Calculating Effect Sizes for Designs with Between-Subjects and Within-Subjects Factors: Methods for Partially Reported Statistics in Meta-analysis

Dolores Albarracin
University of Illinois, Urbana-Champaign

Metodología de las Ciencias del Comportamiento 4(2), xx–xx

Calculating Effect Sizes for Designs with Between-Subjects and Within-Subjects Factors: Methods for Partially Reported Statistics in Meta-analysis

Paul Seignourel[1 2]
Dolores Abarracín
Department of Psychology, University of Florida

Abstract

Con el objetivo de resolver las deficiencias de las técnicas previas para calcular el tamaño del efecto sobre la base de datos insuficientes provenientes de diseños con factores entre- e intra-sujeto, los autores desarrollaron procedimientos para calcular tamaños del efecto a partir de reportes de medias, número de sujetos en cada celdilla y razones F. Específicamente, se derivan términos de error que se utilizan para calcular la desviación típica conjunta y la correlación y estos estadísticos se utilizan para calcular diferencias de medias estandarizadas. Se sugieren correcciones para tamaños del efecto obtenidos sobre la base de distinto número de factores y se extienden todos los procedimientos para diseños desequilibrados y de orden superior.
PALABRAS CLAVE: *Meta-análisis, Análisis de varianza, Diseños mixtos, Desviación típica, Correlación, Reporte incompleto de datos.*

Abstract

To resolve the deficiencies of prior techniques to compute effect sizes from insufficient data of designs with both between-subjects and within-subjects factors, the authors developed procedures to calculate effect sizes from reports of means, number of participants in each cell and $F-$ratios. Specifically, error terms are derived and used to estimate the pooled standard deviation and the correlation, and these statistics can be used to calculate standardized differences. Corrections are suggested for effect sizes obtained from designs with different numbers of factors and all procedures are extended for unbalanced and higher-order designs.
KEY WORDS: *Meta-analysis, Analysis of variance, Mixed designs, Standard deviation, Correlation, Incomplete data reports.*

There is no such thing as a correct study. The sources of error in each individual piece of research arise from the selection of the method, the sampling criteria, and various other factors that clutter the evidence that the study provides (Cooper & Hedges, 1994, Hunter & Schmidt, 1990). Researchers have been aware of this problem for a long time and often attempt to reach conclusions that generalize across the particulars of the individual studies (Olkin, 1990). For example, Pearson (1904) averaged correlations obtained in five samples and was able to estimate the typical effect of inoculation for typhoid fever. More recently, meta-analytic techniques have improved our understanding of the assumptions that underlie such integration of findings (see e.g., Shadish & Haddock, 1994). These synthesis methods allow scientists to discern whether a given phenomenon is robust across conditions, and also

[1]Correspondence concerning this paper should be addressed to Paul Seignourel or Dolores Albarracín, Department of Psychology, University of Florida, Gainesville, Florida 32611. Electronic mail can be sent to <seignour@ufl.edu> or <dalbarra@ufl.edu>

[2]The research was supported by grants from the National Institute of Mental Health (R03-MH58073 and K01-MH01861). We thank Lyle Brenner for conceptual contributions to the paper and G. Tarcan Kumkale, Blair T. Johnson and Alice H. Eagly for insightful comments and suggestions on an earlier version of this manuscript.

to identify between-study differences that moderate the emergence of the phenomenon.

To represent the effect of one variable on another, researchers often obtain a measure of the standardized mean difference d:

$$d = \frac{M_T - M_C}{SD},$$

where M_T and M_C are the means of the treatment and control groups, respectively, and SD is an estimate of the standard deviation. The estimate SD can be defined as (a) the pooled standard deviation of the two groups, (b) the standard deviation of either group under the assumption of homogeneity of variance, (c) the standard deviation of a control group if available, or (d) various "adjusted" standard deviations (see our section on corrected effect sizes). In this paper, we concentrated on S_{pooled}, the pooled standard deviation of the two groups, defined by

$$S_{pooled} = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}},$$

where n_1 and n_2 are the numbers of participants in groups 1 and 2, respectively, and SD_1 and SD_2 are the standard deviations of groups 1 and 2.

