Assigning and combining probabilities in single-case studies

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

There is currently a considerable diversity of quantitative measures available for summarizing the results in single-case studies. Given that the interpretation of some of them is difficult due to the lack of established benchmarks, the current paper proposes an approach for obtaining further numerical evidence on the importance of the results, complementing the substantive criteria, visual analysis, and primary summary measures. This additional evidence consists of obtaining the statistical significance of the outcome when referred to the corresponding sampling distribution. This sampling distribution is formed by the values of the outcomes (expressed as data nonoverlap, R-squared, etc.) in case the intervention is ineffective. The approach proposed here is intended to offer the outcome's probability of being as extreme when there is no treatment effect without the need for some assumptions that cannot be checked with guarantees. Following this approach, researchers would compare their outcomes to reference values rather than constructing the sampling distributions themselves. The integration of single-case studies is problematic, when different metrics are used across primary studies and not all raw data are available. Via the approach for assigning *p* values it is possible to combine the results of similar studies regardless of the primary effect size indicator. The alternatives for combining probabilities are discussed in the context of single-case studies pointing out two potentially useful methods – one based on a weighted average and the other on the binomial test.

Key words: single-case, compound probability, quantitative integration, effect size

The present study deals with a proposal for offering additional numerical evidence on the relevance of single-case study results. This numerical evidence is intended to complement professionals' substantive criteria, visual inspection, and the effect sizes computed and also to serve as a basis for posterior quantitative integrations of individual studies. In the following, the main features of single-case designs are discussed briefly, while for more detail Barlow, Nock, and Hersen (2009) should be consulted. Afterwards, the issues related to the interpretation of the quantitative information available from these designs are commented in order to provide the rationale for the current proposal.

Single-case designs refer to data collection strategies in which measurements are obtained across time from a single unit – an individual or a group (e.g., family, institution) considered as a whole (Onghena & Edgington, 2005). The aim is to assess the effect of a systematic intervention on a behavior of interest while the unit serves as its own control. The utility of single-case designs can be tracked in areas as diverse a clinical psychology (Barlow & Hersen, 1973; Blampied, 2000), special education (Horner et al., 2005; Mastropieri & Scruggs, 1985), neuropsychological rehabilitation (Evans, Emslie, & Wilson, 1998; Perdices & Tate, 2010), and developmental psychology (e.g., Reynhout & Carter, 2006). This type of designs is especially flexible for adjusting to client's context and to professional's objectives (Greenwald, 1976). The longitudinal study of the unit of interest favors an improved understanding of the behavioral process (Fahmie & Hanley, 2008), whereas another positive feature of continuous measurement and assessment is the possibility to interrupt timely any harmful condition (Johnston & Pennypacker, 2009).

The type of control used in idiographic studies differs from nomothetic studies, provided that measurement time is especially relevant (Mace & Kratochwill, 1986).

Particularly important for establishing internal validity (i.e., ruling out alternative explanations of behavioral change) are reversibility –the behavior shifts back to baseline levels when the intervention is withdrawn– and replication of the results observed (Sidman, 1960). In that sense, the simplest design structure AB, an initial evaluation followed by an intervention, is not sufficient for demonstrating experimental control (Wampold & Furlong, 1981). One of the alternatives for converting this pre-experimental AB design to an experimental one is replicating the AB sequence intra-subject (across behaviors or settings) or inter-subjects (across participants) through multiple-baseline designs. This kind of design is especially useful when the behavior studied is not reversible (e.g., learning of skills, internalization of therapist's cognitive schemes). For behaviors that can return to initial (baseline levels), the replication of the AB sequence can be done in the context of an ABAB design, which can also be conceptualized as intra-subject replication. It has the methodological and ethical advantages, as compared to ABA or BAB, of starting with a baseline phase and ending with an intervention. Moreover, for demonstrating reversibility alternating treatment designs can be used when the behavior is susceptible to rapid changes. In these cases it is important that the data pattern match the phases' alternation. Due to the fact that a single unit is studied, the generalization of the results is also subjected to replication; both direct and systematic replications are called for (Sidman, 1960). In that way, specific knowledge can be accumulated taking into account the characteristics of the participant, the behavior of interest, the setting, and the concrete intervention applied. Individual studies integration ought to be based on the similarities among the studies in these features.

Currently there is a great variety of techniques proposed for quantifying the magnitude of effect in single-case data. However, not all of these procedures are accompanied by unquestionable interpretative benchmarks in order to judge the relevance of the results obtained. Among the procedures related to visual analysis, for the Percent of nonoverlapping data there are guidelines offered by the authors (Scruggs & Mastropieri, 1998), but these have not been explained sufficiently. A modification of this index – the Percentage of nonoverlapping corrected data – yields greater values (Manolov & Solanas, 2009) and thus the same criterion cannot be applied. For other nonoverlap indices such as the Improvement rate difference (IRD; Parker, Vannest, & Brown, 2009), the Nonoverlap of all pairs (NAP; Parker & Vannest, 2009) and the Percentage of data points exceeding the median (PEM; Ma, 2006) there are no benchmarks, but confidence intervals can be constructed, as illustrated by Parker and Vannest (2009), to check whether the chance level value is included. For the NAP, a value of .5 (or 50 in terms of percentages) represents the probability that a score drawn at random from a treatment phase will exceed that of a score drawn at random from a baseline phase (Parker & Vannest, 2009) and this is a correct chance level reference. In the case of PEM, the reference value is also 50, which is obtained when half of the intervention phase measurements are above and half below the baseline phase median. For IRD, the authors also point at 50 as the value representing chance level improvement, given that half of the scores overlap (Parker et al., 2009). Nonetheless, a simulation study has shown that the expected value in data with no treatment effect is not 50 and it decreases systematically with the increase of series length (Manolov, Solanas, Sierra, & Evans, 2011).Thus, a better reference is needed for IRD. Another proposal based on data overlap – tau-U – allows assessing the quantification of the magnitude of the effect through *p* values (Parker, Vannest, Davis, & Sauber, 2011).

In the case of regression-based procedures for analyzing single-case data (e.g., Allison & Gorman, 1993; Center, Skiba, & Casey, 1985-1986; White, Rusch, Kazdin, &

Hartmann, 1989), two types of statistical criteria for judging the results are available. On one hand, it is possible to use the statistical significance of the regression coefficients. The validity of these *p* values is doubtful in case of serially dependent residuals due to the fact that ordinary least squares underestimate standard errors (Chaterjee & Price, 1977), which has also been shown in regression models considering the relationship between measurements and time (Manolov, Arnau, Solanas, & Bono, 2010). On the other hand, the R^2 values representing the fit of the model to the data can be interpreted as effect size. However, the appropriateness of Cohen's (1992) guidelines for effect sizes expressed as a coefficient of determination has been questioned in the context of N=1 data where greater effects are usually found (Matyas & Greenwood, 1990; Parker et al., 2005).

Given the importance of replication for external validity, it is critical to integrate the results of similar studies. However, which is the best option for this integration is a question yet to be answered. In that sense, meta-analyses are important for summarizing the empirical support of psychological interventions (Beretvas & Chung, 2008; Jenson, Clark, Kircher, & Kristjansson, 2007; Shadish, Rindskopf, & Hedges, 2008), especially in the context of the recent call for establishing the evidence base for professional practice (APA Presidential Task Force on Evidence-Based Practice, 2006; Wampold, Goodheart, & Levant, 2007) aiming to increase the applicability of research findings to the real context and to improve the way everyday psychological work is done (Kazdin, 2008).

Quantitative integrations need not be restricted to meta-analyses based on effect sizes. An alternative but also classical approach is combining *p* values for obtaining composite significance, also referred to as compound probability (Edgington & Haller, 1984) or probability pooler tests (Darlington & Hayes, 2000). As is the case of meta-analysis, individual studies can be combined using this approach if they meet certain criteria such as studying related problems (Jones & Fiske, 1953), being methodologically sound (Darlington & Hayes, 2000), and when the set of studies selected is either exhaustive or representative (Rosenthal, 1978). Combining probabilities allows also studying moderator variables and reaching more specific conclusions, when the studies are grouped on the basis of relevant dimensions (Darlington & Hayes, 2000).

The present paper focuses on two complementary topics: 1) the additional assessment of intervention effectiveness in an individual single-case study and 2) the quantitative integration of several single-case studies using different numerical indices for representing the magnitude of effect. Regarding the first objective, a proposal for assigning statistical significance values according to the outcome's location in the relevant sampling distribution is presented. For the second purpose, the possibilities for combining probabilities in the context of single-case data are discussed, introducing some modifications to the existing methods. All recommended proposals are illustrated via detailed examples.

Additional evidence for effectiveness in individual studies

The analysis of single-case data ought to be based on the substantive criteria (e.g., knowledge on the discipline, behavior, client), so that the results are interpreted in meaningful terms with practical significance. However, it would also be useful for the researcher to have a statistical criterion complementing the substantive one. Given the lack or inadequacy of benchmarks for most indices, the present section presents an approach for further assessing the relevance of a magnitude of effect value. This approach is based on assigning a statistical significance value to the index computed.

This additional numerical indicator, when combined with substantive evidence such as social validation and self-perceptions, may aid applied researchers to gain more confidence on the relevance of the magnitude of effect value obtained in their studies.

