SECCIÓN METODOLÓGICA

Psicológica (2010), 31, 291-315.

The hard but necessary task of gathering order-one effect size indices in meta-analysis

Carmen Ortego[*1] & Juan Botella[2]

[1] *University of Cantabria, Spain;* [2] *Autonomous University of Madrid, Spain*

Meta-analysis of studies with two groups and two measurement occasions must employ order-one effect size indices to represent study outcomes. Especially with non-random assignment, non-equivalent control group designs, a statistical analysis restricted to post-treatment scores can lead to severely biased conclusions. The 109 primary studies included in 4 meta-analyses were recovered, and their authors were contacted to request the raw data to calculate the order-one effect size indices. From this total we only got 13 primary studies. The results with the raw data analysis were compared with those performed with the order-zero and order-one indices. Despite the difficulties for gathering the data, the few data sets analyzed show that if the meta-analysis is performed with order-zero indices, the results can be severely misleading.

The effect size (ES) has been defined as the degree to which the phenomenon is present in the population (Cohen, 1977), the degree to which the results differ from the null hypothesis (Cohen, 1994; Thompson, 2006), the magnitude or strength of the results (Johnson, Mullen & Salas, 1995), or the size of the relationship between any two variables (Rosenthal, 1991). Meta-analysis can only achieve its goals by effecting good choices in

[*] We thank Julio Sánchez-Meca and two anonymous reviewers for their helpful suggestions during the reviewing process. Financial support for the second author comes from project SEJ2006-12546/PSIC of the Ministerio de Ciencia y Tecnología of Spain. Our special thanks to those who are collaborated sharing their data when requested: Rebecca G. Anderson, Ann Azzollini, Thomas Baranowski, Marcia Bayne-Smith, Felica D. Bradford, Jill E. Bormann, Valerie Burke, Paul Fardy, Mary Gruber, Peter Hannan, Joanne S. Harrell, Chris A. Hopper, Juliana Kain, Cari McCarty, Robert McMurray, Judith Neiding, Dianne Neumark-Sztainer, HyunJu Park, Paul Rohde, Kathleen Sikkema, Eric Stice, Jeffrey Weiss, John R. Weisz and the funding source NINR01837 and those who has helped to make the translation of this document: Soledad Fernández and Nicola Bolton. Correspondence: Carmen Ortego (maria.ortego@unican.es).

the method used to represent the studies outcomes. The weighted combination and analysis of several independent estimates of the effect size indices selected to represent the outcomes allows reaching answers to the questions posed.

There are many alternative indices of ES (Fleiss, 1994; Rosenthal, 1991, 1994; Sánchez-Meca, Marín-Martínez & Chacón-Moscoso, 2003); Huberty (2002) has listed up to 61 different indices. The ES index can be selected from this range taking into account the methodological characteristics of the specific field, the hypothesis of research, and the type of study outcome analyzed. The range of methods to manage the results has no known limits. However, despite this variety, those most frequently employed in psychology are essentially three: the standardized mean difference, the Pearson's correlation coefficient and the odds ratio (Botella & Gambara, 2002; Cooper, 1989; Lipsey & Wilson, 2001; Marín & Sanchez, 1996; Rosenthal, 1994).

The increasing use of multivariate models has conveyed the need of resuming results that reflect the complex relationships existing among a larger number of variables compared to the bivariate case. This is the main reason for putting forward a claim for an increasing use of ES indices that takes into account the influence of a covariate (Keef & Roberts, 2004; Huedo, 2006). These are called order-one indices or partial effect sizes, whereas the three basic indices mentioned in the previous paragraph are called order-zero indices.

A simple multivariate design, very common in applied settings, includes two groups (experimental, control) and two moments of measurement (pre-treatment, post-treatment). If it is a quasi-experimental design, with a non-equivalent (non-randomized) control group, there is no guarantee for the equivalence between the groups before the intervention (Cook & Campbell, 1979). Thus, the pre-treatment measure allows testing the equivalence in the pre-treatment scores. Even when randomization is included in the experimental design and it is expected that it will equate the groups, this process can fail. Especially with small samples, randomization can yield samples with important differences. The order-zero ES indices (based on post- scores) can be severely biased when the two groups are non-equivalent and the pre-treatment scores are ignored. The order-one indices provide a better performance as they do take into account the pre-treatment scores. Huedo (2006) has shown in simulation studies the dramatic effects that ignoring the pre-treatment scores can have on the conclusions of the meta-analysis. In the present study we illustrate this at a practical level

through the statistical information provided by a set of primary studies, included in a published meta-analysis.

The present study

Originally, our plan for the present study included three steps, although finally it had to be changed because of the circumstances described below. First, we chose a problem with a potentially high number of studies published employing the design described above: two groups (experimental, control) and two measures (pre-treatment, post-treatment). Although what we were looking for is better captured in quasi-experimental designs with non-equivalent control group, it could also appear with randomized designs (especially with small samples). In the process of identification of possible studies to be included, we browsed for a meta-analysis already published on a problem for which that design was well suited. Thus, we identified four meta-analysis candidates to be the source of primary studies. Second, the attempt was made to access the primary studies included in these meta-analyses. Third, we performed parallel analyses, with order-zero and order-one ES indices, comparing the conclusions reached.

Our plan was modified because of the insurmountable difficulties in gathering the information from the primary studies. For meta-analyses employing order-zero ES indices, the statistic information for the calculations is often available in the published results. However, order-one indices are based on statistics usually absent in the papers. Thus, we decided to ask the authors of the primary studies for the raw data and then calculate by ourselves all the desired indices. This allowed us a direct analysis of the data in a scenario of primary analysis, rather than in a meta-analytic scenario, giving us the opportunity of comparing the conclusions reached with primary and meta-analytic procedures.

Choosing the Effect Size indices and gathering information from the primary studies

We selected five ES indices: two order-zero and three order-one indices. They were selected according to their frequency of use and their potential advantages. The appendix includes a more detailed description of the indices. The order-zero indices are the two more commonly employed in psychology with quantitative variables: the standardized mean difference (δ) and the Pearson's correlation coefficient (ρ). They are related algebraically and only differ in the easiness for interpreting the results (Rosenthal, 1994).

Our order-one indices are those judged as conceptually closer to the order-zero indices selected; of course, they take into account the pre-treatment scores (two alternatives to δ and one alternative to ρ): the partialized standardized mean difference (δ_p), the difference between the standardized mean change scores (δ_c), and the partial correlation (ρ_p). There are many more alternatives, and indeed some of them have shown interesting properties that make them more desirable than some of those selected. However, we have not taken them into consideration because it is unrealistic to employ the discarded alternatives since the researcher cannot gather all the necessary information from a relevant number of the primary studies selected.

