic trend, respectively).

Finally, it has to be mentioned that the *p* values assigned via the MRA can be sensitive to the performance of the primary indicator: given that the AG usually yields high adjusted R^2 values (Manolov & Solanas, 2008; Parker & Brossart, 2003) the translation via the MRA results in more conservative *p* values which may lead to underestimating the size of the effect observed. Nonetheless, the influence of the deficient performance of the AG on the combined probabilities was shown to be attenuated when the intervention effect is strong as in the Schlosser and Blischak (2004) study.

It has to be mentioned that the application of the MRA to primary indicators which are expressed in the measurement units of the target behavior (e.g., the Slope and level change procedure; Solanas, Manolov, & Onghena, 2010) is more problematic, given that random variability cannot be uniquely specified in order to construct the sampling distributions representing the conditions without intervention effect. On a positive note, the reported deficient performance of the PND with respect to the NAP (Manolov & Solanas, 2008; Manolov, Solanas, Sierra, & Evans, 2011; Parker & Vannest, 2009) was attenuated after translating these primary indicators into *p* values. This is probably due to the fact that PND's sensitiveness to autocorrelation is controlled for using reference values that take autocorrelation into account.

Study limitations and future research

It should be highlighted that only two studies with real data and only three primary indicators of effect size were included. The results on these specific data sets cannot be assumed to be representative for all possible behavioral data, but complementing them with the simulation does offer more solid evidence on the performance of the MRA. Additionally, it should be taken into account that the MRA itself has been so far developed considering continuous rather than count data and that linear or quadratic trends need to be removed from the data before using the probability values assigned via the MRA.

Future studies need to focus on a more extensive comparison between the result of quantitative integrations of the same data sets (e.g., using HLM, simple or weighted averages, and *p* values assigned via the MRA) and the conclusions that practitioners or researchers reach on the basis of visual analysis and client knowledge. On one hand, in search of the optimal integration method, it is important to make explicit the main strengths and limitations of the existing alternatives. On the other hand, the actual use of these methods is subjected to the acceptability by researchers, once their strengths and limitations are known.

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.

APA Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist, 61*, 271-285.

Becker, B. J. (1987). Applying tests of combined significance in meta-analysis. *Psychological Bulletin, 102*, 164-171.

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

Beretvas, S. N., & Chung, H. (2008b). An evaluation of modified R^2-change effect size indices for single-subject experimental designs. *Evidence-Based Communication Assessment and Intervention, 2*, 120-128.

Blampied, N. M. (2000). Single-case research designs: A neglected alternative. *American Psychologist*, *55*, 960.

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.

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

Bulté, I., & Onghena, P. (2012). When the truth hits you between the eyes: A software tool for the visual analysis of single-case experimental data. *Methodology, 8*, 104-114.

Bulté, I., Onghena, P., Salmaso, L., & Solmi, F. (2010, June). *Single-case experiments: A permutation solution to alternation designs*. Scientific meeting of the Italian statistical society: Satellite conference, Padua, Italy.

Burns, M. K., (2012). Meta-analysis of single-case design research: Introduction to the special issue. *Journal of Behavioral Education, 21*, 175-184.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49,* 997-1003.

Darlington, R. B., & Hayes, A. F. (2000). Combining independent *p* values: Extensions of the Stouffer and binomial methods. *Psychological Methods, 5,* 496-515.

Edgington, E. S. (1972). An additive method for combining probability values from independent experiments. *Journal of Psychology, 80,* 351-363.

Edgington, E. S., & Onghena, P. (2007). *Randomization tests* (4th ed.). London: Chapman & Hall.

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

Ferron, J., & Jones, P. K. (2006). Tests for the visual analysis of response-guided multiple-baseline data. *Journal of Experimental Education, 75*, 66-81.

Gedo, P. M. (2000). Single case studies in psychotherapy research. *Psychoanalytic Psychology, 16,* 274-280.

Hedges, L. V., Cooper, H., & Bushman, B. J. (1992). Testing the null hypothesis in meta-analysis: A comparison of combined probability and confidence interval procedures. *Psychological Bulletin, 111*, 188-194.

Holden, G., Bearison, D. J., Rode, D. C., Rosenberg, G., & Fishman, M. (1999). Evaluating the effects of a virtual environment (STARBRIGHT world) with hospitalized children. *Research on Social Work Practice, 9*, 365-382.