Each magnitude of effect index has its sampling distribution and its expected value in the conditions of no intervention effect, that is, when the null hypothesis is true. The value actually obtained (i.e., the “outcome”) can be located in this sampling distribution. Locating the outcome in the sampling distribution can lead to converting the former into the proportion of values as small as and smaller than it (similar to the idea subjacent to percentiles) or the proportion values as large as or larger than it (i.e., a *p* value). Both types of conversion are equivalent and are related to the likelihood of obtaining a value as large as or larger than the outcome only by chance. Out of the two possibilities, throughout this paper the emphasis will be put here on *p* values in order to maintain continuity with the integrative procedures for combining probabilities discussed in following sections.

It should be noted that there is not a unique sampling distribution, given that it depends on factors such as series length, the assumed distribution of the random fluctuations, and the presence or absence of potential confounding variables (e.g., serial dependence, trend, and heteroscedasticy). That is, in order to make a proper assessment of the relevance of the outcome, the correct reference distribution needs to be identified. In the following, we will propose a procedure, called the “maximal reference approach” (MRA), which can be used to locate an individual study’s outcome in an appropriate sampling distribution. In order to explain the rationale of the MRA we will start reviewing an existing approach – simulation modeling analysis (SMA; Borckardt et al., 2008). The SMA is an analytical technique for assessing intervention effectiveness while controlling for autocorrelation. That is, the SMA evaluates whether the change that has occurred between phases is improbable only due to random variation and serial dependence; random variability is represented by a normal distribution and the serial dependence is estimated from the series. Therefore, the reference in the SMA is a population in which there is no intervention effect and the data are distributed normally with the same autocorrelation as the one estimated. The sampling distribution is constructed on the basis of sampling randomly series with the same length as the original ones from the population described above.

Regarding the way in which the non-effect conditions are represented in the SMA, several limitations ought to be pointed out. Firstly, it is supposed that the random fluctuations follow a normal distribution which may or may not be justified, given the evidence from several psychological fields (Bradley, 1977; Micceri, 1989). Additionally, normal variates are continuous and measurements are often discrete in single-case studies. Secondly, the autocorrelation present in the actually obtained data should be estimated precisely to be used afterwards for data generation. This has proven to be problematic when few measurements are available (Huitema & McKean, 1991). Moreover, different estimators may be appropriate for estimating positive and negative serial dependence, while series length is also important in order to select the proper estimator (Solanas, Manolov, & Sierra, 2010). Finally, it is implicitly assumed that the (actual) data generation process is first-order autoregressive – AR(1), which is the one to be followed when simulating the null hypothesis data. However, it is also possible that the underlining process is a first-order moving average (Harrop & Velicer, 1985), although there are claims for the plausibility and utility of the AR(1) model (Gottman, 1981; Simonton, 1977).

In case these assumptions are not met, the non-effect condition generated would not be an appropriate reference. In order to gain confidence on the validity of the sampling distribution as a reference, it may be necessary to construct not one but several sampling distributions. For instance, instead of assuming that the serial dependence is estimated precisely and using these estimates to generate null hypothesis samples, the researcher may specify several degrees of autocorrelation and thus obtain several sampling distributions. In each of these distributions a different statistical significance value will be associated with the outcome and a conservative (and recommended) option is to use the highest *p* value as a reference. Likewise, instead of assuming normally distributed random fluctuations, several models can be used leading to several sampling distributions and as much *p* values. In the end, the researcher ought to use the highest *p* value of all sampling distributions as an indicator to whether an outcome as large as the obtained one is likely to occur only by chance. This conservative option is especially useful when the researcher has no previous evidence on the degree of serial dependence common in his/her discipline. In case such evidence were available, an approach similar to the SMA (but based on literature review rather than on autocorrelation estimation in a single study) would be appropriate.

The simulations described above are not likely to be easily performed by applied researchers. Therefore, we propose that researchers compare their outcomes to previously identified reference values in the MRA. According to the MRA, the index values corresponding to several key *p* values (e.g., .90, .80, .70, .60, .50, .40, .30, .20, .10, .05, and .01) can be identified for a great variety of conditions (i.e., using diverse sampling distributions) including different data generation processes, various degrees of serial dependence, different random variable distributions, data with and without trend, etc. If these *p* values are tabulated, then a researcher can compare his/her outcome to the reference values from the table, according to the study's phase lengths, in order to know in what range of probability is such an outcome expected at random without the need for simulations.

As illustrations of the MRA consider Tables 1, 2, and 3 in which the sampling distributions correspond to three different procedures, respectively, the Slope and level change (SLC; Solanas, Manolov, & Onghena, 2010), the NAP (Parker & Vannest, 2009), and the Allison and Gorman (1993) model. The tables include five combinations of phase lengths ($n = n_A + n_B$) for the two phases being compared. In Table 1, note that the SLC quantifies es slope and level change in the original measurement unit (e.g., frequency of the behavior of interest) after controlling for linear trend. Thus, it is expected that the indicators are equal to zero, whereas greater values become increasing less likely under the hypothesis of no intervention effect. In the case of the SLC, the fact that the level change indicator is less efficient (Solanas, Manolov, & Onghena, 2010) is reflected in the broader ranges of values farther away from zero for ineffective interventions. In Tables 2 and 3, it can be verified that the reference values for a constant probability become larger for $\varphi_1 > 0$, which is indicative of the NAP and the Allison and Gorman (1993) model being distorted by autocorrelation. For all three procedures represented in the tables, longer series are associated with lower reference values, a result similar to the relationship between degrees of freedom and critical values for parametric tests like *t* or *F*.

INSERT TABLES 1, 2, AND 3 ABOUT HERE

It should be noted that the reference values are presented separately for different degrees of serial dependence (φ_1). In a case a researcher has evidence that in his/her discipline the degree of a serial dependence is say .3, then the outcome of the study ought to be compared to the reference value for φ_1 = .3. However, if there is no such evidence the reference value for φ_1 = .3 is likely to be incorrect for all those cases in which the actual autocorrelation is below or above φ_1 = .3. Given that positive autocorrelation is generally associated with overestimating the treatment effect, the *p* value for φ_1 = .3 would be too liberal if the serial dependence in the series is actually say .6. In order to prevent from assigning an excessively low *p* value based on an assumption (φ_1 = .3) difficult to check with guarantees, a conservative approach can be used – the outcome could be compared to the greatest of the tabulated reference values. However, this approach is actually more conservative only for procedures sensitive to autocorrelation. The similarity among the reference values for the SLC entails that the procedure is rather unaffected and thus the loss of power is only slight. In any case, the reader should not lose sight of the fact that the approach is conservative in the sense that the greatest reference value for different disturbance distributions is used.

For constructing Tables 1, 2, and 3, two-phase data sets were generated using Huitema and McKean's (2000) model $y_t = \beta_0 + \beta_1 \cdot T_t + \beta_2 \cdot LC_t + \beta_3 \cdot SC_t + \varepsilon_t$, which allows simulating data with trend, level change, and slope change. Trend is specified via the dummy variable T_t (with values from *1* to *n* representing linearity) and its slope was defined by the β_1 coefficient, here set to 0. Level change is specified using the dummy variable LC_t taking the value of 0 for the first phase and 1 for the second one, whereas SC_t is used for slope change (set to zero for phase A and to 0, 1, 2, … $n_B - 1$ for phase B). The effect size parameters for both level change (β_2) and slope change (β_3) are set to zero, given that the null hypothesis (i.e., lack of intervention effect) is represented in the data. The average level (β_0) is also zero. Therefore, the data generation model is actually reduced to $y_t = \beta_1 \cdot T_t + \varepsilon_t$. In order to take into account another possible confounding variable – serial dependence – the error term is defined as $\varepsilon_t = \varphi_1 \cdot \varepsilon_{t-1} + u_t$, which represents a first-order autoregressive data generation process, that is, each error term is a function of the previous error term plus a random variable. The degrees of serial dependence (i.e., the degrees of autocorrelation between adjacent values) included were –.3, 0, .3, and .6, as they cover the values expected in real data (Parker, 2006). Nonetheless, the final MRA tables should include a broader range (e.g., from –.9 to .9 in steps of .1), which would also make possible assigning more accurate *p* values to the outcomes in case the typical autocorrelation is known in a specific field. It is also possible to generate data with different degrees of autocorrelation in the two phases and arising from different data generation processes. Random fluctuations (or disturbances) are represented via the u_t term and in this study negative exponential, normal, and uniform distributions with zero mean and unity standard deviation were used. The disturbance is usually represented to be distributed normally in a variety of single-case simulation studies (e.g., Ferron & Ware, 1995; Huitema & McKean, 1991, 2007) and this distribution was also used here. However, there have been several calls regarding the need to consider other types of distributions (Bradley, 1977; Micceri, 1989; Sawilowsky & Blair, 1992). We included another symmetric distribution, the uniform, which is platykurtic compared to the normal. Additionally, a skewed distribution like the negative exponential was also included. In this way, a greater range of possible disturbances was included instead of assuming necessarily the normal distribution. When specifying the distribution of the disturbance, it is also possible to set different values for the scale parameter for each phase, which would lead to generating heteroscedastic data. For instance, if the disturbance in the baseline data population (u_A) follows a normal distribution with mean zero and unity standard deviation, heterogeneity of variance can be simulated specifying the u_B distribution (i.e., disturbance for the treatment data population) as N(0, 2). Heterogeneity of variance was not considered for constructing Tables 1, 2, and 3 but it may have to be specified in case there is evidence of this data feature in the concrete psychological field. The simulation was carried out via programming in Fortran and for generating negative exponential, normal, and uniform disturbance the mathematical statistical NAG libraries (http://www.nag.com/numeric/fl/FLdescription.asp) *nag_rand_neg_exp*, *nag_rand_normal*, and *nag_rand_uniform*, respectively, were used. The data controls suggested by Greenwood and Matyas (1990) and Huitema, McKean, and McKnight (1999) were followed.