We searched in the databases *PsycInfo*, *Medline* and *ERIC* for some meta-analyses that met the six following inclusion criteria:

1) The design of the primary studies included in the meta-analysis should have at least two groups, one of which must be a control group, and with at least two measurements on each group, pre- and post-treatment.

2) The equivalence of the groups in the pre-treatment measures is not guaranteed by means of random assignment to the groups.

3) The meta-analysis reports enough details to be replicated step-by-step (Botella & Gambara, 2006; Rosenthal, 1995).

4) The meta-analysis was published between 2006 and 2008.

5) At least more than half of the primary studies included in the meta-analysis had been published after 1995, as recent publication increases the probability of gaining access to the authors and the data.

6) The dependent variable recorded in the pre- and post-treatment must be measured with the same and unique scale.

We selected four meta-analyses. One met the six inclusion criteria, whereas the other three met five of them. One hundred twenty-one primary studies integrated the four meta-analyses; finally, we obtained the written reports of 109 primary studies.

As is easily appreciable in the appendix, for calculating the five ES indices and their variances, many statistical details that are not routinely included in the published reports are needed. In fact, only half of the primary studies provided the values of the pre- and post-treatment sample sizes in both groups (control and experimental), their means and standard deviations. Moreover, none of the primary studies provided the bivariate correlations, the partial correlations, the mean square error of the analysis of

covariance and the adjusted means (when an ANCOVA model was employed).

Since more than half of the 109 primary studies did not provide enough statistical information we decided to contact the main authors (or their co-authors when we were unable to contact the main ones) asking for the pre-post raw data. The years of publication of the 109 primary studies fell between 1980 and 2006, and 75% of the primary studies were published before 2003. However, many authors might have moved to a different job, and furthermore some of the older primary studies did not provide the e-mail address. In fact, 53 main authors had changed their addresses and we were unable to find the current address of twelve authors.

To find out the current address of the main authors we used *PsycInfo* and *Medline* databases, where we sought the main author of the study. When we did not find it or we were not sure or unsatisfied with the findings, we made a new search in *Google*. If this second search was unsuccessful, the same process was followed for the second author. In the case where we did not obtain the address of either the first or second author, we decided to search for any recent publication which he or she might have co-authored. We contacted the main author of this publication requesting the current address of the researcher focus of our attention. As a result, we contacted six main authors successfully, as all of them provided us with the updated information almost immediately.

We then sent e-mails to each of these researchers, asking if they were willing to send us the pre- and post- raw data in both groups from their studies, briefly explaining the purpose of our research and the utility that their data could have for us. Therefore, 83 e-mails and 6 letters were sent. From this total only 48 authors answered this first request, eight authors sent us their data but twenty-five authors did not send their data, giving different reasons.

About a month later, we again sent the same message as on the first occasion along with a new request to those authors who had not replied to the first message or who had answered committing themselves to provide an answer, but had failed to do so at the time. Unfortunately, after this second attempt, only one author sent us the raw data.

Approximately a month after the second request we re-sent the first and second messages together, as a third attempt, to the authors that had not yet answered any of the previous requests. The same procedure was followed with those that had promised to collaborate but had not done so. Following this third request another four authors delivered their data to us.

In short, after contacting the leading authors of 109 primary studies included in four meta-analysis, with the common feature that their primary studies involved a design of at least two groups (control and experimental) and two measure occasions (pre- and post treatment), in both groups, and after sending 237 e-mails and 6 letters, we only got the raw data of 13 primary studies. From these 13 data sets, 8 were sent after the first request, 1 after the second, and 4 after the third request.

The 237 e-mails and 6 letters were distributed as follows: (a) first request, 83 e-mails and 6 letters; (b) second request, 42 e-mails; (c) third request, 24 e-mails; (d) additional information, 41 e-mails (e) acknowledgements, 34 e-mails thanking for their collaboration and 13 e-mails thanking the authors for actually sending their raw data.

We describe this process in some detail because the first conclusion of our study has to do with this. The papers describing the primary studies do not contain, in general, enough information for calculating the estimates of the order-one ES indices, just those more suited for the type of designs we have worked with. Furthermore, the attempts to gather from the authors the raw scores for making the calculations have been exhausting and almost useless. The same feeling has been experienced by others in similar circumstances (Wicherts, Borsboom, Kats & Molenaar, 2006). As far as is reflected in our experience, many psychologists do not adhere in practice to the APA guides about data sharing, a point that has been already highlighted by leading scientific journals (e.g., Botella & Ortego, in press; Nature, 2006).

Data Analysis

Despite the above conclusion, the authors of 13 reports sent us their raw data: seven from the meta-analysis of Stice, Shaw and Marti (2006), whereas the other six (Bormann, Gifford, Shively, Smith, Redwinw, Kelly, Becker, Gershwin, Bone & Belding, 2006; Ettelson, 2002; Neiding, Smith & Brashers 2003; Rohde, Clarke, Mace, Jorgensen & Seeley, 2004; Sikkema, Hansen, Kochman, Tate & DiFranceisco, 2004; Weiss, Mulder, Antoni, de Vroome, Garssen & Goodkin , 2003) were distributed between the other three meta-analyses (Christensen, Kristensen, Bartels, Bliddal & Astrup, 2007; Scott-Sheldon, Kalichman, Carey & Fielder, 2008; Weisz, McCarty & Valeri, 2006). Obviously, the only group of studies that could serve our purposes is that of the primary studies included in the Stice et al's meta-analysis (2006). The sources of the seven studies are marked with an asterisk in the references. Five studies contributed with one data set, whereas one study contributed with two sets, and one study contributed

with three sets. The following calculations and analyses are based on those 10 data sets. Table 1 shows the descriptive statistics.

Table 1. Main descriptive statistics of the 10 data sets included in the study. The first column shows the categorization of each study as randomized (R), non-randomized (NR) or unclear.