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

Horner, R. H., & Kratochwill, T. R. (2012). Synthesizing single-case research to identify evidence-based practices: Some brief reflections. *Journal of Behavioral Education, 21*, 266-272.

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

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

Jones, L. V., & Fiske, D. W. (1953). Models for testing the significance of combined results. *Psychological Bulletin, 50,* 375-382.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., et al. (2010). Single-case designs technical documentation. Retrieved on February 10, 2012 from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/ pdf/wwc_scd.pdf.

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.

Kuppens, S., Heyvaert, M., Van den Noortgate, W., & Onghena, P. (2011). Sequential meta-analysis of single-case experimental data. *Behavior Research Methods, 43,* 720-729.

Manolov, R., & Solanas, A. (2008). Comparing N = 1 effect size indices in presence of autocorrelation. *Behavior Modification, 32,* 860-875.

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

Manolov, R., Solanas, A., Sierra, V., & Evans, J. J. (2011). Choosing among techniques for quantifying single-case intervention effectiveness. *Behavior Therapy, 42*, 533-545.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241-301.

Onghena, P. (1994). *The power of randomization tests for single-case designs*. Unpublished doctoral dissertation, Katholieke Universiteit Leuven, Belgium.

Owens, C. M., & Ferron. J. M. (2012). Synthesizing single-case studies: A Monte Carlo examination of a three-level meta-analytic model. *Behavior Research Methods, 44*, 795-805.

Parker, R. I., & Brossart, D. F. (2003). Evaluating single-case research data: A comparison of seven statistical methods. *Behavior Therapy, 34*, 189-211.

Parker, R. I., & Vannest, K. J. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy, 40,* 357-367.

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. (2012). 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., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy, 42*. 284-299.

Petersen-Brown, S., Karich, A. C., & Symons, F. J. (2012). Examining estimates of effect using Non-overlap of all pairs in multiple baseline studies of academic intervention. *Journal of Behavioral Education, 21*, 203-216.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85*, 185-193.

Rosnow, R. L., & Rosenthal, R. (2009). Effect sizes: Why, when, and how to use them. *Zeitschrift für Psychologie / Journal of Psycholog*y, 217, 6-14.

Schlosser, R. W., & Blischak, D. M. (2004). Effects of speech and print feedback on spelling by children with autism. *Journal of Speech, Language, and Hearing research, 47*, 848-862.

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

Schneider, N., Goldstein, H., & Parker, R. (2008). Social skills interventions for children with autism: A meta-analytic application of percentage of all non-overlapping data (PAND). *Evidence-Based Communication Assessment and Intervention, 2*, 152-162.

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., & 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.

Tate, R. L., McDonald, S., Perdices, M., Togher, L., Schultz, R., & Savage, S. (2008). Rating the methodological quality of single subject designs and n-of-1 trials: Introducing the Single-Case Experimental Design (SCED) Scale. *Neuropsychological Rehabilitation, 18,* 385-401.

Taylor, L. K., & Weems, C. F. (2011). Cognitive-behavior therapy for disaster-exposed youth with posttraumatic stress: Results from a multiple-baseline examination. *Behavior Therapy, 42*, 349-363.

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., & Onghena, P. (2003). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, & Computers, 35,* 1-10.

Whitlock, M. C. (2005). Combining probability from independent tests: The weighted Z-method is superior to Fisher's approach. *Journal of Evolutionary Biology, 18*, 1368-1373.

Table 1. Simulation comparisons (conditions without treatment effect) based on series lengths appearing in the real data used. The numerical values represent the proportion of *p* values as small as or smaller than the value depicted in the respective column header.