In order to construct the sampling distributions (one for each combination of n, φ_1, and u_t), 100,000 samples were drawn and for each sample the techniques (SLC, NAP, the Allison and Gorman model) were computed. This amount of iterations was chosen to guarantee sufficient estimation accuracy (Robey & Barcikowski, 1992). Thus for each condition there were 100,000 values forming the sampling distribution in absence of intervention effect for each indicator for a specific series length, degree of serial dependence, random variable distribution, and trend (zero or nonzero). The values were sorted in ascending order and, for instance, the one that is 90,000th in that order represents percentile 90 and thus a p value of .10. This approach is consistent with directional hypothesis tests in which an increment in the behavior of interest is expected. In case decrease in the behavior is intended, the values in the sampling distribution can be sorted in descending order to obtain the correct statistical significance.

It should be noted that the ranges in Tables 1, 2, and 3 were constructed representing the maximum p values across all three u_t distributions. If the outcome is greater than the appropriate reference, then there is evidence that such a value will not be commonly found in data with no intervention effect, where "commonly" is defined by the user via the nominal alpha (α) chosen *a priori*.

Such tables could serve as additional numerical evidence to assess the relevance of the findings and as a complement to educational, social, clinical, etc. criteria. It should be noted that no clear cut-off points or compulsory benchmarks are suggested here, as Tables 1, 2, and 3 are only an illustration of the information that the MRA could provide. Additionally, the *p* value assigned to the outcome would be a more precise piece of evidence if several references are available, for instance, .90, .80, .70, .60, .50, .40, .30, .20, .10, .05, and .01.

Finally, Tables 1, 2, and 3 illustrate that the MRA can be applied to both regression-based and nonoverlap indices, as well as to other quantitative techniques such as the SLC. Therefore, reference values can also be obtained for other procedures beyond the ones included in the tables if they are commonly used or are considered potentially useful (e.g., Parker et al.'s [2009] IRD, Parker et al.'s [2011] Tau-U, Maggin et al.'s [2011] generalized least squares effect size, McKnight, McKean, and Huitema's [2000] double bootstrap method).

It should be mentioned that in the current paper, the emphasis is put heavily on a comparison between two phases – a baseline (A) and a treatment condition (B). However, this does not necessarily imply focusing solely on a pre-experimental AB design. In that sense, we agree with the comments by Kromrey and Foster-Johnson (1996) and Schlosser, Lee, and Wendt (2008) that when a more complex design structure is used, a comparison and a quantification of behavioral change can be computed for each change in phase. For instance, in an ABAB design, there are three directly meaningful comparisons: between the initial baseline and the first treatment phase, between the first B phase and the withdrawal phase, and between the withdrawal and the re-introduction of the treatment. Accordingly, for a multiple baseline design (regardless of whether it is across participants, behaviors, or settings) the comparison ought to be performed for each baseline (Busse, Kratochwill, & Elliott, 1995). This discussion concurs also with the claim that comparisons should only be made between adjacent phases (Gast & Spriggs, 2009). Therefore, a single study involving either of these more complex design structures would report more than one effect size and there would be one probability assigned to each outcome.

If the researcher is willing to combine the two-phase comparisons within an ABAB or a multiple-baseline design into a single effect it is crucial that such pooling should take place only when for each comparison the difference is in the direction expected. For instance, the three comparisons in the ABAB design are to be combined only when the actual data pattern matches the expected one (e.g., a child with autism whose self-injurious behavior is greater during baseline and withdrawal than during the initial intervention and the reintroduction of the intervention). If experimental control is not demonstrated, that is, if the behavior does not change as a function of the change in phase, it would not be meaningful to combine the two-phase comparisons. More detail will be given when discussing the meta-analytical application of the proposal, given that providing a single indicator out of several comparisons is itself a quantitative integration.

As a limitation of the approach presented here it should be highlighted that it is not readily applicable to alternating treatments designs (ATD), due to the fact that their initial quantitative analysis is also problematic. For instance, if the individual phases are short, numerical summaries would not be reliable – a mean would be unlikely a representative value of the average level and trend could hardly be estimated. In this sense, for ATD visual inspection becomes especially important for assessing whether the behavior changes as expected according to phases' alternation. The same kind of whole pattern assessment is actually useful for all single-case design structures.

Additional evidence for effectiveness in individual studies: Example

As an illustration of the MRA, suppose a hypothetical study in which the insomnia of a patient diagnosed with depression is treated with a selective serotonin reuptake inhibitor (SSRI) antidepressant. The fictitious two-phase data (n_A = 5 and n_B = 5) are graphed in Figure 1 and represent the hours of sleep per day before and after the intervention. In order to quantify the amount of change in the behavior between the two phases suppose that the SLC is applied, yielding a slope change quantification of 0 and a level change quantification of 2.7 (i.e., an average of almost 3 hours of sleep more with the SSRI, after controlling for a potential phase A trend). Also suppose that the author chooses a 5% nominal significance for testing the effect sizes.

INSERT FIGURE 1 ABOUT HERE

An initial step consistent with the SMA would be to estimate the degree of serial dependence in data. In the present case, a modified estimator referred to as r_1^+ by Huitema and McKean (1991) was used, given that in general it shows better performance in terms of mean square error and bias when only five measurements are available (Solanas, Manolov, & Sierra, 2010). The autocorrelation estimated is −.100 for phase A and .165 for phase B. To illustrate the application of the SMA 99,999 samples were simulated with these values for five-measurement phases, a normal random variable, and no intervention effect. In each of these samples the SLC was applied and thus two sampling distributions were obtained, for level and slope change, respectively, each with 100,000 values. These 100,000 values include the 99,999 outcomes from the simulated samples and the outcome for the example data, following Noreen's (1989) comments on statistical validity. The *p* value associated with slope change is approximately .50 (the outcome is the 49,871st greatest value in its set), whereas the *p* value associated with level change is approximately .024 (the outcome is the 97,643rd greatest value in its set). Considering this latter value it can be stated that there is evidence that the difference in the average amount of hours of sleep between phase A and phase B is not likely to be due solely to random fluctuations.

However, we should not lose sight of the fact that these results are conditional on meeting the assumptions for simulating data. In order to obtain a *p* value that is not subjected to such assumptions and does not require simulation, the MRA can be followed, that is, the researcher can use a table with reference values for the quantitative index applied and their associated probability under conditions with lack of effect. In the running example, given that the SLC is used to provide summary measures, the appropriate table with reference values would be Table 1. In order to assign a *p* value to the estimates, a researcher first has to look at the section with the corresponding phase lengths (i.e., $n_A = n_B = 5$). For slope change (SC = 0), it does not matter what the degree of autocorrelation is, since the associated *p* value is always .50. If the researcher has no basis for assuming any specific φ_1, then a conservative approach would be to compare the level change (LC = 2.7) to the highest reference value in each row. 2.7 is smaller than all references for $p = .01$ and so the *p* value assigned should be greater. In fact, 2.7 is greater than the highest reference for $p = .05$, which is 2.55, and thus, the level change quantification can be assigned a *p* value of .05. This means that such a large level change is not likely in case of lack of intervention effect. Hence, following the MRA, the same conclusion can be reached as the one expressed above on the basis of the approach related to the SMA. However, there is a greater degree of confidence that the result is not conditioned on restrictive assumptions, that is, the conclusion is founded on a broader set of conditions.

Quantitative integration of N=1 studies using different metrics

Integrating results from individual single-case studies is apparently straightforward if the raw data from all relevant studies are available. The researcher carrying out the quantitative synthesis would be able to choose the summary index, apply it to all the raw data series, and integrate index's values meta-analytically. In this specific case, the use of different summary measures in the primary studies would not be a problem. In that sense, applied researchers are encouraged to publish their raw data in their scientific reports for further analysis. However, the access to raw data is not always trouble-free. Specifically, a recent single-case designs review (Shadish & Sullivan, 2011) shows that even the use of "un-graph" techniques for retrieving the original data is not automatic or perfect. Moreover, this review suggests that reaching the original authors with data requests may be problematic. An additional issue is that there are no guarantees that all past studies include raw data. The proposal for integration presented here is potentially useful for those cases in which the systematic review is intended to include a) only studies with raw data available; b) only studies for which raw data cannot be obtained, but there is a summary index (i.e., an effect size) computed or a *p* value is available (e.g., via a randomization test); and c) both types of studies.