	Experimental Group						Control Group					
Data set	N	$\overline{BMI}_{pre}$	S^2_{pre}	$\overline{BMI}_{post}$	S^2_{post}	$r_{pre/post}$	N	$\overline{BMI}_{pre}$	S^2_{pre}	$\overline{BMI}_{post}$	S^2_{post}	$r_{pre/post}$
1 - R	18	21.37	19.26	21.97	17.06	.989	14	26.33	65.01	26.84	64.71	.997
2 - R	404	22.64	16.52	22.66	16.50	.969	177	23.69	25.51	23.83	26.64	.960
3 - U	265	18.32	7.78	18.72	8.59	.964	220	17.59	5.71	18.07	5.88	.959
4 - U	251	17.96	7.62	18.57	8.56	.964	220	17.59	5.71	18.07	5.88	.959
5 - U	41	17.29	5.48	17.50	5.64	.902	48	17.07	8.64	17.19	7.78	.929
6 - NR	2098	19.63	14.26	19.73	13.30	.965	888	19.20	13.07	19.39	12.09	.965
7 - U	261	21.52	25.00	21.74	25.69	.990	244	21.46	23.22	21.67	23.05	.995
8 - U	312	22.55	29.20	22.73	29.51	.988	244	21.46	23.22	21.67	23.05	.995
9 – U	298	22.28	27.93	22.50	27.60	.991	244	21.46	23.22	21.67	23.05	.995
10 - NR	80	27.80	49.97	28.19	55.93	.981	99	25.85	34.57	26.21	34.86	.982
Overall	4028	20.43	20.46	20.61	20.13	.978	2398	20.20	21.49	20.45	21.02	.983

The studies involve programs for prevention of obesity. One of the dependent variables analyzed is the Body Mass Index (BMI), defined as the weight in kilograms divided by the squared height in meters. It is expected that the participants that receive the programme will increase their BMI less (or it may even remain stable or decrease) than those in the control group. As we have the raw scores, we can provide separate analyses based either on the raw scores or on the group statistics. In fact, one of our main purposes was to compare the conclusions reached with both analyses. The results are presented in three sections. First, we describe the analysis of the

raw scores. Second, we reproduce a meta-analytic scenario in which the meta-analysts have only the group statistics and employ order-zero ES indices. And finally, we employ all the information to calculate the order-one ES indices and repeat the meta-analysis with these indices.

Analysis of the direct scores

First of all, we have analyzed the scores separately for each moment. We ran an ANOVA on the post- scores to test whether there are significant differences between the experimental and control groups when ignoring the pre- scores. Table 2 shows that setting α in 0.05 there are 3 programs (5, 7 and 9) that show no significant differences. For the remaining seven programs, the significant difference in 2 studies is in the expected direction (1 and 2) and in 5 programs the difference appears to be in the opposite direction (3, 4, 6, 8 and 10). The same analysis on the simple aggregation of data, ignoring the program, showed no significant difference (table 2, last row).

Then we analyzed (also with ANOVAs) for any difference between the groups in the pre-treatment scores. Table 2 shows that there are no significant differences in 4 programs (4, 5, 7 and 9; three of them are the same that show no significant differences in the post- scores). About the rest, in two programs the difference is in the expected direction for the post-scores (1 and 2, the same as in the post- scores) and opposite in the other four (3, 6, 8 and 10). The aggregated whole sample also shows a significant effect (table 2, last row). This first set of analysis reveals that an analysis that ignores the pre-scores probably leads to erroneous conclusions because there are important pre-treatment differences that should be considered in some way.

The most appropriate way to analyze the raw scores is by a factorial ANOVA (2 occasions in 2 groups) to check for statistical significance of the interaction (a statistically identical choice consists in analyzing the change scores, as the difference between the individual post- and pre-scores). Last column of Table 2 shows the results for each data set. Only one data set (program 6) shows statistical significance for the interaction. Graphing the means for that program (Table 1) shows that the interaction reveals a beneficial effect of the intervention (BMI increases from pre- to post- moment to a lesser extent in the experimental than in the control group). Despite that only one data set shows a significant interaction, the aggregated sample shows also a significant effect (table 2, last row). The inspection of the grand means (see table 1, last row) reveals that this is again due to a beneficial effect of the programs. The experimental group

increases the mean BMI, from the pre- to the post-treatment occasions, significantly less than the control group.

Table 2. ANOVAs of the pre- and post-treatment scores. Last column shows the test for the interaction in the two-way ANOVA.

Data set	ANOVA Pre- scores		ANOVA Post- scores		F for the interaction effect	
1	$F_{1,30} = 4.961$ (.034)	+	$F_{1,30} = 4.94$ (.034)	+	$F_{1,30} = 0.188$ (.668)	
2	$F_{1,579} = 6.991$ (.008)	+	$F_{1,579} = 8.629$ (.003)	+	$F_{1,579} = 1.456$ (.228)	
3	$F_{1,483} = 9.443$ (.002)	-	$F_{1,483} = 6.972$ (.009)	-	$F_{1,483} = 1.383$ (.240)	
4	$F_{1,469} = 2.400$ (.122)	-	$F_{1,469} = 4.001$ (.046)	-	$F_{1,469} = 3.552$ (.060)	
5	$F_{1,87} = 0.151$ (.699)	-	$F_{1,87} = 0.324$ (.571)	-	$F_{1,87} = 0.170$ (.681)	
6	$F_{1,2984} = 8.339$ (.004)	-	$F_{1,2984} = 5.548$ (.019)	-	$F_{1,2984} = 5.519$ (.019)	+
7	$F_{1,503} = 0.015$ (.901)	-	$F_{1,503} = 0.021$ (.884)	-	$F_{1,503} = 0.032$ (.859)	
8	$F_{1,554} = 6.094$ (.014)	-	$F_{1,554} = 5.797$ (.016)	-	$F_{1,554} = 0,176$ (.675)	
9	$F_{1,540} = 3.447$ (.064)	-	$F_{1,540} = 3.616$ (.058)	-	$F_{1,540} = 0.087$ (.768)	
10	$F_{1,177} = 4.037$ (.046)	-	$F_{1,177} = 3.904$ (.050)	-	$F_{1,177} = 0.027$ (.871)	
Overall	$F_{1,6424} = 3.899$ (.048)	-	$F_{1,6424} = 1.771$ (.183)	-	$F_{1,6424} = 10.806$ (.001)	+

It should be highlighted that the use of an ANCOVA model for the post- scores with the pre- scores as the covariate is controversial. While it is a correct choice for randomized designs, it is not for non-randomized, non-equivalent, quasi-experimental designs (Miller & Chapman, 2001; Van Breukelen, 2006). The studies integrated here employ both designs and, in fact, in the primary reports ANCOVA is the choice in several studies. That is why we have not used it for our integrative purposes.

Furthermore as we have the data from all the studies available, we can perform a single, integrated analysis. We have performed a multilevel analysis of the change scores with random effects in the intercepts and the slopes. The conclusion is that there is no effect of the treatment ($F<1$).