n_A+n_B	Reference for φ	Indicator				*p* value			
			.01	.05	.10	.20	.30	.50	1.00
3+6	0	NAP	.01	.08	.08	.19	.35	.55	1.00
		PND	.01	.05	.05	.24	.42	.42	1.00
		AG	.01	.04	.09	.22	.31	.50	1.00
3+6	.6	NAP	.01	.01	.02	.09	.20	.55	1.00
		PND	.01	.01	.05	.13	.24	.41	1.00
		AG	.01	.02	.04	.12	.19	.38	1.00
2+9	0	NAP	.02	.07	.11	.22	.36	.55	1.00
		PND	.02	.05	.11	.27	.38	.51	1.00
		AG	.01	.05	.10	.21	.31	.51	1.00
2+9	.6	NAP	.02	.02	.03	.16	.21	.54	1.00
		PND	.02	.02	.05	.11	.26	.50	1.00
		AG	.02	.07	.14	.26	.37	.56	1.00
4+11	0	NAP	.01	.06	.08	.17	.32	.52	1.00
		PND	.01	.05	.15	.24	.36	.52	1.00
		AG	.01	.06	.11	.22	.29	.51	1.00
4+11	.6	NAP	.00	.01	.02	.07	.17	.53	1.00
		PND	.00	.00	.03	.10	.24	.53	1.00
		AG	.00	.02	.04	.11	.20	.39	1.00
3+22	0	NAP	.01	.06	.11	.23	.31	.52	1.00
		PND	.02	.06	.13	.20	.30	.43	1.00
		AG	.01	.05	.10	.20	.31	.49	1.00
3+22	.6	NAP	.00	.01	.03	.10	.22	.52	1.00
		PND	.00	.01	.04	.09	.19	.42	1.00
		AG	.02	.08	.13	.24	.33	.53	1.00

Table 2. Simulation comparisons based on series lengths appearing in the real data used. The numerical values represent the proportion of *p* values as small as or smaller than the value depicted in the respective column header. The conditions simulated correspond to change in level ($\beta_1 = 1$) and change in slope ($\beta_2 = .1$).

n_A+n_B	Effect	Indicator	*p* value						
			.01	.05	.10	.20	.30	.50	1.00
3+6	$\beta_1 = 1$	NAP	.13	.13	.21	.45	.67	.93	1.00
		PND	.13	.13	.30	.49	.67	.81	1.00
		AG	.01	.02	.05	.13	.21	.41	1.00
3+6	$\beta_2 = .1$	NAP	.02	.02	.05	.15	.30	.68	1.00
		PND	.02	.02	.08	.19	.34	.54	1.00
		AG	.01	.02	.05	.13	.20	.39	1.00
2+9	$\beta_1 = 1$	NAP	.15	.15	.24	.58	.67	.91	1.00
		PND	.15	.15	.30	.45	.69	.86	1.00
		AG	.02	.07	.13	.25	.36	.56	1.00
2+9	$\beta_2 = .1$	NAP	.05	.05	.08	.29	.37	.72	1.00
		PND	.05	.05	.12	.21	.43	.68	1.00
		AG	.01	.07	.14	.26	.37	.56	1.00
4+11	$\beta_1 = 1$	NAP	.03	.09	.18	.47	.74	.95	1.00
		PND	.03	.09	.18	.38	.61	.89	1.00
		AG	.00	.02	.04	.12	.21	.39	1.00
4+11	$\beta_2 = .1$	NAP	.01	.02	.05	.19	.43	.81	1.00
		PND	.01	.02	.05	.17	.37	.76	1.00
		AG	.00	.02	.04	.11	.20	.39	1.00
3+22	$\beta_1 = 1$	NAP	.02	.19	.34	.60	.79	.95	1.00
		PND	.05	.15	.33	.50	.67	.85	1.00
		AG	.02	.07	.13	.23	.32	.53	1.00
3+22	$\beta_2 = .1$	NAP	.01	.14	.28	.58	.79	.95	1.00
		PND	.03	.12	.30	.51	.71	.89	1.00
		AG	.03	.08	.13	.24	.33	.54	1.00

Table 3. Primary indicator values computed and *p* values assigned following the MRA and the SMA-based approach.