Meta-analysis using effect sizes vs. combining probabilities

Although Rosenthal (1978) advised using effect sizes *and* composite *p* values, such practice has been deemed redundant, given that both can be used to test the same hypothesis that in every study the average effect is zero (Hedges, Cooper, & Bushman, 1992). Against combining probabilities it has been stated that the meta-analytical pooling gives information about the magnitude of effect (Hedges et al., 1992) and about the homogeneity of effects across studies (Becker, 1987). It should also be taken into account that indicators such as effect sizes, confidence intervals around point estimates (Cumming, 2008; 2011), and the probability of replication p_{rep} (Killeen, 2005) are currently more supported by journals than *p* values (Rosnow & Rosenthal, 2009). This preference for effect sizes instead of *p* values is due to the fact that no single value such as .05 is sanctified for distinguishing between "statistically significant" and "statistically nonsignificant" results and emphasis is put on the strength of association between the variables. However, the interpretation of effect size measures is not trouble-free, given that the numerical values should still be translated into substantive terms (Cortina & Landis, 2011). In that sense, the rigid use of benchmarks, like the ones provided by Cohen (1992), is also flawed (Thompson, 2001) and each value should be interpreted when there is already evidence on the typical effect sizes in each field (Grissom & Kim, 2012).

Combined significance is useful when effect sizes are not reported or there is not an established effect size measure for some data analytic procedure (Rosnow & Rosenthal, 2009) and when the effect size parameters differ across studies (Becker, 1987). A related favorable condition for combining *p* values arises due to the great proliferation of analytical techniques (some of which are expressed in different metrics) and the lack of consensus on which technique to use in order to summarize the results of an N=1 study – a situation which hinders carrying out meta-analyses (Beretvas & Chung, 2008). In the current proposal *p* values are only used to add further evidence and as a common metric for the sake of integration. Therefore, using a composite *p* value does not imply that practical significance is left aside, as it might be argued (Kirk, 1996), nor is focus only placed on the likelihood of the observed differences between phases under the null hypothesis (Hallahan & Rosenthal, 1996). Instead, for each individual study the quantification of the strength of relationship between intervention and behavioral change (e.g., expressed as amount of data overlap or as a quantification of slope and level change) is still available.

Reporting requirements for combining probabilities

Following the MRA, the outcome is not assigned an individual exact *p* value, which would entail having tabulated each and every possible outcome and its location in the sampling distribution. Instead, the outcome is compared to the reference values associated with several key *p* values to explore whether the outcome is more or less likely when there is no intervention effect. In cases in which exact probabilities are not available (e.g., a study only reports that $p < .05$), Edgington (1972a) suggests a conservative solution which allows combining probabilities. This solution consists in taking the upper limit of the probability given by a table as the probability itself, hence if the information available is that $p < .05$, then $p = .05$ should be the value used when integrating studies. Accordingly, when a statistical table is used and the outcome has a probability between .05 and .10, then $p = .10$ should be used. The use of tables was of course more common in the past, but it might be important to be able to also integrate results obtained farther away in time. Following Edgington's (1972a) proposal and the MRA, it is possible to combine the probabilities and carry out a statistical test on the compound *p*, as it will be shown in the following section.

Methods for combining probabilities

In this section several methods will be discussed considering their strengths and potential limitations for application in single-case studies. It is necessary to stress that in order to carry out the statistical tests on the combined probability the individual studies must be independent (Jones & Fiske, 1953), given that lack of independence may lead to inflating Type I errors (Strube, 1985). The lack of statistical dependency between the measurements may be specifically important to be checked in the case of multiple baselines across behaviors and across settings.

The most frequently used method for combining probabilities is the Stouffer method (Hedges et al., 1992; Noble, 2006) consisting in converting the *p* values from the *n* studies to integrate into *Z*s. The test statistic is computed as $\sum_{i=1}^{n} Z_i \Big/ \sqrt{n}$ and is referred to a standard normal distribution. Rosenthal (1978) recommended using this method in most cases. Group designs simulations suggest that this method is more powerful than adding *t*s and than the mean *Z* when less than five studies are being integrated (Strube & Miller, 1986). According to Rice (1990), adding *Z*s is appropriate for testing consensus, with the null hypothesis stating that the average *p* value is greater than or equal to .50. However, it needs to be considered whether this null hypothesis provides interesting information and sufficient evidence on the relevance of the effects in the studies being integrated. Moreover, in the context of N=1 studies, the assumption implicit in the conversion (i.e., that the index used for quantification is distributed normally) is questionable.

The Fisher method, consisting in adding the logarithms of the *p* values, is also considered a standard due to its simplicity and wide acceptance (Mathew, Sinha, & Zhou, 1993). The test statistic is expressed as $-2\sum_{i=1}^{n} \log_e p_i$ and is referred to a chi-square distribution with degrees of freedom equal to twice the number of studies integrated. This method has been claimed to give more information than the binomial test (Jones & Fiske, 1953) and to be potentially more powerful (Rosenthal, 1978). As a limitation, it should be stated that the test favors the alternative hypothesis (small *p* values) due to the use of logarithms (Rice, 1990; Whitlock, 2005). This characteristic is related to the fact that the test does not actually assess whether the group of data sets collectively supports or refutes a common null hypothesis; it rather tests whether there is at least one significant component (Rice, 1990).

It is also possible to add directly *p* values using $\left(\sum_{i=1}^{n} p_i\right)^n \Big/ n!$ to estimate the overall probability of a sum of *p*s as small as the one obtained (Edgington, 1972a). Its author suggests that it is more effective than adding logs, but the greater power also suggested by Rosenthal (1978) was not found in a simulation study (Strube & Miller, 1986). A small amount of studies are recommended to be integrated in order to avoid that the sum of probabilities exceed 1, which would lead to conservative results (Rosenthal, 1978).

The mean of the added *p* values can be tested for significance using $\left(\bar{p} - .5\right)\sqrt{12n}$ as a test statistic and referring to the standard normal distribution (Edgington, 1972b). One of the test’s bases is the assumption that under the null hypothesis each *p* value has the same probability of reaching any value between zero and one and thus their expected value and variance can be modeled by a uniform distribution. The simplicity of the method has been highlighted (Rosenthal, 1978) and it has shown acceptable power with as few as two studies being integrated (Strube & Miller, 1986).

The binomial test has been deemed both quick (Rosenthal, 1978) and simple (Darlington & Hayes, 2000). The main advantage is that it allows using studies which report only information on whether the *p* value was above or below .05 (Rosenthal, 1978). In the case of N=1 studies, this translates into the possibility to follow the MRA and use tables like Tables 1, 2, and 3. Due to this possibility, in the following paragraph the application of the binomial test will be further explained.

The test consists in comparing each *p* value associated with the outcome to the nominal level α predetermined by the researcher. Following the discussion by Darlington and Hayes (2000), all outcome *p* values lower than α are counted as "successes" (versus "failures" in case of greater values) in the terms commonly used when working with the binomial distribution. After that the probability of obtaining as many successes, denoted by *s*, is referred to a binomial distribution with *n* (the number of trials), which is equal to the number of independent individual studies, and *π* (the probability of success in each trial), which is equal to α. The probability of obtaining *s* or more successes can be calculated directly from $\sum_{x=s}^{n}\binom{n}{x}\alpha^{x}\left(1-\alpha\right)^{n-x}$.

Rosenthal (1978) presents "positive results" for the binomial test as those with *p* lower than .50, which is similar to the null hypothesis of the Stouffer method (Rice, 1990). Suppose that the result of one study's statistical test is a *p* value of .49, and in another study the *p* value is .01. Both these results would then be labeled as "positive results" according to this approach and thus would be treated as equivalent. This would imply a great loss of information. However, in the binomial test the nominal significance need not be always specified as α = .50. Actually, one of the main advantages of the binomial test is that the researcher has the possibility to choose any α for distinguishing between "positive" and "negative" results and for testing whether the amount of positive results can reasonably be explained only by chance.

Combining weighted vs. unweighted probabilities

It has been stated that one of the possible sources of invalidity when integrating results is weighting equally studies with different sample sizes (Noble, 2006). However, weighting is not only restricted to sample size (Rosenthal, 1978), as possible weights include other factors that may influence the reliability and validity of the results (Noble, 2006). Weighting gives more information and proved to be more powerful when combining studies with different sample sizes (Whitlock, 2005), but none of the methods commented above – adding logs, *Z*s, or probabilities – includes weighting explicitly. In that sense, it has been claimed that *p* values are already weighted indirectly by sample size, as obtaining the *p* value associated with the test statistic depends both on the effect size and the sample size (Hedges et al., 1992, for the Stouffer method). However, Darlington and Hayes (2000) claimed that in the case of the Stouffer method the weight is actually the square root of *n* and so it is not proportional to the sample size.

In accordance with the validity statements mentioned above, we consider that weighting is important and it is already inherent to the meta-analytical combination of effect sizes. Although a minimal requirement would be to use series length (the single-case equivalent of sample size) as a weight, it should be considered that the ideal weight is the inverse of the standard error of the effect size measure, that is, the Fisher information (Whitlock, 2005). In the current proposal we follow this recommendation and thus propose to use the inverse of the standard error of the summary index used when combining the *p* values assigned to them. Therefore, the weighted combination of *p* values proposed here matches closely the approach followed in the meta-analytical integration of studies' findings (Cumming, 2011; Hedges & Olkin, 1985, but see Hunter & Schmidt, 2004, for a different perspective on weighting) and so there is continuity between the integrative approaches. In the current proposal, the standard error of an index can be estimated as the standard deviation of its sampling distribution. However, the different indices are expressed in different metrics and, thus, the comparison between standard errors cannot be made directly. In order to express the variability of the indices' distributions in the same metric, the coefficient of variation (CV) can be used, provided that it is a nonndimensional measure of dispersion. Hence, the inverse of each index's standard error should not be used as a weighting factor, given that it would lead to weights expressed in different metrics, since the indices themselves are not directly comparable. Instead the inverse of the CV is proposed as for weighting purposes.