In short, the analysis of the raw scores at the data set level generates an inconsistent and contradictory view. However, the overall analysis of raw data indicates that these programs have, on the average, a significant beneficial effect, although this effect is so small that it is doubtful whether it is relevant. In fact, to achieve statistical significance of the interaction in the aggregated sample it has been necessary to use a really large sample size, by aggregating all the data sets. Whereas the mean difference in the post-scores is 0.16 BMI points (the average in the experimental group being larger!), the difference between the differences at the pre- and post- occasions is only 0.07 BMI points, favouring the experimental group (smaller difference than the control group). The multilevel statistical model offers a different conclusion.

Meta-analysis with order-zero ES indices

As mentioned above, estimates of two order-zero ES indices are used in this phase of our study: the standardized mean difference (d_g) and the Pearson's correlation (r). As they are order-zero estimates they do not take into account the pre- scores. Table 3 shows the d_g and r indices calculated for each data set, as well as the combined estimate (weighted by the inverse of the variance) and the test of homogeneity. As d and r are algebraically related the conclusions are identical. In fact, the correlation between the 10 pairs of d_g and r values is 0.999.

In line with the findings in previous analyses there are only two data sets (1 and 2) in which the effect has a positive sign, indicating a beneficial effect of such programs, whereas the other programs show negative effects. The meta-analytic integration appears on the last row of table 3. As can be seen in table 3, the estimates with both ES indices, assuming a random effects model (RE), suggest that this group of programs does not generate an effect significantly different from 0. On the contrary, using a fixed effects model (FE) the estimate is significantly different from 0 (but reflecting a detrimental effect of the programs!). Assuming a random effects model, we must conclude that the programs are ineffective or, even worse, that they have a negative effect, as all point estimates (both indices and models) are negative.

Table 3. Meta-analysis with order-zero ES indices.

Data set	d_g	r
1	0.772	.376
2	0.264	.121
3	-0.240	-.119
4	-0.184	-.092
5	-0.120	-.061
6	-0.094	-.043
7	-0.013	-.007
8	-0.206	-.102
9	-0.164	-.082
10	-0.296	-.147
Meta-analysis	d_g = -0.091 (RE) 95%CI[-0.200; 0.018] Q(9)=29.665 (p=.0005) d_g = -0.095 (FE) 95%IC[-0.147; -0.044]	r = -.045 (RE) 95%CI[-.099; .009] Q(9)=30.582 (p=.000) r = -0.044 (FE) 95%IC[-.069; -.020]

Meta-analysis with order-one ES indices

We calculated estimates for the three order-one ES indices selected for the present study: the partialized standardized mean difference (d_p), the difference between the standardized mean change scores (d_c) and the partial correlation coefficient (r_p). Unlike what happened with the order-zero ES indices, in the order-one indices the pre-treatment scores are taken into

account. Table 4 shows the three estimates for each data set. The last row of table 4 shows the overall effect size and the test for homogeneity. As with the order-zero indices, given that the d_p and r_p are algebraically related, their results are similar (the correlation between the 10 pairs of values of d_p and r_p is 0.999).

Table 4. Meta-analysis with order one ES indices

Data set	d_p	d_c	r_p
1	0.0216	-0.072	.0110
2	0.1325	0.024	.0610
3	0.1047	0.058	.0520
4	-0.1718	-0.019	-.0860
5	-0.1127	-0.049	-.0570
6	0.0657	0.027	.0300
7	-0.0161	-0.000	-.0080
8	0.0244	0.009	.0120
9	-0.0430	0.001	-.0210
10	-0.0039	0.006	-.0020
Meta-Analysis	d_p = 0.015 (RE) 95%CI[-0.044; 0.074] Q(9)=25.075 (p=.003) d_p = 0.049 (FE) 95%IC[0.021; 0.076]	d_c = 0.009 (RE) 95%CI[-0.003; 0.022] Q(9)11.470 (p=.245) d_c = 0.010 (FE) 95%IC[-0.0003; 0.019]	r_p = .015 (RE) 95%CI[-.010; .04] Q(9)=8.801 (p=.456) r_p = .015 (FE) 95%IC[-.01; .04]

A salient feature of the figures draws our attention. We expected that the values of the order one ES were different from the order-zero values, but what is more interesting is the fact that in several programs the sign of the effect is reversed. That is, whereas some data sets showed differences favouring the experimental group (or control group) with the indices based only on post-treatment scores, when they are analyzed with indices that take into account the pre- scores some of them show the opposite sign.

Looking at the d_p, five data sets show a positive effect of the intervention and five show a negative effect. If we assume a random effects model the overall size from these values reflects a positive effect, though not significantly different from 0 (d_p = 0.015). With a fixed effects model it reaches statistical significance (d_p = 0.049), but this last model should not be accepted because the test for homogeneity is also significant and we want to make inferences about the population of studies within which our data sets is considered a random sample (Hedges & Vevea, 1998).

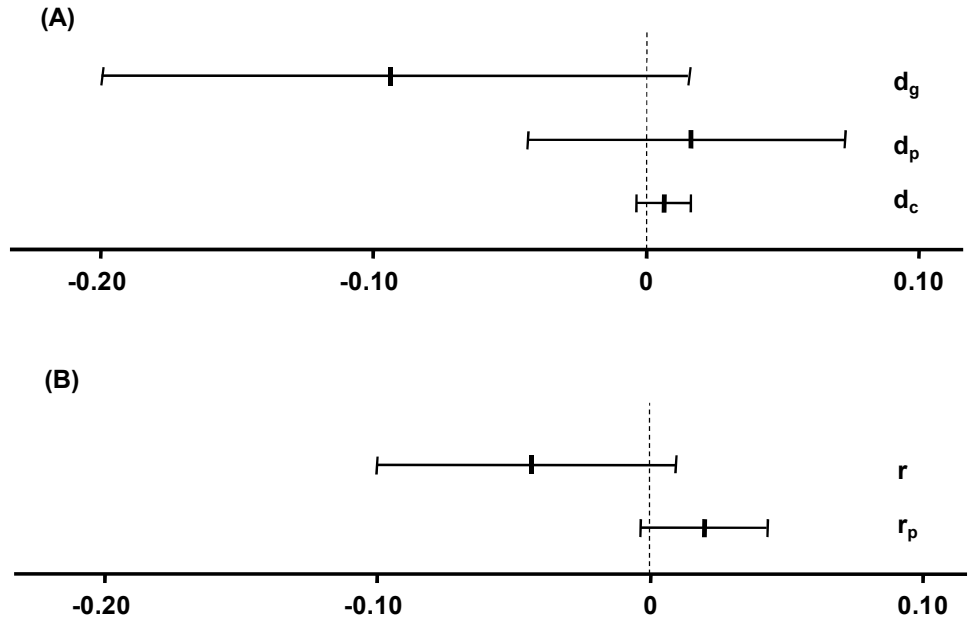

Figure 1. Overall estimate and confidence intervals for ES indices (order-zero and order-one) of the *d* and *r* families (panels A and B, respectively), assuming a random effects model.

