Assigning and combining probabilities in single-case studies: A second study

Rumen Manolov · Antonio Solanas

Published online: 23 March 2013
© Psychonomic Society, Inc. 2013

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 by comparing the outcome to previously tabulated reference values, and they reflect the likelihood of the results in the case that no intervention effect is present. In the present study, we explored how well different metrics are translated into p values in the context of simulation data. Furthermore, two published multiple-baseline data sets were used to illustrate how well the probabilities might reflect the intervention effectiveness, as assessed by the original authors. Finally, the importance of which primary indicator would be used in each data set to be integrated was explored; two ways of combining probabilities were used: a weighted average and a binomial test. The results indicated that the translation into p values worked 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, when either taken individually or combined, were well aligned with effectiveness for the real-life data. These results suggest that assigning probabilities can be useful for translating the primary measure into the same metric, using these probabilities

as additional evidence of the importance of behavioral change, complementing visual analysis and professionals' judgments.

Keywords Single-case designs · Quantitative integration · Combined probabilities · Effect size

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 ongoing effort for developing further methodological improvements (Kratochwill et al., 2010; Tate et al., 2008). Considering the need to support these interventions empirically (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 section, we review briefly some alternatives for integrating several SCD studies.

R. Manolov · A. Solanas
Department of Behavioral Sciences Methods,
Faculty of Psychology, University of Barcelona,
Barcelona, Spain

R. Manolov · A. Solanas
Institute for Research in Brain, Cognition, and Behavior (IR3C),
University of Barcelona,
Barcelona, Spain

R. 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
e-mail: rrumenov13@ub.edu

Combining SCD studies' results

Meta-analysis Meta-analysis is currently the predominant way of integrating results, as it provides information on the magnitudes of effects (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 (HLMs; Van de Noortgate & Onghena, 2003), by either using raw data or combining standardized effect size measures. HLMs 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-interval 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 had been discussed thoroughly (Becker, 1987; Edgington, 1972; Rosenthal, 1978) before the widespread use of meta-analysis via effect sizes. In the context of SCD, this possibility is potentially useful, given the lack of consensus on what is the most appropriate summary measure (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 into an R package, making it more practical to carry out visual and statistical comparisons and meta-analyses (Bulté & Onghena, 2008, 2012).

We considered that combining probabilities might be a useful alternative to meta-analysis when raw data were not available, and thus researchers could not 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 that was reported by Shadish and Sullivan (2011).

In this context, the maximal reference approach (MRA) was proposed (Manolov & Solanas, 2012), with the aim of aiding 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 an 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 index values that would have been obtained in the absence of an 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 the 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 can compare their outcomes to these reference values and assign a p value accordingly. In this way, carrying out simulations is not required in order to construct a sampling distribution in which to locate the outcome, as would be necessary if the SMA were followed.

The p values assigned to the primary indicators through the MRA can be integrated using the binomial test, as has been described by Darlington and Hayes (2000). A cutoff p value is established for distinguishing between “successes” (i.e., p values lower than or equal to the cutoff) and “failures.” Afterward, the probability of obtaining as many successes out of the number of studies to integrate is computed. A small binomial probability would be indicative that a true positive effect does exist. Another option (Manolov & Solanas, 2012) for integrating would be to just 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 was suggested by Whitlock, 2005).

Three aspects have to be highlighted, with regard to the possibility of assigning 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 for visual criteria or the professional’s judgment about 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 the magnitude of an effect, 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 as being more useful for quantifying treatment effectiveness (Hedges et al., 1992), especially considering the common inappropriate use p values in the context of hypothesis testing or of the assessment of the practical significance or replicability of results (Cohen, 1994; Nickerson, 2000). Specifically, separate quantifications of slope change and level change have been deemed necessary (Beretvas & Chung, 2008b), apart from the importance of considering data overlap (Parker, Vannest, & Davis, 2012).

Second, it is necessary to distinguish between p values as a common metric for integrating studies (as in the MRA) and p values as the primary means 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 being 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 a key outcome. Finally, the p values and their interpretation are subject to an assessment of internal validity, as will be 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 purpose, 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 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 to be accompanied by 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 an SCD data analysis or integration.