The MRA tables would need to include the CV as computed out of the standard deviation and the average value of the index in the sampling distribution constructed to obtain the *p* value associated with the outcome. The weights obtained via the CV could then be used to compute a weighted mean of the *p* values as illustrated later. The weighted mean would probably be a more accurate summary representation of the global effect and can be compared to a predefined reference value, but given that its sampling distribution is unknown no further statistical test can be carried out. The use of Edgington's (1972b) test with the weighted values is not justified, but in any case it would only permit a comparison with *p* = .50, which might not be sufficiently informative for applied researchers.

Combining probabilities for ABAB and multiple-baseline designs

As discussed previously, when the single-case design chosen by the researcher is ABAB or multiple-baseline, the individual comparisons between adjacent phases can be combined in order to obtain one numerical indicator for the whole study. In this section we will illustrate a tentative approach for such a combination. First, we revisit the first example involving an SSRI drug and the hours of sleep as a dependent variable, but now administering the drug according to an ABAB design. If the NAP is applied to each two-phase comparison, it is possible to assign a *p* value to each percentage according to the MRA (i.e., using Table 2). Afterwards, we could combine the three *p* values following the weighted mean approach described above. Thus, we do not just compute the average of the *p* values, we rather given different weights to each comparison according to the variability of the index for that specific comparison, which is related to the amount of data involved. However, there is an additional requirement. We advise against averaging the *p* values if the data pattern is not consistent with the expected one. In the example presented in the upper panel of Figure 2 we see that both after the first introduction of the SSRI and after its reintroduction we have a (tentatively labeled) “large” increase in the hours of sleep. Thus, for the A_1B_1 comparison and the A_2B_2 comparison we would have “large” NAP values and “small” *p* values. However, after the withdrawal of the SSRI we also have an increase (although not as “large”) in the hours of sleep, which is against the idea of a functional relationship between the intervention and the behavior. Hence, for the B_1A_2 comparison we would have a NAP value below what is expected by chance and a “large” *p* value. If we were to compute the weighted mean of the three *p* values, the “small” *p* values would be attenuated (here made larger) by the unexpected increase of hours of sleep during A_2, which would penalize for the unexpected direction of the difference between B_1 and A_2. Nonetheless, we consider that such a “numerical correction” of the *p* value is not sufficient, given that we actually do not know whether the SSRI is the cause of the change in the dependent variable or it is the individual’s natural tendency toward improvement. Therefore, we recommend computing the weighted average only when the actual data pattern resembles the one represented in the lower panel of Figure 2. The same reasoning can be applied to multiple baseline designs. The *p* values for each baseline can be combined using the mean weighted according to the CV, but only in case all the baselines show the expected pattern (i.e., an improvement in the behavior only in the moment in which the intervention is introduced for that specific baseline). In this way, a single a *p* value can be obtained even for studies using ABAB and multiple-baseline designs and used as additional evidence for judging the effect of the intervention. Such a *p* value (formed meta-analytically) can be then incorporated in quantitative integrations of several single-case studies on the same topic via the binomial test. The weighted mean method can also be applied for combining probabilities of several studies; in this case the weight for an ABAB or a multiple-baseline design could be the inverse of the average of the CVs for each two-phase comparison. Finally, this is only a tentative proposal and further developments are required on the optimal way in which to obtain a single indicator per design. It would be especially relevant to discuss how to combine two-phase indicators computed for the same study without assuming independence between the comparisons.

INSERT FIGURE 2 ABOUT HERE

Quantitative integration of N=1 studies using different metrics: Example

Suppose that a researcher wants to perform a quantitative integration of single-case studies on the effectiveness of a specific cognitive-behavior treatment on people diagnosed with major depression. Ten independent but closely related studies have been identified on the topic, none of which offers raw data. In five of these studies the adjusted R^2 of Allison and Gorman’s (1993) regression model is used and in the remaining five the NAP (Parker & Vannest, 2009) is computed and so direct integration is not possible given that the indices do not provide the same information and are not expressed in the same scale.

The ten possible outcomes are provided via a simulated example in the third column of Table 4. For each of the studies, the researcher carrying out the quantitative integration compares the outcome to the corresponding sampling distribution following the MRA, i.e., using references as the ones presented in Tables 2 and 3. In the following we provide examples of how this is done for one study using the regression model and one using the nonoverlap index always assuming, for the sake of simplicity, that the data are not serially related. For the remaining studies the procedure is analogous. For instance, in the third study the adjusted R^2 value is .34. Using Table 3 and considering the phase lengths, the *p* value assigned would be .50 for independent series. That is, the outcome is as large as the reference value for this level of probability and this is the *p* value to be used for integrating results; it logically corresponds to the one in the fourth column of Table 4. Accordingly, for the ninth study, the *p* value assigned to the outcome computed via the NAP can be traced back to the reference table for the NAP – Table 2. Given the phase lengths, it can be seen that a NAP value of 72.00 is actually associated with a *p* value of .05 (the same as in Table 4) for unrelated data. Note that in case the researcher had not had any information on the degree of serial dependence and it was actually positive, the *p* values thus assigned may be too liberal. Checking Tables 2 and 3, it can be verified that the *p* values would have been excessively low for such data. For instance, NAP = 72.00 should have been assigned $p = .20$ instead of .05 in order to avoid overestimating the effect in case φ_1 were as large as .6.

INSERT TABLE 4 ABOUT HERE

Actually, in order to construct Table 4, the *p* value was determined prior to simulating independent series with normal disturbance. The outcome value corresponding to this *p* value was then identified. Therefore, the values in Table 4 represent accurately the association between an outcome and its probability under non-effect conditions, rather than being purely fictitious. The CV values included in the last column of Table 4 were also obtained in the corresponding sampling distributions of independent normal disturbance data.

Weighted mean

Suppose that the researcher wants to check whether the combined probability in the studies on average is lower than .25. This consensus question (Rice, 1990) can also be expressed as checking whether the outcomes on average are greater than 75% of the outcomes expected by chance. Using the MRA *p* values from the fourth column of Table 4, the unweighted average of the *p* values is obtained to be .297. The weighted mean can be computed through $\sum_{i=1}^{n} w_i p_i \Big/ \sum_{i=1}^{n} w_i$, where the w_i elements represent the weights for each *p* value. Given that the weights represent indices' efficiency, in the running example the values in the fifth column of Table 4 are needed. The weights are computed as the inverse of the coefficient of variation, obtaining 1.56, 1.12, .93, 1.14, .77, 3.13, 3.45, 3.13, 3.85, and 2.44 for the ten studies, respectively. The weights add up to 21.52, which is the denominator of the weighted mean. The numerator is obtained adding the products of each *p* value and its weight, a total of 5.21. Thus, the weighted average *p* value is .242; lower than the a *priori* specified reference of .25, indicating that the weighted average outcome is greater than 75% of the outcomes in absence of effect. This weighted *p* value is lower than the unweighted average of .297, as in this case the indices show less relative variation in the studies in which the outcome is less likely under the null hypothesis, as it can be seen from Table 4. Note that both the unweighted and the weighted means are likely to be conservative estimates, as that the *p* values used for computing these means are not exact *p* values, given that the MRA is followed.

Binomial test

In order to carry out the binomial test, consider that the nominal significance chosen is .05 and thus the test will be carried out to obtain the probability of getting as many *p* values lower than .05 only by chance. This reference has been commonly used in statistical decision making and seems useful, because the range of "successful" studies seems sufficiently narrow as to group studies with similar effects. In the running example there are three positive results, that is, there are three outcomes with *p* values of .05 or less according to the MRA (see Table 4). The binomial test is carried out with parameter n = 10 (i.e., ten studies) and a probability of success in each trial $\pi = \alpha = .05$. The test yields a statistical significance value of p = .0115, which denotes the probability of having 3 or more "successes" of ten studies only by chance. This result could be interpreted as positive evidence on the relevance of the cognitive-behavior treatment effect on major depression across studies. In case the focus of the study were then put on these three studies for which the outcome is greater than the critical value, it would be possible to identify certain features responsible for the positive results (Darlington & Hayes, 2000).

Assigning probabilities and publication bias

It has been suggested that when selecting the studies to be integrated, the possibility of a publication bias should be taken into account (Whitlock, 2005), given that if the studies to combine are not a random sample (i.e., if they are not representative), then the validity of the results is threatened (Vevea & Woods, 2005). Rosenthal (1979) referred to nonsignificant results being obtained but not published as the “file drawer problem” and proposed a correction. For the binomial approach, Darlington and Hayes (2000) illustrated how the results obtained with the available studies can be confronted with different amounts of unavailable studies with negative results (i.e., failures). For the running example, it can be verified that even if 7 “failure” studies with *p* values greater than .05 were added to the ten studies available, the probability of 3 or more positive results would be .0503. Thus even an omission of as many studies would not make results completely explicable only by chance, if we consider the result of .0503 as marginally significant and do not draw a strict line at .05, following Cohen’s (1990) and Rosnow and Rosenthal’s (1989) comments.