Related to the d_c index, table 4 shows that for six data sets (2, 3, 6, 8, 9 and 10) the estimates are positive, which means that the BMI is reduced in the experimental group as compared to the control group after the intervention. But the other four programs show the opposite sign. The overall effect size in these indices indicates a positive effect, though not significantly different from 0 (d_c = 0.009).

As to the r_p index there are also studies with positive and negative estimates. Of course, the signs agree with d_p and with the overall effect size. Positive correlations are expected from the intervention effect. Table 4 shows that this happens in five programs (1, 2, 3, 6 and 8) while in the other five the opposite happens. The meta-analysis on these values indicates a positive overall effect, though not significantly different from 0 (r_p = 0.015).

In short, looking at the two order-zero indices (d_g and r) there are positive estimations only for data sets 1 and 2, whereas looking at the three order-one indices (d_p, d_c, and r_p) there is a larger number of programs reflecting positive estimations. This fact will affect the combined estimations producing larger order-one ESs than order-zero ESs, as shown in Figure 1.

The main difference between the order-zero indices and the order-one indices is that the second group yield positive point estimates (for all indices and models), whereas the first group show negative estimates.

Discussion

A specific difficulty for employing the methodology of meta-analysis, and one of its main challenges, shows up when the statistics from primary studies to be included are absent in the published report and must be previously gathered from the authors. A common complaint among meta-analysts is that the statistical description of the data is scarce. The information available is often enough for calculating order-zero ES indices, but the specific statistics needed for calculating order-one indices are almost always absent.

The ES for the outcome of an intervention with two occasions (e.g., pre and post treatment) and two groups (e.g., experimental and control) should not be measured with order-zero indices. This is especially important when there is no random assignment of participants to groups. In quasi-experimental designs with non-equivalent groups a significant effect on the post-treatment scores is not enough evidence for concluding about the treatment impact (e.g., Keppel & Wickens, 2004; Kirk, 1995; Maxwell & Delaney, 1990). Potential differences in the pre-treatment scores must be taken into account in the statistical analysis, which must rely on the interaction of the factorial ANOVA or on the direct comparison of the change scores. Of course, if this is true for the analysis of any primary study the same must be said for a meta-analysis that integrates studies that use this type of designs. The way to do this is using order-one ES indices instead of

the more frequent order-zero ES indices. Many previous papers (e.g., Becker, 1988; Dunlap, Cortina, Vaslow & Burke, 1996; Gibbons, Hedeker & Davis, 1993; Keef & Roberts, 2004; Morris, 2000; 2008; Morris & DeShon, 2002; Viechtbauer, 2007) focus on the benefits of using order-one effect size indices to control for differences in the baseline when comparing two groups at the post-test; we show it with a group of 10 data sets included in a published meta-analysis.

However, our experience indicates that the information needed to calculate most of the order-one estimates is absent in the published reports, despite the increasing insistence of journal editors on its importance (Peterson & Brown, 2005). Even worse, the authors of the primary studies are reluctant to share their data, even after the paper has been published. Currently, performing meta-analysis with primary studies that need order-one indices is a very difficult task. Only one of the order-one ES indices employed in the present study (d_c) can be calculated with the sample statistics sometimes reported, and it is not the one with better statistical properties, as shown in the simulation study of Huedo (2006). The results of studies with the designs we are dealing with cannot be performed with the order-one ES indices with better properties.

A possible course of action is carrying out the calculations taking as a basis the subsample of studies that provide the information needed. But this restriction increases the sampling error, reduces the accuracy and the generalizability of the population ES estimates and could bias the overall ES estimation (Glaser, 2002; Peterson & Brown, 2005). An alternative is ignoring the pre- scores and assume the previous equivalence of the groups in the pre- scores, but this can have even larger costs. On one hand, it is precisely the known threat to internal validity labelled as "selection of samples" (Cook & Campbell, 1979) that motivates the use of two-moment designs. Ignoring this in the meta-analysis would not make any sense. A better choice is taking an order-one index, d_c, which is better achieved with the information usually available in the reports although it is not the best choice in statistical terms. It could represent a good balance between availability and statistical properties. Furthermore, there are formulas for calculating the statistics needed (specially the standard deviation of the change scores or the correlation between the pre- and post- scores) based on the t-test for the change. The average value can also be employed for imputation to the studies where it is not available[1]. However, in order to have a good idea of how this way would work in our study, we must

[1] Thanks are due to the associate editor, Julio Sánchez-Meca, who directly suggested in the process of review this way to complement the calculations.

highlight that none of the primary studies reported the correlations between the pre- and post- scores. It is necessary for calculating the variance of d_c (see the appendix). The variance of the change scores can also be obtained if a t-test statistic for the pre- post difference is available; only 7% of the comparisons offered this in our study. Alternatively, most of them included F ratios for tests with several groups, without the MCE, or the F for ANCOVA.

A possible solution to the shortage of statistical information is requesting the authors of the primary studies for the statistics needed to calculate other desired ES indices. Taking as a basis our experience, we think that it is better asking for the raw data instead of the statistics, for several reasons. First, most of the ES indices (order-zero or order-one) are parametric, and in order to get a correct inference some assumptions must be made; they can be tested if the raw data are available. Second, when conducting a meta-analysis it is necessary to make a number of decisions along the process. To reduce the bias in the process of decision making, it is essential that the meta-analysis is carried out by more than one coder and that these coders discuss their disagreements. The fewer things left to be guessed, the more accurate findings will be reached. The fact of having the raw data from the primary studies not only allows to check the assumptions, but also offers the possibility of performing the same statistical data. Besides this, different experimental designs can lead to estimate different parameters. This must be taken into account in the meta-analysis, as some adjustments are often needed for the ES calculated on data from different experimental designs (Morris & DeShon, 2002).

The alternative chosen in the present study has been to request the authors to send the raw data, a strategy often recommended when the published reports do not provide the statistical information (Orwin, 1994). However, this has shown to be a tedious and unproductive task, as others have complained (Wicherts, Borsboom, Kats & Molenaar, 2006). In our case, although we requested the raw data from the main authors of 109 primary studies we only obtained the information from 13 primaries studies (11.9%).

Obtaining the raw data should be an easy target itself, being an ethical principle of research that should be observed by researchers, regardless of their scientific discipline. Institutions and documents, such as the APA, the Code of Good Scientific Practice, the Fifth Reform of the Helsinki Declaration (WMA), professional organizations as the American Medical Association (AMA) or editorial policies of leading scientific journals as

Nature or *Science*, point the need to share information between researchers in a more or less explicit way.