In the case of a within-subjects design,

$$d = \frac{M_2 - M_1}{SD},$$

where M_1 and M_2 are the means at Time 1 and Time 2, respectively, and SD is an estimate of the standard deviation. As in the case of between-subjects designs, we used S_{pooled} as an estimate of the standard deviation.[3]

Despite the simplicity of the standardized mean difference d, the statistic is difficult to compute when the studies report insufficient data. Despite these deficiencies, in the case of between-subjects designs, one can derive S_{pooled} from a single F−ratio, accompanied by the means and the numbers of participants in each cell (see Johnson, 1993). However, researchers who analyze data from a design with both between-subjects and within-subjects factors are expected to know the mean and the standard deviation in each group, as well as the correlation between the repeated measures (see e.g., Nouri & Greenberg, 1995; Cortina & Nouri, 2000; Johnson & Eagly, 2000). Unfortunately, the literature only rarely presents such detailed statistical reports (Cooper & Hedges, 1994; Johnson, 1993).

The present paper concerns a possible treatment of the problem of insufficient report in designs with between-subjects and within-subjects factors. The methods were developed in the context of a meta-analysis that Kumkale, Albarracín, and Seignourel (2001) conducted to determine attitude change maintenance and decay following exposure to a persuasive message. The researchers were interested in obtaining

[3]Dunlap, Cortina, Vaslow and Burke (1996) recommended using the pooled standard deviation to estimate effect sizes, regardless of whether estimates correspond to between- or within-subjects differences. However, Johnson and Eagly (2000) suggested using the standard deviation of the difference in the case of within-subjects designs. In our view, using S_{pooled} has the advantage of providing effect size indexes that are on the same metric as those calculated from between-subject designs. This property, in turn, is crucial for pooling effect sizes across different experimental designs. We nonetheless included the computation of S_{diff} in our programs, which can be accessed from www.psych.ufl.edu/˜albarrac/effectsizes.htm.

S_{pooled} to test change over time within each condition of interest, but some studies, especially those published in the sixties and seventies, failed to report standard deviations and correlations between repeated measures (e.g., Johnson & Watkins, 1971; Watts & McGuire, 1964; Weber, 1971). However, the means, the number of participants in each cell, and at least two F-ratios were almost always reported. Instead of discarding these studies, the researchers developed methods to calculate the pooled standard deviation, an estimate of the correlation and the effect sizes of interest, on the basis of the available information.[4]

This paper describes the procedures we derived to calculate effect size estimates from reports of means, number of participants and F-ratios. However, pooling statistical data of any sort can provide biased estimates if the primary data violated the statistical assumptions that underlie the correct use of a statistical procedure. In the case of analysis of variance designs with within-subjects factors, homogeneity of variance, normality and compound symmetry should be verified before reporting results. To this extent, the procedures we derived are useful only if one can select quality reports and if the primary authors took steps to guarantee the scientific and statistical integrity of their data.

Two-Factor Experiments with Repeated Measures on One Factor[5]

In certain designs, some, but not all of the variables are within-subjects factors. Imagine that a researcher conducts an experiment to investigate differential decay in attitude change among males and females. There are two factors in this design. The first one (between-subjects), which we denote as factor A, distinguishes between males and females. The second, within-subjects factor, B, corresponds to participants' reports of their attitudes at two points in time. This design appears in Table 1 and can be represented by the model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \rho_{k(i)} + \epsilon_{ijk}, \tag{1}$$

where Y_{ijk} is an observation; $i = 1, 2$ are the two levels of factor A; $j = 1, 2$ are the two levels of factor B; μ is the overall mean; α_i is the mean of group i deviated from the overall mean μ; β_j is the mean of measure j deviated from the overall mean μ; $(\alpha\beta)_{ij}$ is the mean of measure j for group i, minus the mean for group i, minus the mean for measure j, plus the overall mean μ; and the $\rho_{k(i)}$ and ϵ_{ijk} are error terms independent of each other, independent across observations, and distributed $\mathcal{N}(0, \sigma_p^2)$ and $\mathcal{N}(0, \sigma^2)$, respectively. The $\rho_{k(i)}$ correspond to the independent effects of each participant, which are constant over the different measures. The ϵ_{ijk}, instead, are different for each participant and for each measure.[6]

---

[4] For programs associated with these methods, see www.psych.ufl.edu/˜albarrac/effectsizes.htm.

[5] For all notations and definitions in the following section, we refer to the presentation of two-factor designs with repeated measures on one factor given by Neter et al. (1990). Note that, in our model, the between- and within-subjects variables are both considered to be fixed effects.