		NAP	NAP	NAP	AG	AG	AG	PND	PND	PND
Data set	n_A+n_B	value	*p* MRA	*p* SMA[a]	value	*p* MRA	*p* SMA[a]	Value	*p* MRA	*p* SMA[a]
John	3+8	83.33	.20	.05	.47	1.00	1.00	75	.20	.05
Michael	3+11	92.42	.10	.05	.77	1.00	1.00	9.91	.05	.01
Kelly	4+11	43.18	1.00	1.00	.98	.05	.01	0	1.00	1.00
Jennifer	2+9	100.00	.01	.01	.79	1.00	1.00	100	.01	.01
Elizabeth	4+11	89.77	.10	.20	.82	.50	.50	72.73	.10	.20
Sarah	3+11	57.58	.50	.50	.93	.30	.30	27.27	.50	.50
Scott speech-print	3+6	91.67	.20	.20	.81	.50	.50	83.33	.10	.10
Scott speech	3+6	91.67	.20	.20	.96	.20	.20	83.33	.10	.10
Scott print	3+6	100.00	.01	.01	.91	.30	.30	100.00	.01	.01
Fred speech-print	3+14	96.43	.05	.05	.79	1.00	1.00	92.85	.05	.05
Fred speech	3+14	96.43	.05	.05	.84	1.00	1.00	92.85	.05	.05
Fred print	3+14	96.43	.05	.10	.84	1.00	1.00	92.85	.05	.05
Justin speech-print	3+22	93.18	.05	.20	.87	1.00	1.00	86.36	.05	.10
Justin speech	3+22	95.45	.05	.05	.72	1.00	1.00	90.91	.05	.05
Justin print	3+22	95.45	.05	.10	.75	1.00	1.00	90.91	.05	.05
Carl speech-print	3+22	84.09	.20	.20	.75	1.00	1.00	68.18	.20	.30
Carl speech	3+22	88.64	.10	.20	.84	1.00	1.00	77.27	.10	.20
Carl print	3+22	81.82	.20	.05	.37	1.00	1.00	63.64	.20	.10

[a] For the SMA-based approach, for the Taylor and Weems (2011) data (represented above the separation line), simulated data with phase-specific autocorrelation were generated, except for Jennifer. For the Schlosser and Blischak (2004) data, below the separation line, the whole series were generated with the autocorrelation as estimated from phase B.