Despite the effort and time invested in getting the raw data and the small number of data sets obtained, we have been able to verify the differential estimations of the order-zero and order-one ESs. In our sample of data sets, when using order-zero indices, that do not take into account the pre-scores, a negative overall effect size, although non-significant, is obtained. Therefore, the use of order-zero indices provides a result that goes against what was expected. On the contrary, the order-one indices yield a positive effect, although again non-significant. The order-one indices which include both the post-scores and the pre-scores increase the sensitivity of the indices, and are able to detect a positive effect, in the expected direction. This indicates that, even when there is no equivalence in the groups at the pre- occasion, the intervention has caused the desired effect in the experimental group as compared with the control group; this is reflected in the order-one indices but not in the order-zero indices.

As pointed out in our goals, the double analysis presented in this paper gives us the opportunity of comparing the conclusions reached with primary and meta-analytic procedures. The comparison shows that the meta-analysis with order-one indices yields conclusions close to those from the primary analysis with the raw data, whereas the meta-analysis with order-zero indices does not. The source of the discrepancy between them is that the ANOVAs of the pre- scores shown in Table 2 show the non-equivalence between the groups before the intervention. This lack of equivalence is simply ignored when order-zero indices are employed for the meta-analysis.

In spite of the small number of data sets involved, our study has shown that order-one indices have to be employed instead of the order-zero ones in order to reach conclusions about the effect of an intervention in a meta-analysis that includes primary studies with this type of designs.

Conclusions

In 1976 Gene V. Glass, in the presidential communication of the annual meeting of the American Research Association, coined the term meta-analysis to refer to *"the analysis of analysis*" (Glass, 1976, pg. 3) defining it as "*the statistical analysis of a large analysis collection of analysis results from individual studies for the purpose of integrating the findings*" (Glass, 1976, pg.3). Since then many researchers keep trying to improve this methodology. Thus, today we know that order-one indices are more appropriate for meta-analyses of primary studies with two groups and

two measure occasions than the order-zero indices, which involve only the post-treatment scores. This is especially important with quasi-experimental designs with a non-equivalent control group as pre-treatment equivalence is not guaranteed with random assignment. However, for this practice to be successfully implemented the primary studies need to provide much more statistical information than they have provided so far.

Achieving the cooperation of primary studies authors is a must. Their contribution is essential for a further development of the meta-analysis. In fact, without the fair cooperation of some colleagues, actively involved in providing the raw data from their primary studies, it would have been impossible for us to finish the present research successfully. At the end, our analyses show clearly that we cannot trust the conclusions if order-zero ES indices are employed. But they also show the practical difficulties that in fact are encountered when trying to gather the statistical information or the raw data. This process would be much easier if these were already available in data repositories. The APA and other institutions should foster a better cooperation between psychologists and encourage the move towards routine data sharing by open access data bases. The statistical information included in the reports is so far clearly insufficient. That is why we think that, although order-one ES indices are completely necessary for a proper meta-analysis of studies with some types of designs**,** it is still too difficult, and this will remain the same as lone as fair data sharing is not assumed among psychologists (Botella & Ortego, in press; Nature, 2006).

RESUMEN

La difícil pero necesaria tarea de reunir índices de orden uno en meta-análisis. El meta-análisis de estudios primarios con diseño de dos grupos y medidas en dos momentos debe emplear índices de tamaño de efecto de orden uno. Especialmente si la asignación no es aleatoria, con diseños de grupo control no equivalente, las conclusiones alcanzadas pueden estar fuertemente sesgadas si sólo se incluyen las medidas post tratamiento. 109 estudios primarios incluidos en 4 meta-análisis fueron recopilados y se contactó con sus autores para pedirles los datos originales con fin de estimar los índices de orden uno. De este total sólo se consiguió 13 estudios primarios. Los resultados obtenidos con los datos originales fueron comparados con los estimados con los índices de orden cero y uno. A pesar de las dificultades para conseguir los datos, el pequeño grupo de datos analizados mostró que si el meta-análisis se realizaba con índices de orden cero las conclusiones eran erróneas.

REFERENCES

(References with an asterisk (*) indicate the studies from which the raw data were gathered from the authors and take part of the data sets analyzed here).

American Medical Association (1994) *Current opinions of the medical council on ethical and judicial affairs. E-9.08 new medical procedures.* Chicago: American Medical Association.

American Psychological Association (2006). *Publication Manual of the American Psychological Association* (5th ed.). Washington, DC: American Psychological Association.

*Baranowski, T., Baranwski, J. C., Cullen, K. W., Thompson, D. I., Nicklas, T., Zakeri I. F. & Rochon, J. (2003). The fun, food, and fitness Project (FFFP): The Baylor GEMS pilot study. *Ethnicity & Disease, 13* (Suppl.1), S1-30-S1-39.

*Bayne-Smith, M., Fardy, P. S., Azzollini, M. S., Magel, J., Schmitz K. H. & Agin, D. (2004). Improvements in heart health behaviors and reduction in coronary artery disease risk factors in urban teenaged girls through a school-based intervention: the PATH program. *American Journal of Public Health, 94*(4), 1538-1543.

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical & Statistical Psychology, 41,* 257-278

Bormann, J. E., Gifford, A. L., Shively, M., Smith, T. L., Redwine, L., Kelly, A., Becker, S., Gershwin, T. L., Bone, P. & Belding, W. (2006). Effects of spiritual mantram repetition on HIV outcomes: A randomized controlled trial. *Journal of Behavioral Medicine, 29* (4), 359-376.

Botella, J. & Gambara, H. (2002). *Qué es el meta-análisis*. Madrid: Biblioteca Nueva.

Botella, J. & Gambara, H. (2006). Doing and reporting a meta-analysis. *International Journal of Clinical and Health Psychology, 6*(2), 425-440.

Botella, J. & Ortego, C. (in press). Compartir datos: hacia una investigación más sostenible. *Psicothema*.

*Burke, V., Milligan, R., Thompson, C., Taggart, A., Dunbar, D., Spencer, M., Medland, A., Gracey, M., Vandongen, R. & Beilin, L. (1998). A controlled trial of health promotion programs in 11-year-olds using physical activity "enrichment" for higher risk children. *The Journal of Pediatrics, 132*(5), 840-848.

Christensen, R., Kristensen, P. K, Bartels, E. M, Bliddal, H. & Astrup, A. (2007). Efficacy and safety of weight-loss drug rimonabant: a meta-analysis of randomised trial. *The Lancet, 370*(17), 1706-1713.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (2ª ed.). Hillsdale, NJ: Lawrence Erlbaum.