However, it should be noted that the logic behind this approach for dealing with publication bias has been questioned (Vevea & Woods, 2005). Moreover, it is possible that such a calculus may not be necessary in the context of single-case studies. The file drawer problem essentially considers unpublished studies in which the main (or only) numerical focus is put on *p* values, something not common in N=1 studies. Indeed, including a *p* value in the report of a single-case study (although only as additional indicator associated with the primary effect size measure such as a percentage of nonoverlap) might lead to authors feeling more inclined to submit only “significant” results. However, in the current publication policy endorsed by scientific journals all outcomes are welcome and a greater amount of information is required (e.g., the confidence intervals about effect sizes; Wilkinson & The Task Force on Statistical Inference, 1999). The current proposal is well-aligned with such calls for more complete reporting regardless of statistical significance (Rosenthal & Rubin, 1985). In that sense, statistical tests have already been included associated with effect size measures (e.g., Maggin et al., 2011). As long as the editorial emphasis is put on effect size and the accompanying numerical indicators are supplementary, publication bias is less likely (although still possible) and published studies may contain both low and high magnitude of effect index values, especially considering that it is not clear what is low and what is high for the variety of existing indices (Parker et al., 2005). Moreover, in the extreme case in which current additional evidence (*p* values) is not endorsed by journal editors, publication bias will not take place. However, the information given by the MRA could still be used by meta-analysts who would convert the original metric reported in the articles (e.g., a percentage of nonoverlap) into a *p* value and use it for integrating the studies chosen.

Discussion

The present paper proposes a numerical indicator which is potentially useful for interpreting an effect size index and for combining effect size indices expressed in different metrics in a meta-analytical style. Given that this indicator is *additional*, it should be considered supplementary to the primary assessment tools used by psychologists – their substantive criteria and experience, visual analysis of experimental unit's progression via graphs, and quantitative analysis (e.g., a nonoverlap index, a regression-base procedure). On one hand, whether an intervention effect is relevant is more closely related to clinical, educational, social, etc. criteria. On the other hand, the magnitude of an outcome (e.g., whether it is "large" or "small") can be evaluated considering three different aspects: a) the bounds of the quantitative indicator, for instance, a nonoverlap index ranges from 0 to 100; b) the typical values for that same indicator in the specific psychological field, if such data are available; and c) the location of the outcome in a non-effect sampling distribution (i.e., via the *p* values of the current proposal). In the following, we focus on the strengths and limitations of the proposal.

Strengths and limitations of the additional evidence for effectiveness in individual studies

The main idea of the proposal is to add a numerical indicator to enhance the interpretation of the summary index computed in a single-case study. This additional evidence has been presented in terms of *p* values. *p* values show several limitations, among which we have to mention the likelihood of flawed interpretation and the dichotomous decision making (Cohen, 1994), the lack of attention to Type II errors (Schmidt, 1996), and the unreliability of *p* values across replications (Cumming, 2008). Therefore, in the context of the current proposal *p* values should not be used as the sole indicator, but rather as a complement to the index quantifying behavioral change, that is, effect size measures, regardless whether these are expressed in original or standardized units (Cumming, 2011). Apart from the assessment of the outcome obtained in an individual study, the numerical indicator proposed is potentially useful for comparing the results across studies, even when these results are expressed in different metrics (e.g., based on data overlap, based on regression).

The attainment of the *p* value is similar to SMA (Borckardt et al., 2008) – it is subjected to comparing to a suitable non-effect reference. In order to avoid depending on the assumptions and the problematic autocorrelation estimation, several plausible scenarios are represented in the MRA, which allows researchers to make a more solidly supported decision. Given that the greatest *p* value of all conditions is used, this approach is rather conservative. Nonetheless, with the typical single-case data series lengths statistical procedures tend to have low power (e.g., Ferron & Ware, 1995), so this feature is not unique to the MRA. Moreover, if only strong interventions are assigned “small” *p* values, it is more likely that the effects of these interventions be of practical (and not only statistical) significance, as was suggested also for visual analysis (Parsonson & Baer, 1986). In terms of the efforts required from the researcher, using the MRA involves only making a comparison to an already available maximal reference for the index and phase lengths used. Thus the loss of information (the exact *p* value is not known) is compensated by the ease of use this approach for applied researchers.

Strengths and limitations of the quantitative integration of N=1 studies using different metrics

The MRA allows combining studies whose results are summarized using different indices. This is important, given that the lack of a common effect size index precludes conducting meta-analyses of studies reporting only summary measures but not raw data. A compound probability approach can be followed integrating the *p* values assigned via the MRA. The MRA can actually be followed both when the raw single-case data are available and when there is only a summary measure (e.g., a nonoverlap index) computed.

The present paper focuses specifically on two ways of combining individual studies' *p* values. The first consists in using a weighted mean, where the weight represents the efficiency of the index used to obtain the outcome and gives greater numerical importance to studies in which the index shows less relative variability. An equivalent integration is performed in meta-analysis when, for instance, the standardized mean differences to be combined are weighted according to their standard error. This weighted mean has the drawback of the impossibility to test the composite *p* value statistically. It only allows a comparison with a reference *p* value determined by the researcher prior to carrying out the combination of individual studies' results. Moreover, the CVs required for weighting cannot be computed for certain techniques (e.g., SLC) for which the expected value under the null hypothesis is zero. However, for most of the procedures mentioned here this is not the case, as Table 4 illustrates.

The second highlighted alternative for combining results is a binomial test based on counting the amount of studies in which the *p* values is below a predefined nominal α. This integrative approach makes possible including studies for which only a *p* values is available, for instance, the statistical significance of a regression coefficient representing the change in phase, or the one provided by a randomization test. Primary studies for which only a summary measure is reported can also be included, once this measure is converted into a *p* value following the MRA. Another advantage of this method is that its logic is relatively simple to understand and its use is straightforward. In comparison to the Stouffer method whose test is focused on a *p* value of .50, the binomial test allows specifying any reference value, given that it uses this reference as the probability of a positive result in each independent trial. Additionally, although the test is performed on *p* values, the original information expressed in terms of the magnitude of effect indices commonly used in single-case research is not lost. The binomial test applied jointly with the MRA is straightforward both in terms of conduction and interpretation, which is why we considered that it might be more attractive to applied researchers, who are also willing to test statistically the integrated result.

Future research

As discussed above, the MRA would avoid making decisions about the specific features of the data to be generated, given that a broad variety of conditions is reflected in the reference values, when these become available for the most commonly used and most promising procedures. As a limitation of the current study, it has to be pointed out that the MRA was not illustrated with all available procedures. Instead a nonoverlap, a regression-based procedure, and a slope/level change technique were used in the examples as representative of a broader set of procedures. In case the idea subjacent to the MRA were considered useful by applied researchers, it would be necessary to obtain the reference values for a variety of techniques. These reference values would be made available by means of web-based tables or free software like R.

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.

Barlow, D. H., & Hersen, M. (1973). Single-case experimental designs: Uses in applied clinical research. *Archives of General Psychology, 29,* 319−325.

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

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

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

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.

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

Bradley, J. V. (1977). A common situation conducive to bizarre distribution shapes. *American Statistician, 31*, 147−150.

Busse, R. T., Kratochwill, T. R., & Elliott, S. N. (1995). Meat-analysis for single-case consultation outcomes: Applications to research and practice. *Journal of School Psychology, 33*, 269−285.

Center, B. A., Skiba, R. J., & Casey, A. (1985-1986). A methodology for the quantitative synthesis of intra-subject design research. *The Journal of Special Education, 19*, 387−400.

Chaterjee, S., & Price, B. (1977). *Regression analysis by example.* New York: John Wiley & Sons.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304−1312.

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

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

Cortina, J. M., & Landis, R. S. (2011). The Earth is not round (*p*=.00). *Organizational Research Methods, 14*, 332−349.

Cumming, G. (2008). Replication and *p* intervals: *p* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3*, 286−300.

Cumming, G. (2011). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. London: Routledge.

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. (1972a). An additive method for combining probability values from independent experiments. *Journal of Psychology, 80*, 351−363.

Edgington, E. S. (1972b). A normal curve method for combining probability values from independent experiments. *Journal of Psychology, 82*, 85−89.

Edgington, E. S., & Haller, O. (1984). Combining probabilities from discrete probability distributions. *Educational and Psychological Measurement, 44*, 265−274.

Evans, J., Emslie, H., & Wilson, B. A. (1998). External cueing systems in the rehabilitation of executive impairments of action. *Journal of the International Neuropsychological Society, 4,* 399−408.

Fahmie, T. A., & Hanley, G. P. (2008). Progressing toward data intimacy: A review of within-session data analysis. *Journal of Applied Behavior Analysis, 41*, 319−331.

Ferron, J. M., & Ware, W. (1995). Analyzing single-case data: The power of randomization tests. *The Journal of Experimental Education, 63*, 167−178.

Gast, D. L., & Spriggs, A. D. (2009). Visual analysis of graphic data. In D. L. Gast (Ed.), *Single subject research methodology in behavioral sciences* (pp. 199−233). London: Routledge.