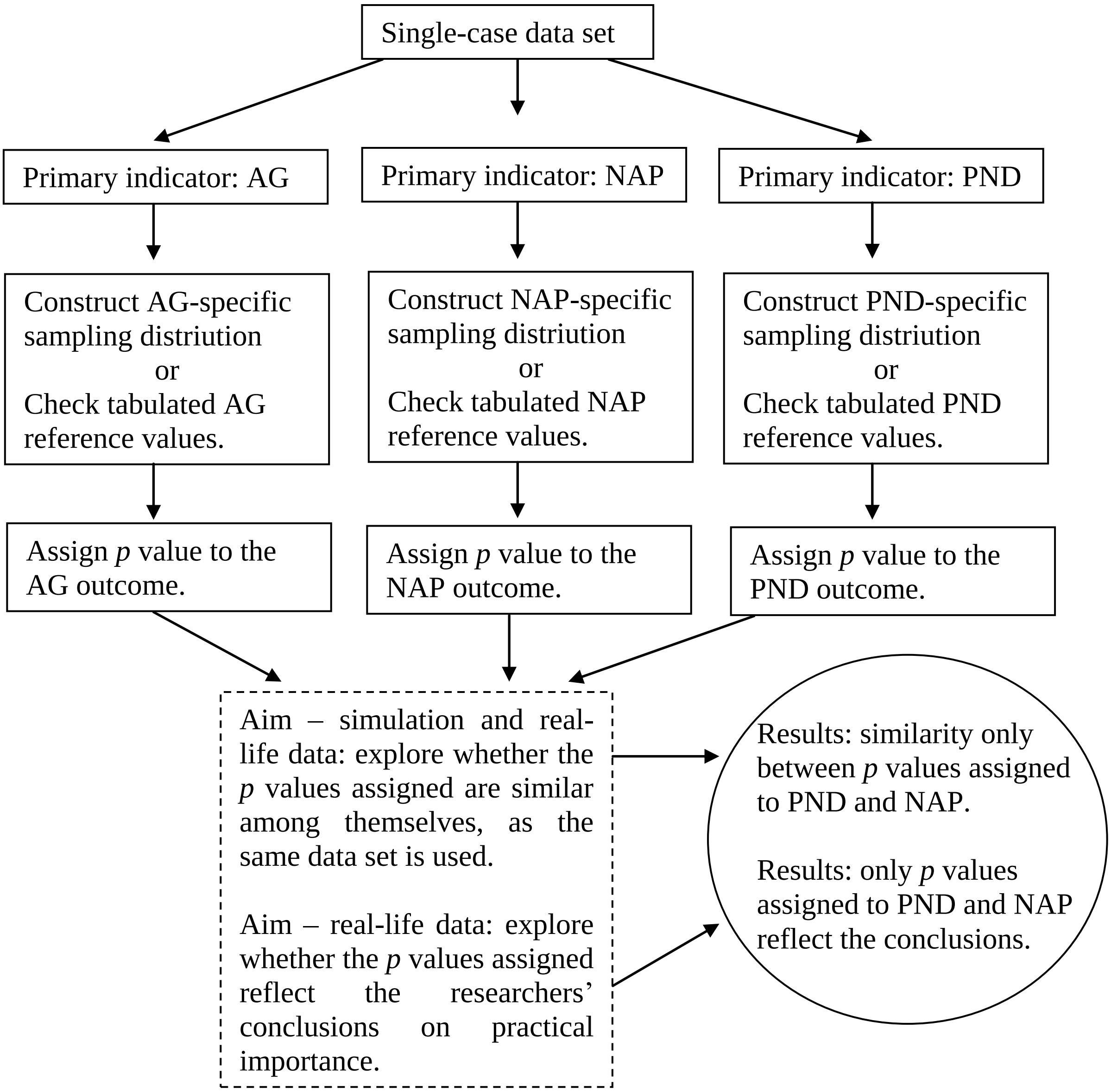
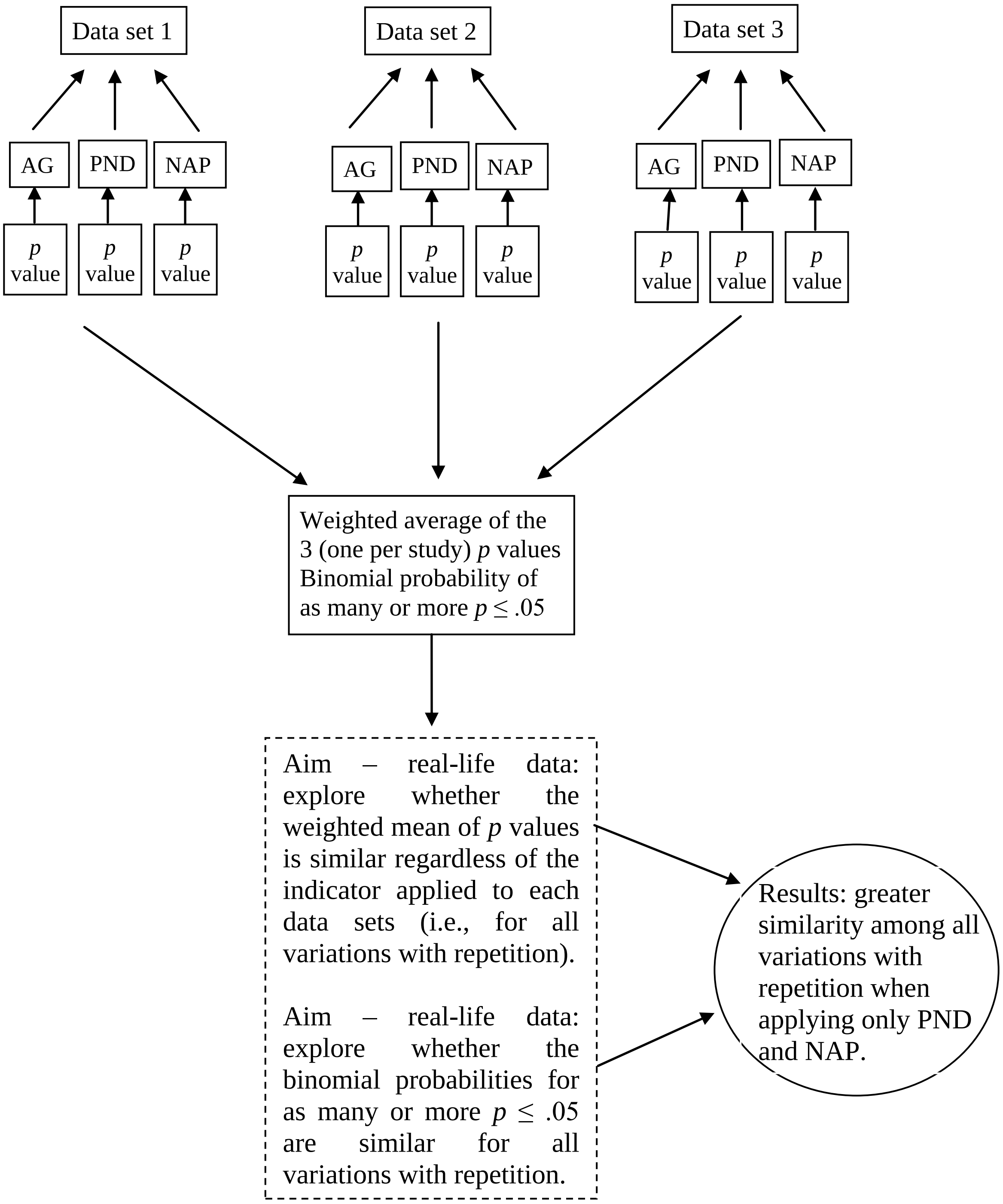