[6] Some readers may notice that the term $\beta_{jk(i)}$ introduced by Winer (1971) is missing in equation (1). However, because there is only one observation per participant and per measure in this model, the additional term $\beta_{jk(i)}$ is confounded with ϵ_{ijk}. That is, given a set of data, it would be impossible to distinguish between the two terms, and only their sum could be uniquely identified. We therefore refer to the model and notations found in Neter et al. (1990), which we find more parsimonious. However, the same derivations, with slight modifications, could be obtained using the model introduced by Winer.

Table 1. A 2 x 2 Design with Repeated Measures on One Factor

	Factor *B* (Within-subjects)	
Factor *A* (Between-subjects)	Time 1	Time 2
Males	$\overline{Y}_{11\cdot}$	$\overline{Y}_{12\cdot}$
Females	$\overline{Y}_{21\cdot}$	$\overline{Y}_{22\cdot}$

The sum of the variance of $\rho_{k(i)}$ and the variance of ϵ_{ijk} provides the variance of the observation Y_{ijk}:

$$\sigma\{Y_{ijk}\} = \sqrt{\sigma_p^2 + \sigma^2} \tag{2}$$

As mentioned before, in this paper we assume homogeneity of variance, and we focus on S_{pooled} as an estimate of $\sigma\{Y_{ijk}\}$. Therefore, for the design in Table 1, the estimate of the difference between Time 1 and Time 2 among males, for example, is

$$g = \frac{\overline{Y}_{12\cdot} - \overline{Y}_{11\cdot}}{S_{pooled}}, \tag{3}$$

where $\overline{Y}_{11\cdot}$ and $\overline{Y}_{12\cdot}$ are the means for males at Time 1 and Time 2, respectively.[7] Likewise, the estimate of the difference between females and males at Time 1 is

$$g = \frac{\overline{Y}_{21\cdot} - \overline{Y}_{11\cdot}}{S_{pooled}}. \tag{4}$$

Computing the Pooled Standard Deviation and the Correlation in a 2 × 2 Design

We first analyzed a sample of 20 males and 20 females who participated in Project RESPECT, a multi-site study funded by CDC. Project RESPECT (Kamb et al., 1998; Kamb, Dillon, Fishbein, Willis, & Project RESPECT Study Group, 1996) was a randomized control trial comparing three separate face-to-face HIV (Human Immunodeficiency Virus)/STD (Sexually Transmitted Disease) prevention interventions (i.e., educational messages typical of current practice, brief counseling and enhanced counseling) with approximately 1,500 participants in each. The project was longitudinal and measures of behavioral and psychosocial variables were obtained at the (a) baseline, (b) immediate follow-up, (c) 3-month follow-up, (d) 6-month follow-up, (e) 9-month follow-up, and (f) 12-month follow-up. For our first analysis, we restricted consideration to participants' reports of their attitudes towards condom use at baseline and the immediate follow-up (within-subjects factor) across the two gender groups (between-subjects factor).

The descriptive statistics and F−ratios for these 40 participants appear in Tables 2 and 3. Using the standard deviations reported in Table 2, we estimated the pooled standard deviation for this design at 1.24. We used the raw data to calculate the correlation between attitudes at Time 1 and 2, which were $r = .45$ and $r = .54$ for males and females, respectively. The pooled correlation was therefore .50.

[7] With respect to equation (1), $\overline{Y}_{12\cdot}$ is an unbiased estimate of $(\alpha\beta)_{12} + \alpha_1 + \beta_2 + \mu$, while $\overline{Y}_{11\cdot}$ is an unbiased estimate of $(\alpha\beta)_{11} + \alpha_1 + \beta_1 + \mu$. Therefore, the numerator of equation (3) is an unbiased estimate of $(\alpha\beta)_{12} - (\alpha\beta)_{11} + \beta_2 - \beta_1$. Likewise, the numerator of equation (4) is an unbiased estimate of $(\alpha\beta)_{21} - (\alpha\beta)_{11} + \alpha_2 - \alpha_1$.