Gottman, J. M. (1981). *Time-series analysis: A comprehensive introduction for social scientists*. Cambridge: Cambridge University Press.

Greenwald, A. G. (1976). Within-subject designs: To use or not to use? *Psychological Bulletin, 8,* 314−320.

Greenwood, K. M., & Matyas, T. A. (1990). Problems with application of interrupted time series analysis for brief single-subject data. *Behavioral Assessment, 12*, 355−370.

Grissom, R. J., & Kim, J. J. (2012) *Effect size for research: Univariate and Multivariate applications* (2nd ed.). London: Routledge.

Hallahan, M., & Rosenthal, R. (1996). Statistical power: Concepts, procedures, and applications. *Behaviour Research and Therapy, 34*, 489−499.

Harrop, J. W., & Velicer, W. F. (1985). A comparison of alternative approaches to the analysis of interrupted time-series. *Multivariate Behavioral Research, 20*, 27−44.

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.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York: Academic Press.

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.

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.

Huitema, B. E., & McKean, J. W. (2007). An improved portmanteau test for autocorrelated errors in interrupted time-series regression models. *Behavior Research Methods, 39*, 343−349.

Huitema, B. E., McKean, J. W., & McKnight, S. (1999). Autocorrelation effects on least-squares intervention analysis of short time series. *Educational and Psychological Measurement, 59*, 767−786.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage Publications.

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.

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

Johnston, J. M., & Pennypacker, H. S., Jr. (2009). *Strategies and tactics of behavioral research* (3rd ed.). New York: Routledge.

Kazdin, A. E. (2008). Evidence-based treatment and practice: New opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *American Psychologist, 63*, 146−159.

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science*, *16,* 345−353.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746−759.

Kromrey, J. D., & Foster-Johnson, L. (1996). Determining the efficacy of intervention: The use of effect sizes for data analysis in single-subject research. *The Journal of Experimental Education, 65,* 73−93.

Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject research: Percentage of data points exceeding the median. *Behavior Modification, 30,* 598−617.

Mace, F. C., & Kratochwill, T. R. (1986). The individual subject in behavior analysis research. In J. Valsiner (Ed.), *The individual subject and scientific psychology* (pp. 153−180). London: Plenum Press.

Maggin, D. M., Swaminathan, H., Rogers, H. J., O'Keefe, B. V., Sugai, G., & Horner, R. H. (2011). A generalized least squares regression approach for computing effect sizes in single-case research Application examples. *Journal of School Psychology, 49,* 301−321.

Manolov, R., Arnau, J., Solanas, A., & Bono, R. (2010). Regression-based techniques for statistical decision making in single-case designs. *Psicothema, 22,* 1026−1032.

Manolov, R., & Solanas, A. (2009). Percentage of nonoverlapping corrected data. *Behavior Research Methods, 41,* 1262−1271.

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

Mastropieri, M. A., & Scruggs, T. E. (1985). Early intervention for socially withdrawn children. *Journal of Special Education, 19*, 429−441.

Mathew, T., Sinha, B. K., & Zhou, L. (1993). Some statistical procedures for combining independent tests. *Journal of the American Statistical Association, 88*, 912−919.

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis for single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23*, 341−351.

McKnight, S. D., McKean, J. W., & Huitema, B. E. (2000). A double bootstrap method to analyze linear models with autoregressive error terms. *Psychological Methods, 3*, 87-101.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156−166.

Noble, J. H. (2006). Meta-analysis: Methods, strengths, weaknesses, and political uses. *Journal of Laboratory and Clinical Medicine, 147*, 7–20.

Noreen, E. W. (1989). Computer-intensive methods for testing hypotheses: An introduction. New York: Wiley.

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

Parker, R. I. (2006). Increased reliability for single-case research results: Is bootstrap the answer? *Behavior Therapy, 37*, 326−338.

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., & 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., & Brown, L. (2009). The improvement rate difference for single-case research. *Exceptional Children, 75*, 135−150.

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.*

Parsonson, B. S., & Baer, D. M. (1986). The graphic analysis of data. In A. Poling & R. W. Fuqua (Eds.)*, Research methods in applied behavior analysis: Issues and advances* (pp. 157−186). New York: Plenum Press.

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

Reynhout, G., & Carter, M. (2006). Social stories for children with disabilities. *Journal of Autism and Developmental Disorders, 36*, 445–469.

Rice, W. R. (1990). A consensus combined *p*-value test and the family-wide significance of component tests. *Biometrics, 46*, 303−308.

Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology, 45,* 283−288.

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

Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. *Psychological Bulletin, 86*, 638−641.

Rosenthal, R., & Rubin, D. B. (1978). Statistical analysis: Summarizing evidence versus establishing facts. *Psychological Bulletin, 97*, 527−529.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276−1284.

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.

Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the *t* test to departures from population normality. *Psychological Bulletin, 111*, 352−360.

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.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 2,* 115–129.

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

Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention, 2,* 188−196.

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.

Sidman, M. (1960). *Tactics of scientific research*. New York: Basic Books.

Simonton, D. K. (1977). Cross-sectional time-series experiments: Some suggested statistical analyses. *Psychological Bulletin, 84,* 489−502.

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.

Strube, M. J. (1985). Combining and comparing significance levels form nonindependent hypothesis tests *Psychological Bulletin, 97*, 334−341.

Strube, M. J., & Miller, R. H. (1986). Comparison of power rates for combined probability procedures: A simulation study. *Psychological Bulletin, 99*, 407−415.

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education, 70*, 80−93.

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods, 10,* 428–443.

Wampold, B. E., & Furlong, M. J. (1981). Randomization tests in single-subject designs: Illustrative examples. *Journal of Psychopathology and Behavioral Assessment, 3,* 329−341.

Wampold, B. E., Goodheart, C., & Levant, R. (2007). Clarification and elaboration on evidence-based practice in psychology. *American Psychologist, 62,* 616–618.

White, D. M., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P. (1989). Applications of meta-analysis in individual subject research. *Behavioral Assessment, 11*, 281−296.

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.

Wilkinson, L., & The Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 694−704.

Table 1. Reference values obtained from 100,000 samples in absence of treatment effect for the slope and level change indicators of the SLC (Solanas et al., 2010). For each degree of serial dependence φ_1 (−.3, 0, .3, and .6), the value represented is the greatest one obtained after exploring separately exponential, normal, and uniform random variable distributions.

		SLC-LC				SLC-SC			
Phase length	*p* value	$\varphi_1 = -.3$	$\varphi_1 = 0$	$\varphi_1 = .3$	$\varphi_1 = .6$	$\varphi_1 = -.3$	$\varphi_1 = 0$	$\varphi_1 = .3$	$\varphi_1 = .6$
n_A=5, n_B=5	.50	0	0	0	0	0	0	0	0
	.30	.87	.78	.76	.75	.26	.27	.30	.34
	.20	1.37	1.25	1.20	1.20	.41	.44	.48	.55
	.10	2.03	1.87	1.81	1.81	.62	.65	.72	.83
	.05	2.55	2.35	2.30	2.33	.80	.83	.91	1.05
	.01	3.62	3.31	3.22	3.26	1.28	1.29	1.35	1.49
n_A=5, n_B=10	.50	.01	.07	.04	.03	0	0	0	0
	.30	.84	.77	.73	.76	.20	.21	.23	.26
	.20	1.33	1.22	1.17	1.22	.33	.34	.375	.41
	.10	1.98	1.81	1.78	1.85	.49	.51	.55	.62
	.05	2.47	2.29	2.25	2.35	.62	.64	.69	.79
	.01	3.49	3.23	3.27	3.48	1.02	1.02	1.06	1.11
n_A=10, n_B=10	.50	.05	.01	0	−.01	0	0	0	0
	.30	.72	.65	.63	.69	.11	.12	.13	.16
	.20	1.14	1.03	1.02	1.11	.18	.19	.22	.26
	.10	1.71	1.56	1.54	1.68	.28	.29	.32	.39
	.05	2.17	1.99	1.98	2.18	.36	.37	.41	.50
	.01	3.13	3.05	3.08	3.41	.57	.57	.60	.72
n_A=15, n_B=15	.50	.05	.01	.01	0	0	0	0	0
	.30	.68	.61	.61	.66	.07	.08	.09	.10
	.20	1.08	.97	.97	1.05	.12	.12	.14	.17
	.10	1.62	1.46	1.46	1.59	.18	.19	.21	.25
	.05	2.05	1.88	1.85	2.07	.23	.24	.26	.32
	.01	3.01	2.91	2.91	3.19	.37	.37	.39	.45
n_A=50, n_B=50	.50	.02	0	.03	.04	0	0	0	0
	.30	.58	.56	.60	.67	.02	.02	.02	.02
	.20	.93	.89	.93	1.07	.04	.04	.04	.04
	.10	1.37	1.34	1.40	1.61	.05	.05	.05	.06
	.05	1.73	1.73	1.78	2.07	.07	.07	.07	.08
	.01	2.42	2.72	2.77	3.00	.11	.11	.11	.11

Table 2. Reference values obtained from 100,000 samples in absence of treatment effect for the NAP (Parker and Vannest, 2009). For each degree of serial dependence φ_1 (−.3, 0, .3, and .6), the value represented is the greatest one obtained after exploring separately exponential, normal, and uniform random variable distributions.