Single-case data set
Primary indicator: AG
Primary indicator: NAP
Primary indicator: PND
Construct AG-specific sampling distriution or Check tabulated AG reference values.
Construct NAP-specific sampling distriution or Check tabulated NAP reference values.
Construct PND-specific sampling distriution or Check tabulated PND reference values.
Assign p value to the AG outcome.
Assign p value to the NAP outcome.
Assign p value to the PND outcome.
Aim – simulation and real-life data: explore whether the p values assigned are similar among themselves, as the same data set is used.
Aim – real-life data: explore whether the p values assigned reflect the researchers' conclusions on practical importance.
Results: similarity only between p values assigned to PND and NAP.
Results: only p values assigned to PND and NAP reflect the conclusions.

Figure 1. Representation of the way *p* values are assigned to the outcome computed through the three primary indicators. The bottom rectangular box includes two of the main aims of this study, whereas the oval box synthesized the corresponding results.



Weighted average of the 3 (one per study) *p* values
Binomial probability of as many or more $p \leq .05$

Aim – real-life data: explore whether the weighted mean of *p* values is similar regardless of the indicator applied to each data sets (i.e., for all variations with repetition).

Aim – real-life data: explore whether the binomial probabilities for as many or more $p \leq .05$ are similar for all variations with repetition.

Results: greater similarity among all variations with repetition when applying only PND and NAP.

Figure 2. Representation of the way *p* values are integrated across studies (in case only three studies were available). Given that any of the three primary indicators (AG, PND, and NAP) can be applied to any study, all possible variations with repetition are considered. The bottom rectangular box includes two of the main aims of this study, whereas the oval box synthesized the corresponding results.