Study aims

The present study had four aims. The first objective was 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 effectiveness of an intervention. In the present study, three different primary indicators were applied to the same data, generated via Monte Carlo methods. Thus, adequate performance of the MRA as a

translation method would require that similar p values be assigned across all primary indicators.

The second aim was 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 being assigned to studies in which the effect was judged to be larger. The study based on real data would also focus on whether the combined probabilities reflected the general assessment of the results reported in the original articles. Note that this second aim was intended to be a pilot study into the relationship between p values and researchers' assessments, rather than a large-scale field test. The process for assigning p values and a graphical summary of the first two aims are represented in Fig. 1.

The third aim was to compare the p values obtained through the MRA- and SMA-based approaches in the context of the real-life data for which the “true” degree of autocorrelation was unknown. We explored what would be the gains, if any, of constructing sampling distributions that matched as closely as possible the autocorrelation estimated from the data at hand (as is done in the SMA), instead of assuming a conservative, moderately high degree of autocorrelation (as is done in the context of the MRA).

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

To sum up, the aim of this study was 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 can 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 of a variety of data patterns with known characteristics, a Monte Carlo study was carried out.

Fig. 1 Representation of the way that p values are assigned to the outcome computed via the three primary indicators. The bottom rectangular box represents two of the main aims of this study, whereas the oval box synthesizes the corresponding results

The phase lengths were chosen to match the ones present in the real-life data sets detailed below. Short baseline phases were included, with n_A ranging from two to four measurements, in order to represent less favorable but potentially common situations (e.g., Shadish & Sullivan, 2011, found that almost 50 % of their baselines contained fewer than five data points). Complementarily, the 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.

For data generation, we simplified the Huitema and McKean (2000) model to $y_t = \beta_1 LC_t + \beta_2 SC_t + \varepsilon_t$. This model allowed for 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 ($\beta_1 = 1$), whereas for the slope change, an increase of .1 per measurement occasion was simulated ($\beta_2 = .1$). The ε_t term was specified by a first-order autoregressive model $\varepsilon_t = \phi_1 \varepsilon_{t-1} + u_t$ 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, a 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, 2008a; 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 effects, as well as trends. However, this model represents continuous data, which may not always be available in single-case studies in which count data (e.g., numbers of behaviors) are used.

Real-life data selection

Two studies (Schlosser & Blischak, 2004; Taylor & Weems, 2011)¹ using multiple-baseline designs (MBDs) across participants were included here. MBDs across participants have three characteristics that make them suitable for quantitative integration: Several data series are available, these series are independent (Jones & Fiske, 1953), and the same intervention and goals are present for each of them (Rosnow & Rosenthal, 2009). Note that not all data sets are independent in the Schlosser and Blischak (2004) study as four participants are studied in three different conditions. The data sets were

¹ We thank Leslie K. Taylor and Carl F. Weems for kindly providing their raw data for the analyses presented here.

Fig. 2 Representation of the way that p values are integrated across studies (in the case that only three studies are 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 represents the other two main aims of this study, whereas the oval box synthesizes the corresponding results

chosen to represent different psychological fields: The Schlosser and Blischak study deals with autism, aiming to increase spelling performance, whereas the Taylor and Weems study is focused on reducing posttraumatic stress disorder symptoms. Apart from the substantive differences, the studies are also distinct in terms of their baseline measurements, with the former study presenting no data variability (all of the baseline measurements were zero). Another reason for including these data sets was the range of treatment phase lengths. Nonetheless, it should be stressed that as the present study includes few data sets, it can only serve as an initial approximation to the degree of agreement between MRA's p values and the primary authors' conclusions.