		NAP			
Phase length	*p* value	$\varphi_1 = -.3$	$\varphi_1 = 0$	$\varphi_1 = .3$	$\varphi_1 = .6$
n_A=5, n_B=5	.50	52.00	52.00	52.00	48.00
	.30	60.00	60.00	64.00	72.00
	.20	64.00	68.00	72.00	80.00
	.10	68.00	76.00	84.00	92.00
	.05	76.00	84.00	92.00	100.00
	.01	88.00	92.00	100.00	100.00
n_A=5, n_B=10	.50	50.00	50.00	50.00	50.00
	.30	56.00	58.00	62.00	68.00
	.20	60.00	64.00	70.00	76.00
	.10	66.00	72.00	80.00	88.00
	.05	72.00	78.00	86.00	94.00
	.01	80.00	88.00	96.00	100.00
n_A=10, n_B=10	.50	50.00	50.00	50.00	50.00
	.30	55.00	57.00	60.00	65.00
	.20	59.00	61.00	67.00	72.00
	.10	63.00	67.00	75.00	83.00
	.05	67.00	72.00	80.00	89.00
	.01	75.00	81.00	90.00	97.00
n_A=15, n_B=15	.50	50.22	50.22	50.22	50.22
	.30	54.22	56.00	57.78	61.78
	.20	56.89	59.56	62.67	68.44
	.10	60.44	64.44	68.89	76.89
	.05	63.56	68.00	74.22	84.00
	.01	69.33	75.11	82.67	92.89
n_A=50, n_B=50	.50	50.00	50.16	50.04	50.36
	.30	52.36	53.04	54.32	56.52
	.20	53.80	54.96	57.04	60.24
	.10	55.76	57.44	60.40	65.16
	.05	57.28	59.68	63.48	69.28
	.01	60.32	63.64	69.12	75.84

Table 3. Reference values obtained from 100,000 samples in absence of treatment effect for the Allison and Gorman (1993) model. For each degree of serial dependence φ_1 (−.3, 0, .3, and .6), the value represented is the greatest one obtained after exploring separately exponential, normal, and uniform random variable distributions.

		Allison and Gorman's adjusted R^2			
Phase length	*p* value	$\varphi_1 = -.3$	$\varphi_1 = 0$	$\varphi_1 = .3$	$\varphi_1 = .6$
n_A=5, n_B=5	.50	.20	.34	.50	.65
	.30	.44	.58	.71	.81
	.20	.57	.70	.80	.87
	.10	.72	.81	.87	.92
	.05	.81	.87	.91	.95
	.01	.92	.94	.96	.98
n_A=5, n_B=10	.50	.40	.52	.62	.71
	.30	.63	.73	.80	.85
	.20	.73	.80	.86	.90
	.10	.83	.88	.92	.94
	.05	.89	.92	.95	.96
	.01	.95	.96	.97	.98
n_A=10, n_B=10	.50	.06	.15	.29	.48
	.30	.20	.33	.49	.67
	.20	.29	.44	.59	.75
	.10	.43	.57	.71	.83
	.05	.53	.67	.79	.88
	.01	.70	.80	.87	.93
n_A=15, n_B=15	.50	.03	.10	.20	.38
	.30	.12	.23	.38	.57
	.20	.20	.32	.48	.67
	.10	.30	.45	.61	.77
	.05	.40	.55	.69	.83
	.01	.57	.69	.81	.89
n_A=50, n_B=50	.50	.01	.03	.07	.16
	.30	.04	.08	.15	.28
	.20	.06	.12	.21	.37
	.10	.10	.18	.30	.49
	.05	.14	.24	.38	.58
	.01	.24	.37	.52	.71

Table 4. Fictitious example representing a summary of ten individual studies with the outcome obtained for Allison and Gorman's (1993) adjusted R^2 and Parker and Vannest's (2009) NAP (column three). Column four presents the smallest *p* value for which the outcome is as large as or larger than the Maximal Reference Approach (MRA) reference value. Column five contains the coefficient of variation of the index for the phase lengths (n_A and n_B) represented.

Phase length	Index used	Outcome	MRA *p* value	Index CV
n_A=4, n_B=10	R^2	.83	.30	.64
n_A=5, n_B=7	R^2	.28	.60	.89
n_A=5, n_B=5	R^2	.34	.50	1.08
n_A=8, n_B=14	R^2	.90	.01	.88
n_A=20, n_B=20	R^2	.15	.40	1.30
n_A=7, n_B=7	NAP	55.10	.40	.32
n_A=7, n_B=11	NAP	70.13	.10	.29
n_A=6, n_B=8	NAP	85.42	.01	.32
n_A=10, n_B=10	NAP	72.00	.05	.26
n_A=4, n_B=5	NAP	50.00	.60	.41

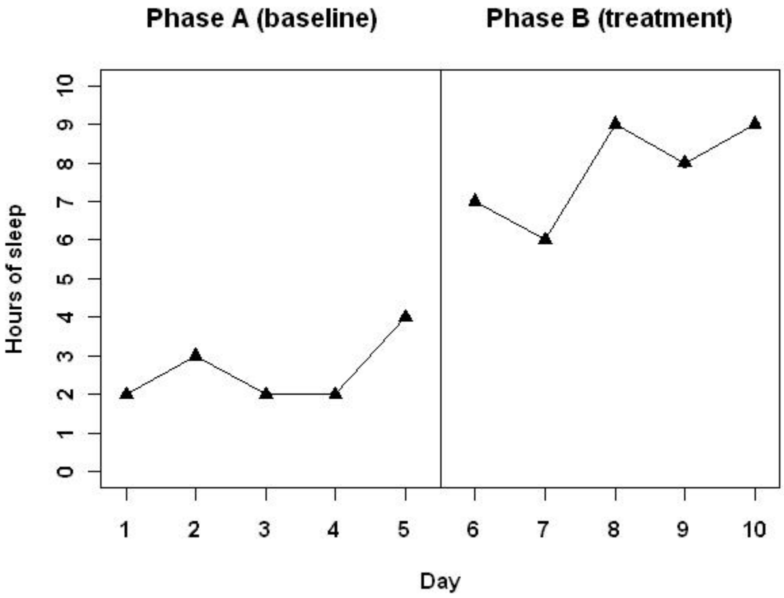

Figure 1. Fictitious single-case data for the insomnia example.

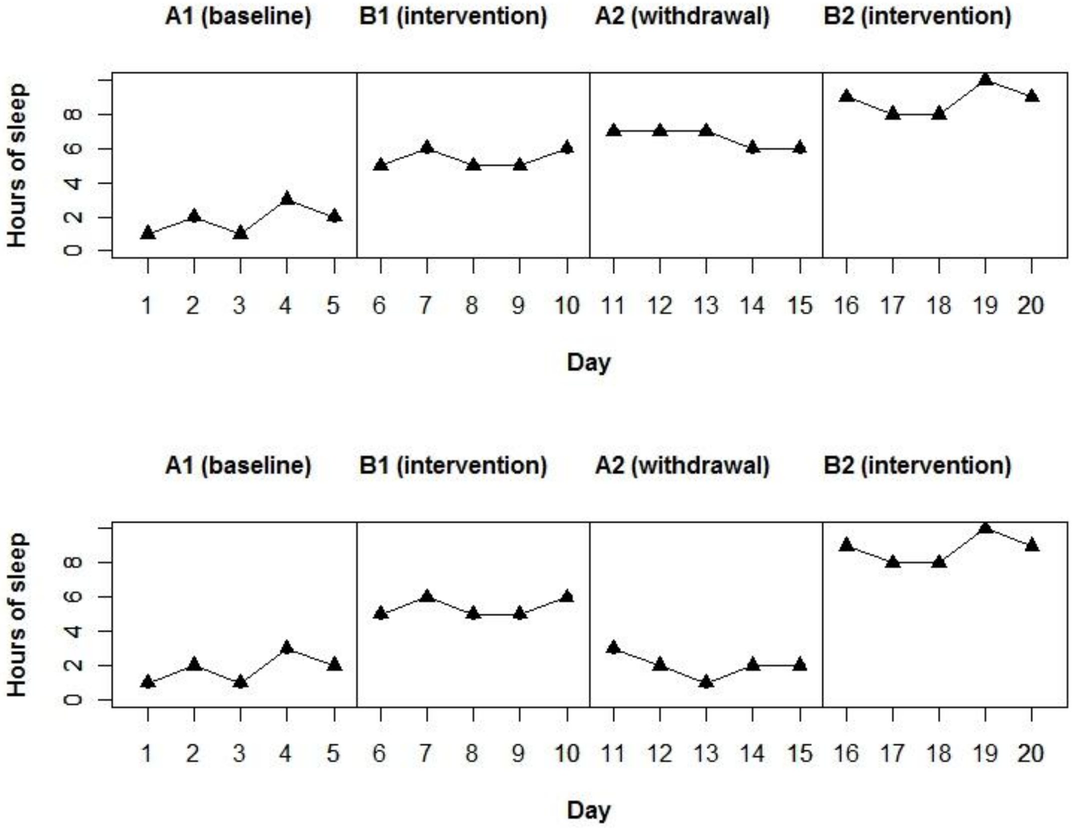

Figure 2. Fictitious single-case data, representing a data pattern illustrative of experimental control (lower panel) and lack of experimental control (upper panel) in the context of an ABAB design.