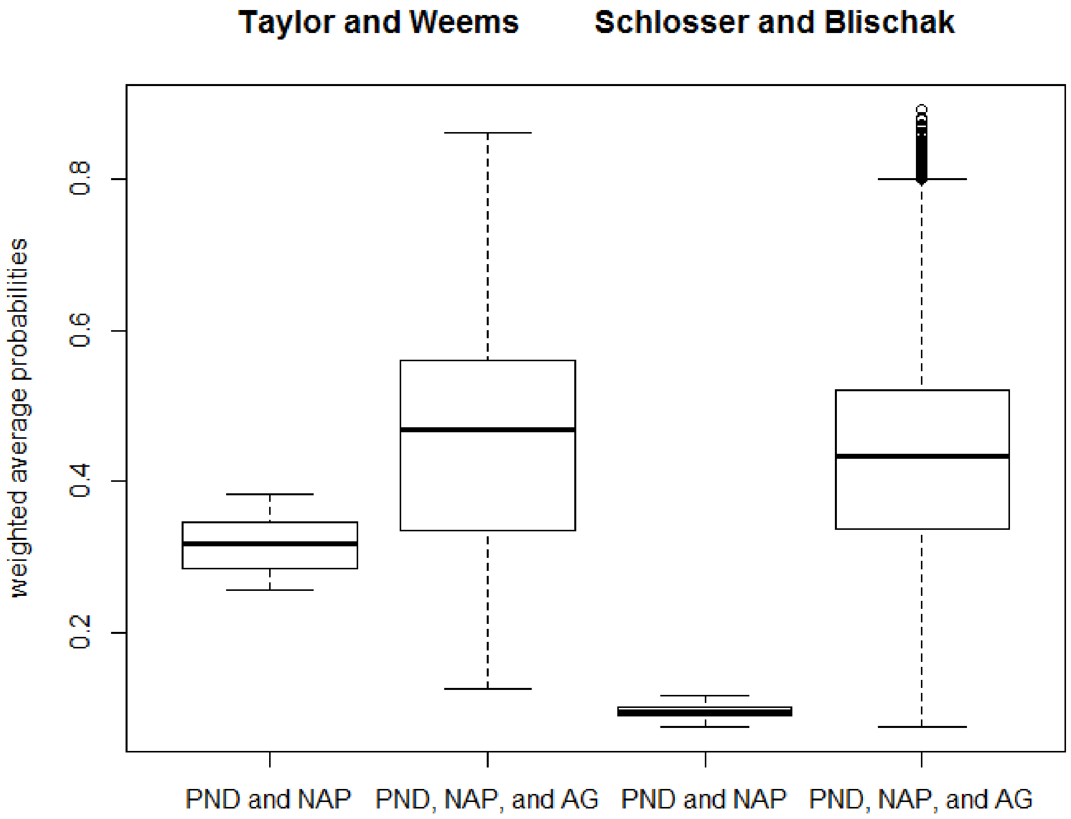


Figure 3. Quantitative integration of *p* values assigned through the MRA for the six Taylor and Weems (2011) and the twelve Schlosser and Blischak (2004) data sets via the weighted mean of probabilities; displaying all possible variations with repetition.

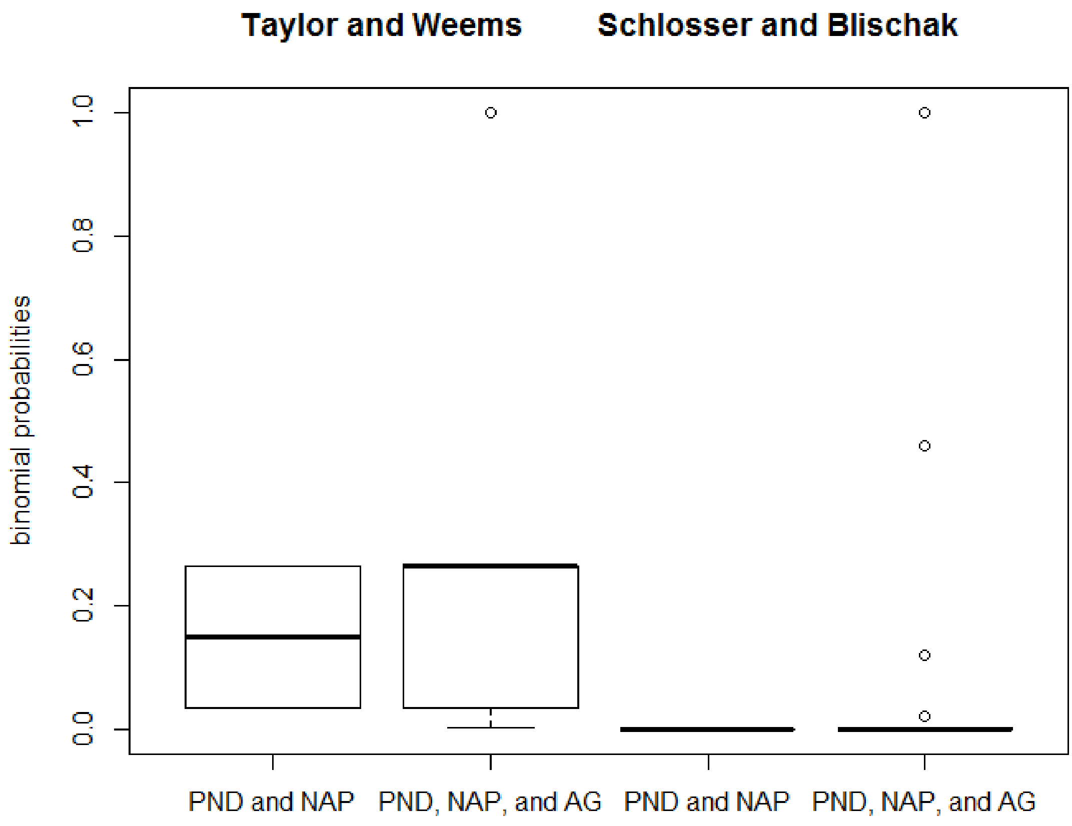


Figure 4. Quantitative integration of *p* values assigned through the MRA for the six Taylor and Weems (2011) and the twelve Schlosser and Blischak (2004) data sets via the binomial test; displaying all possible variations with repetition.