Data analysis

Each simulated or real-life data set was analyzed via three primary indicators: the percentage of nonoverlapping 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

a more classical, parametric model were included. The PND was chosen (instead of, e.g., TauU; Parker, Vannest, Davis, & Sauber, 2011) just for illustration purposes, as an extreme case—an example of a nonoverlap procedure with deficiencies (e.g., distortions due to outliers or improving trends) to be compared with a more supported procedure such as the NAP, given that both are based on data overlap. 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 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, baseline linear trends are initially controlled for. Afterward, the detrended series are modeled using a dummy variable representing the change in phase and a variable representing slope change (i.e., LC and SC, respectively, in the data generation model presented above). Therefore, 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 MRA² reference values specific for the phase lengths and for the indicator. Three informal comparisons were carried out in the present investigation: (a) for the simulated data, the distribution of MRA's p values was compared across primary indicators, and considering whether or not an effect was present; (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 in Fig. 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 were used. The aim was to answer the following two questions: (a) Are the combined probabilities similar if all studies are analyzed using PND versus the NAP versus 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 used here, consider the following example. Suppose that we were studying only two primary indicators (PND and NAP) and that the results of only two data sets were intended to be integrated (Data Set 1 and Data Set 2). In such a situation, there would be four possible, distinct ways in which the indicators could be applied to the different data sets: (1) PND could be applied to both Data Sets 1 and 2; (2) NAP could be applied to both Data Sets 1 and 2; (3) PND could be applied to Data Set 1 and NAP applied to Data Set 2; and (4) NAP could be applied to Data Set 1 and PND to Data Set 2. Thus, there would be four variations with repetition, obtained after elevating the amount of indicators to a power equal to the number of data sets, 2². Using the same logic, there would be 36 ways in which three indicators could be applied to the six data Taylor and Weems (2011) data sets—that is, a total of 729 variations with repetition (or, 2⁶ = 64, if only the two nonoverlap procedures are included). For the Schlosser and Blischak (2004) data sets, there would be 312 = 531,441 variations with repetition (212 = 4,096, when only the nonoverlap procedures are included). For each of these variations with repetition, we computed the combined probability and studied the variability in distribution of the combined probability values. The desired result was low heterogeneity,

indicating similarity regardless of which indicator was computed for 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 Fig. 2.

Results

First aim: Assigning probabilities to generated data

When no intervention effect is simulated, the proportion of each p value is equal to the p value itself (e.g., approximately 1 % of the p values are .01 or less, 5 % are .05 or less, etc.) 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 corresponded to different p values, due to the discreteness of the primary indicator values. However, the p values generally became conservative when the MRA value for $\phi = .6$ was used for assigning probabilities to the outcomes.

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

First aim: Assigning probabilities to the real-life data

Assigning MRA p values to the primary indicators was done by comparing the MRA reference values for $p = .01, .05, .10, .20, .30$, and $.50$ ($p = 1$, in the case that the outcome was smaller than the reference value for $p = .50$). The results of applying the primary indicators and the probabilities assigned to them are presented in Table 3. Comparing among the 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 the absence of an intervention effect.

Second aim: Probabilities and the primary authors' judgments

For the Schlosser and Blischak (2004) data, the p values assigned to the NAP and the PND were $\leq .20$, which is well aligned with the authors' claim that the intervention was

² Note that for the present methodological study, identifying 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 were not the primary focus of this study.

Table 1 Simulation comparisons (conditions without a treatment effect) based on series lengths appearing in the real data used

$n_A + n_B$	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
	.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
	.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
	.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
	.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

The numerical values represent the proportions of p values as small as or smaller than the value depicted in the respective column header. NAP, nonoverlap of all pairs (Parker & Vannest, 2009); PND, percentages of nonoverlapping data (Scruggs et al., 1987); AG, regression-based procedure of Allison and Gorman (1993)

effective for all children regarding acquisition. For the data sets corresponding to the participants referred to as Jennifer, Elizabeth, and Michael in the Taylor and Weems (2011) study, all p values assigned to the NAP and the PND were $\leq .10$. These were the three children pointed out by Taylor and Weems as displaying more evident results.