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

Cook, T. D. & Campbell, D. T. (1979). *Quasi-experimentation: design and analysis issues for field setting.* Rand McNally: Chicago.

Cooper, H. M. (1989). *Integrating research: a guide for literature reviews* (2nd ed.). Newbury Park, CA: Sage.

Dunlap W. P., Cortina J. M., Vaslow J. B., & Burke M. J. (1996). Meta-Analysis of experiments with matched groups or repeated measures. *Psychological Methods, 1*, 170-177.

Ettelson, R. G. (2002). *The treatment of adolescent depression*. Unpublished doctoral dissertation. Illinois State University.

Fisher, R. A, (1928). The general sampling distribution of the multiple correlation coefficient. *Proceeding of Royal Society of London, Series A, 121*, 654-673

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper, and L.V. Hedges (Eds.), *The Handbook of Research Synthesis* (pp. 245-260). New York: Sage.

Gibbons, R. D., Hedeker, D. & Davis, J. M. (1993). Estimation of effect size from a series of experiments involving paired comparisons. *J. Educat. Statist. 18*, 271-279.

Glaser, R. R (2002). *Accuracy of effect size calculation methods for repeated measures.and ANCOVA data*. Unpublished doctoral dissertation. Universidad of Memphis.

Glass, G. (1976). Primary, secondary and meta-analysis of research. *Educational Research, 5*, 3-8.

Hedges, L. V. (1981). Distribution theory for Glass's estimador of effect size and related estimators. *Journal of Statistics, 6*(2), 107-126.

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

Hedges, L. V. & Vevea, J. L. (1998). Fixed- and random-effects models in meta analysis. Psychological Methods, 3, 486-504.

*Hopper, C. A, Munoz, K. D., Gruber, M. D., MacConnie, S., Schonfeldt, B. & Shunk, T. (1996). A school-based cardiovascular exercise and nutrition program with parent participation: an evaluation study. *Children's Health Care, 25*(3), 221-235

Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement, 62,* 227-240.

Huedo, T. (2006). *Índices de tamaño del efecto para meta-analizar diseños pretest-postest con dos grupos: Un estudio de simulación Monte Carlo.* [Effect Size indices for meta-analysis with two-groups pretest-posttest designs: a Monte Carlo simulation study]. Unpublished Doctoral Dissertation. Universidad Nacional de Educación a Distancia (UNED), Spain.

Johnson B.**,** Mullen B. & Salas E**. (**1995**).** Comparisons of three major meta-analytic approaches**.** *Journal of Applied Psychology 80***,** 94–106

*Kain, J., Uauy, R., Albala, Vio F., Cerda R. and Leyton B. (2004). School-based obesity prevention in Chilean primary school children: methodology and evaluation of a controlled study. *International Journal of Obesity, 28,* 483-493.

Keef, S. P. & Roberts, L. A. (2004). The meta-analysis of partial effect sizes. *British Journal of Mathematical and Psychology, 57*, 97-129.

Keppel G. & Wickens, T. D. (2004). *Design and analysis: a researcher's handbook* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Kirk, R. E. (1995). *Experimental design. Procedures for the behavioural sciences* (3rd ed.). Belmont, CA: Brooks/Cole.

Lipsey, M. W. & Wilson, D. B. (2001). *Practical Meta-analysis*. Thousand Oaks, CA: Sage.

Marín. F. & Sánchez, J. (1996). Estimadores del tamaño del efecto en meta-análisis: Un estudio Monte Carlo del sesgo y la eficiencia. *Psicológica, 17*, 467-482.

Maxwell, S. E. & Delaney, H. D. (1990). *Designing experiments and analyzing data: a model comparison perspective.* Belmont, CA: Wadsworth.

*McMurray, R. G., Harrell. J. S., Bangdiwala, S. I., Bradley, C. B., Deng, S. & Levine, A. (2002) A school-based intervention can reduce body fat and blood pressure in young adolescents. *Journal of Adolescent Health, 31*(2), 125-132.

Miller G. M. & Chapman J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology, 110*, 40-48.

Morris, S. B. (2000). Distribution of the standardized mean change effect size for meta-analysis on repeated measures. *British Journal of Mathematical and Statistical Psychology, 53*, 17-29.

Morris, S.B. (2008). Estimating effect sizes from pretest-posttest- control group designs. *Organizational Research Methods, 11*(2), 364-386.

Morris, S. B. & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*(1), 105-125.

Nature (2006). Editorial. A fair share. 444. 653-654. Retrieved February 28, 2008, from: http://.nature.com/nature/journal/v444/n7120full/444653b.html

Neidig, J. L., Smith, B. A. & Brashers, D. E. (2003). Aerobic exercise training for depressive symptom management in adults living with HIV infection. *Journal of the Association of Nurses in AIDS Care, 14*(2), 30-40.

*Neumark-Sztainer, D., Story, M., Hannan, P. J., Stat, M. & Rex, J. (2003). New Moves: a school-based obesity prevention program for adolescent girls. *Preventive Medicine, 37*, 41-51.

Orwin, R. G. (1994). Evaluating coding decisions. In: Cooper y Hedges (eds.). The handbook of research synthesis. (pp.139-162). The Rusell Sage Foundation: New YorK

Peterson, R. A. & Brown, S. P. (2005). On the use of beta coefficients in meta-analysis. *Journal of Applied Psychology, 90*(1), 175-181.

Rohde, P., Clarke, G. N., Mace, D. E., Jorgensen, J. S & Seeley, J. (2004). An efficacy/effectiveness study of cognitive-behavioural treatment for adolescents with comorbid major depression and conduct disorder. *Journal of the American Academy of Child & Adolescent Psychiatry, 43,* 660-668

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (rev. ed.). Newbury Park, CA: Sage.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (eds), *The Handbook of Research Synthesis* (pp. 231-244). New York: Russell Sage Foundation.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin, 18*, 183-192.

Sánchez-Meca, J., Marín-Martínez, F. & Chacón-Moscoso, S. (2003). Effect-size índices for dichotomized outcomes in meta-analysis. *Psychological Methods, 8*(4), 448-467.

Scott-Sheldon, L. A. J., Kalichman S. C, Carey, M. P. & Fielder, R. L. (2008). Stress management interventions for HIV+ adults: A meta-analysis of randomized controlled trials, 1989 to 2006. *Health Psychology, 27*(2), 129-139

Shadish W. R. & Haddock C. K. (1994). Combining estimates of effect size. In: H. Cooper y L. V. Hedges (Eds.). *The Handbook of Research Synthesis* (pp.261-283). New-York: Sage.