Third aim: MRA-versus-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 ten 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 (2004) data and for the participant called Jennifer by Taylor and Weems (2011), 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 that had been 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 (2011) data sets, given that autocorrelation estimates ranged from $-.20$ to $.32$ for the baseline data, and only two estimates were greater than .60 for the treatment data. However, due to the high positive autocorrelation estimated for the Schlosser and Blischak (2004) 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 weighted means

According to the internal validity discussion, certain data sets (e.g., participants John, Sarah, and Kelly in the Taylor and

Table 2 Simulation comparisons based on series lengths appearing in the real data used

$n_A + n_B$	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
	$\beta_2=.1$	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
	$\beta_2=.1$	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
	$\beta_2=.1$	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
	$\beta_2=.1$	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

The numerical values represent the proportions of p values as small as or smaller than the value depicted in the respective column header. The conditions simulated correspond to changes in level ($\beta_1=1$) and changes in slope ($\beta_2=.1$). NAP, nonoverlap of all pairs (Parker & Vannest, 2009); PND, percentages of nonoverlapping data (Scruggs et al., 1987); AG, regression-based procedure of Allison and Gorman (1993)

Weems, 2011, study) probably should not be included in the quantitative integration, given that it was not clear whether the expected data pattern was 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 was applied to all data sets, the weighted average probabilities for the Taylor and Weems (2011) data were .30 for the PND, .32 for the NAP, and .65 for the AG. For the Schlosser and Blischak (2004) data, these values were .09, .10, and .86, respectively. The lower-weighted means for the Schlosser and Blischak data agree with the fact in 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 did 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 would vary according to the

indicator used for each of the data sets. The two boxplots on the left of Fig. 3 represent the distribution of the weighted averages for all variations with repetition for the Taylor and Weems (2011) data. When only the two nonoverlap procedures were used (leftmost boxplot), the combined p values were very similar, ranging from .26 to .38, with a median value of .32. When the AG was also used, the variability was greatly increased, with a minimum of .12 and a maximum of .86. The results for the Schlosser and Blischak (2004) data (the two boxplots on the right of Fig. 3) are similar, and illustrate even more clearly how the use of the AG (and the usually greater probabilities assigned to it) leads to increased variability in the weighted average probabilities.

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

When the same indicator was applied to all data sets and the binomial test was used for integrating, with “successes”

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

Data Set	$n_A + n_B$	NAP			AG			PND		
		Value	<i>p</i> MRA	<i>p</i> SMA ^a	Value	<i>p</i> MRA	<i>p</i> SMA ^a	Value	<i>p</i> MRA	<i>p</i> SMA ^a
John	3+8	83.33	.20	.05	.47	1.00	1.00	75.00	.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	100.00	.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 in the first six rows above), simulated data with phase-specific autocorrelation were generated, except for Jennifer. For the Schlosser and Blischak (2004) data, below the first six rows, the whole series were generated with the autocorrelation estimated from Phase B. NAP, nonoverlap of all pairs (Parker & Vannest, 2009); PND, percentages of nonoverlapping data (Scruggs et al., 1987); AG, regression-based procedure of Allison and Gorman (1993)

being defined as $p \leq .05$, the (binomial) probabilities of getting as many successes by chance were as follows.

- (a) For the six Taylor and Weems (2011) data sets: .0328 for the PND (two successes) and .2649 for the NAP and the AG (one success)

- (b) For the 12 Schlosser and Blischak (2004) data sets: $4.95 \cdot 10^{-7}$ for the PND and the NAP (seven successes) and 1 for the AG (zero 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 the single difference between NAP and PND for the Taylor and Weems data was associated with a considerable change in the binomial probability, given the few data sets being integrated.

When considering all possible variations with repetition (Fig. 4), for both the Taylor and Weems (2011) and Schlosser and Blischak (2004) data, the presence of AG as a primary indicator leads to a broader range of results. However, the proportions of binomial probabilities at or below .05 remained very similar: 40 %–50 % in the former data set, and 95 % in the latter, in which the intervention had more evident effects.