Sikkema, K. J., Hansen, N. B, Kochman, A., Tate, D. C, & DiFranceisco, W. (2004). Outcomes from a randomized controlled trial of a group intervention for HIV positive men and women coping with AIDS-related loss and bereavement. *Death Studies, 28*, 187-209.

Stice, E., Shaw, H. & Marti, C. N. (2006). A meta-analytic review of obesity prevention programs for children and adolescents: the skinny on interventions that work. *Psychological Bulletin, 132*(5), 667-691.

Thompson, B. (2006). Research synthesis: Effect sizes. In: J.Green, G. Camilla y PB Elmore (eds). *Handbook of complementary methods in education research* (pp. 583-603). Washington, DC: American Educational Research Association.

Van Breukelen, G. J. P. (2006). ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. *Journal of Clinical Epidemiology, 59*, 920-925

Viechtbauer, W (2007). Approximating confidence intervals for standardized effect sizes in two-independent and two-dependent samples designs. *Journal of Educational and Behavioral Statistics, 32*, 39-60

Weiss, J. J., Mulder, C. L., Antoni, M. H., de Vroome, E. M., Garssen, B., & Goodkin, K. (2003). Effects of a supportive-expressive group intervention on long-term psychosocial adjustment in HIV-infected gay men. *Psychotherapy and Psychosomatics, 72*, 132-140.

Weisz, J. R., McCarty, C. A.& Valeri, S. M. (2006). Effects of Psychotherapy for depression in children and adolescents: a meta-analysis. *Psychological Bulletin, 132(*1), 132-149.

Wicherts, J. M., Borsboom, D., Kats, J. & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist, 61,* 726-728

WMA (2000). *Word Medical Association Declaration of Helsinki. Ethical Principles for Medical Research involving Human Subjects*. Edimburg: World medical Association General Assembly.

APPENDIX

Effect Size indices employed.

It is assumed a design with two groups, Experimental (E) and Control (C), measured in two occasions, pre-treatment (1) and post-treatment (2). Thus, the mean and the variance are $\overline{X}_{E1}$ and S^2_{E1} for the pre-treatment in the experimental group, whereas the subscripts change to C and/or to 2 for the other combinations of groups and occasions; N_E and N_C stand for the group sizes.

Order-zero indices

- *Standardized mean difference* (d_g), is the difference between the means in the post-treatment scores, divided by the common standard deviation, corrected for bias,

$$d_g = c(m_{d_g}) \cdot \frac{\overline{X}_{C2} - \overline{X}_{E2}}{S_{pooled}}$$

where S_{pooled} is the pooled standard deviation and $c(m_{d_g})$ is the well known correction factor for bias (Hedges, 1981, pg. 114),

$$c(m_{d_g}) \approx 1 - \frac{3}{4 \cdot (N_E + N_C - 2) - 1}$$

The (large sample approximation) variance of the index is (Hedges y Olkin, 1985, pg. 86, ec. 15),

$$S^2_{d_g} = \frac{N_E + N_C}{N_E \cdot N_C} + \frac{d_g^2}{2 \cdot (N_E + N_C)}$$

- *Pearson's correlation*, between the post-treatment scores in the dependent variable and the dichotomous codes for the group (Y),

$$r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y}$$

As it is usual, we will analyze it after the transformation to Fisher's Z (Fisher, 1928),

$$Z_r = \frac{1}{2} \log_e \frac{1 + r_{XY}}{1 - r_{XY}}$$

The (large sample approximation) variance of the index is (Shadish y Haddock, 1994, pg.268),

$$S^2_{Z_r} = \frac{1}{(N_E + N_C) - 3}$$

Order-one indices

- *Partialized standardized mean difference* (d_p), is the difference between the means, previously adjusted for differences between the pre-treatment scores (Keef y Roberts, 2004, pg. 100, ec.5),

$$d_p = c(m_{d_p}) \cdot \frac{\overline{X}^a_C - \overline{X}^a_E}{\sqrt{MCE^a}}$$

where the superscript *a* stands for *adjusted* and the correction is (Hedges, 1981, pg. 114),

$$c(m_{d_p}) \approx 1 - \frac{3}{4 \cdot (N - p - 1) - 1}$$

and its (large sample approximation) variance is (Keef y Roberts, 2004, pg. 100, ec.10),

$$S^2_{d_p} = \frac{d^2_p}{(N - p - 1) - 2} \cdot \left(\frac{N - p - 1}{t^2} + (N - p - 1) \cdot \left[c\left(m_{d_p}\right) \right]^2 - (N - p - 1) + 2 \right)$$

- *Difference between the standardized change scores* (d_c), is the difference observed between the changes in each group from the mean pre- to the mean post- scores, after standardizing each change by its own standard deviation (Becker, 1988, pg. 259),

$$d_c = d_{cC} - d_{cE}$$

For the experimental group (for the control group, the under script changes from E to C), d_{cE} stands for the standardized mean change in that group,

$$d_{cE} = c(m_{d_{cE}}) \cdot \frac{\overline{X}_{E1} - \overline{X}_{E2}}{S_{E1}}$$

The bias correction is (Hedges, 1981, pg. 114) ,

$$c(m_{d_{cE}}) \approx 1 - \frac{3}{[(4 \cdot (N_E - 1) - 1]}$$

and its (large sample approximation) variance is (Morris 2000, pg.21, ec.9),

$$S^2_{d_{cE}} = \left[c\left(m_{d_{cE}}\right)\right]^2 \cdot \left(\frac{2(1 - r_{E1E2})}{N_E}\right) \cdot \left(\frac{N_E - 1}{N_E - 3}\right) \cdot \left(1 + \frac{N_E}{2 \cdot (1 - r_{E1E2})} d^2_{cE}\right) - d^2_{cE}$$

The variance of the index is the sum of the variances (Morris, 2000, pg. 26, ec.16),

$$S^2_{d_c} = S^2_{d_{cE}} + S^2_{d_{cC}}$$

- *Partial correlation coefficient*, is the correlation between the post-scores and the dichotomous codes for the group (Y), partializing for the pre-treatment scores,

$$r_{Y2.1} = \frac{r_{Y2} - r_{Y1} \cdot r_{12}}{\sqrt{1 - r^2_{Y1}} \sqrt{1 - r^2_{12}}}$$

The Fisher's Z transformation is applied as in the bivariate case, and the (large sample approximation) variance is,

$$S^2_{Z_{r_p}} = \frac{1}{N - 3 - (p - 1)}$$

being *p* the number of predictors partialized (here, one).

(Manuscript received: 22 December 2008; accepted: 25 March 2009)