Fig. 3 Quantitative integration of *p* values assigned by means of the MRA for the six Taylor and Weems (2011) and the 12 Schlosser and Blischak (2004) data sets via the weighted mean probabilities, displaying all possible variations with repetition

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 were close to the ones

Fig. 4 Quantitative integration of p values assigned by means of the MRA for the six Taylor and Weems (2011) and the 12 Schlosser and Blischak (2004) data sets via the binomial test, displaying all possible variations with repetition

expected by chance when no intervention effect was simulated. Therefore, the conversion of the primary indicators into probabilities properly represented 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 increased when a level or slope change was simulated. The low probability of observing an extreme indicator value under the null hypothesis was well aligned with the fact that the null hypothesis was actually false, given that an effect was present (simulated) in the data. Hence, the p values did reflect the features of the data, even after an intermediate step in which data were summarized by a primary measure (i.e., here, a nonoverlap index), which was afterward converted into probabilities.

Concerning the second aim, both the individual and the combined p values assigned via the MRA agreed well with the original authors' conclusions for the data included here. That is, lower probabilities were obtained in the cases in which a greater effect was considered to be present. Note that these results were applicable only to the two nonoverlap 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 effectiveness.

As regards the third aim, the difference between the conservative MRA- and the SMA-based approaches was not found to be consistently in favor of either of the two. Specifically, the more data-specific SMA-based approach did not systematically lead to increased detection of existing intervention effects. In fact, the SMA was 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, we showed that the combined probabilities differed only slightly, regardless of which of the two indicators was applied to each data set. Thus, when the PND and the NAP were translated into p values and afterward combined, their similarity across all variations with repetition reflected well the fact that the same data sets had been analyzed and integrated. Once again, introducing the conservative AG made the variability of the p values greater.

Methodological implications

In the present 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 the 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 will necessarily be 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 a p value of, say, .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 can 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 of 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 by assuming that moderately high positive autocorrelation may be present in the data. Another issue that needs consideration when integrating SCD data is the baseline trend (Van den Noortgate & Onghena, 2003). The reference values obtained via the MRA can also be applied to data with trends, if for such data the trend is removed before computing the primary indicator, either via ordinary least squares regression or through first- or second-order differencing (for linear and quadratic trends, 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 was strong, as in the Schlosser and Blischak (2004) study.

It has to be mentioned that the application of the MRA to primary indicators that are expressed in 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 an 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 sensitivity to autocorrelation is controlled for by 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 in this study. The results with 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 so far has been developed to consider 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 results 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 the search for an optimal integration method, it is important to make explicit the main strengths and limitations of the existing alternatives. On the other, the actual use of these methods will be subject to their acceptability among researchers, once their strengths and limitations are known.

Author note This research was partially supported by the Spanish Ministerio de Ciencia e Innovación, Grant No. PSI2009-07076, and by the Agència de Gestió d'Ajust Universitaris i de Recerca de la Generalitat de Catalunya, Grant No. 2009SGR1492.

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). An evaluation of modified R^2 -change effect size indices for single-subject experimental designs. *Evidence-Based Communication Assessment and Intervention*, 2, 120–128.
- Beretvas, S. N., & Chung, H. (2008b). A review of meta-analyses of single-subject experimental designs: Methodological issues and practice. *Evidence-Based Communication Assessment and Intervention*, 2, 129–141.
- 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. doi:10.3758/BRM.40.2.467
- 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. doi:10.1037/0003-066X.49.12.997
- 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, UK: 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. doi:10.3758/BRM.42.3.930
- 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., & Shadish, W. R. (2010). Single-case designs technical documentation [Technical Report]. Retrieved February 10, 2012, from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/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. doi:10.1037/a0017736
- Kuppens, S., Heyvaert, M., Van den Noortgate, W., & Onghena, P. (2011). Sequential meta-analysis of single-case experimental data. *Behavior Research Methods*, 43, 720–729. doi:10.3758/s13428-011-0080-1
- 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. doi:10.1037/a0029248
- 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. doi:10.3758/s13428-011-0180-y
- 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. *Journal of Psychology*, 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. doi:10.3758/s13428-011-0111-y
- Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods*, 17, 510–550. doi:10.1037/a0029312
- Solanas, A., Manolov, R., & Onghena, P. (2010a). Estimating slope and level change in $N = 1$ designs. *Behavior Modification*, 34, 195–218.
- Solanas, A., Manolov, R., & Sierra, V. (2010b). 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. doi:10.3758/s13428-012-0213-1
- 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. doi:10.3758/BF03195492
- 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.