Table 2. Descriptive Statistics for Attitudes at Time 1 and 2

	Attitude 1		Attitude 2	
Gender	*M*	*SD*	*M*	*SD*
Females	0.364	1.374	1.536	1.389
Males	1.750	1.243	2.371	0.905

Table 3. Analysis of Variance for Attitude Change at Time 1 and 2

Source	*df*	*F*
Between-subjects		
Gender (*A*)	1	10.84
Within-subjects		
Time (*B*)	1	19.80
A x *B*	1	1.86

Imagine, however, that the researchers presented only the means, the sample size and the F–ratios. In that case, meta-analysts who wanted to calculate S_{pooled} must first compute various sums of squares using the means provided in Table 2, and then use the F–ratios to derive the standard deviation. Following that procedure,

$$MSA = 2n\sum_i(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot})^2 = 40 * (.555^2 + .555^2) = 24.7, \quad (5)$$

$$MSB = 2n\sum_j(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdot\cdot\cdot})^2 = 40 * (.448^2 + .448^2) = 16.1, \quad (6)$$

and

$$\begin{aligned} MSAB &= n\sum_{i,j}(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdot\cdot\cdot})^2 \\ &= 20 * (.138^2 + .138^2 + .138^2 + .138^2) = 1.5 \end{aligned} \quad (7)$$

We can then derive the error terms associated with this design. First,

$$\begin{aligned} MSS(A) &= \frac{1}{n-1}\sum_{i,k}(\overline{Y}_{i\cdot k} - \overline{Y}_{i\cdot\cdot})^2 \\ &= \frac{MSA}{F_A} = \frac{24.7}{10.8} = 2.28 \end{aligned} \quad (8)$$

gives us the mean square error associated with the participants (often referred to as between-subjects error variance). We can also calculate:

$$MSB.S(A) = \frac{1}{2(n-1)}\sum_{i,j,k}(\overline{Y}_{ijk} - \overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot k} + \overline{Y}_{i\cdot\cdot})^2, \quad (9)$$

often referred to as within-subjects error variance, by using F_B, or, alternatively, F_{AB}. That is,

$$MSB.S(A) = \frac{MSB}{F_B} = \frac{16.1}{19.8} = 0.81, \quad (10)$$

or

$$MSB.S(A) = \frac{MSAB}{F_{AB}} = \frac{1.5}{1.9} = 0.81. \tag{11}$$

Once we obtain $MSS(A)$ and $MSB.S(A)$, we are only one step away from obtaining S_{pooled}, because $MSS(A)$ and $MSB.S(A)$ are estimates of $\sigma^2 + 2\sigma_p^2$ and σ^2, respectively (see Neter, Wasserman, & Kutner, 1990). Therefore, an estimate of the pooled standard deviation is:

$$S_{pooled} = \sqrt{\frac{MSS(A) + MSB.S(A)}{2}} = 1.24, \tag{12}$$

which is equal to the value we computed by pooling the standard deviations in Table 2.

The correlation between the within-subjects measures can be derived with equally-straightforward procedures.[8] From Equation (1),

$$\text{Cov}\{Y_{ijk}, Y_{ij'k}\} = \sigma_p^2, \tag{13}$$

for every i, k and $j \neq j'$. Hence, the correlation between attitude at Time 1 and attitude at Time 2 is given by

$$\frac{\sigma_p^2}{\sigma_p^2 + \sigma^2} = \frac{\text{Cov}\{Y_{ijk}, Y_{ij'k}\}}{\sigma^2\{Y_{ijk}\}} \tag{14}$$

Therefore, an estimate of the correlation is

$$r = \frac{MSS(A) - MSB.S(A)}{MSS(A) + MSB.S(A)} = \frac{2.28 - 0.81}{2.28 + 0.81} = 0.47 \tag{15}$$

Note that the value .47 is slightly different from the pooled correlation .50 that we computed directly from the raw data. In order to show why r and r_{pooled} differ in general, we related their expressions to the variances and covariances of the different cells and groups in the sample (see Appendix B). As a consequence, we also showed that their values are identical whenever the different variances, as well as the different covariances, are equal in the sample. Even though this condition is unlikely to be met in any given data set, this result shows that we can expect the two estimates r and r_{pooled} to be close when the variances and covariances do not differ by much in the sample.

[8] Although we do not use the correlation in computing our effect size estimates, the correlation is necessary to calculate the sampling variance of the estimates (see e.g., Becker, 1988; Morris, 2000), which is required for many meta-analytic procedures. We should point out, however, that the expression found in Becker requires an estimate of the pre-test standard deviation, which is unavailable in the situation we describe in this paper. Future research is needed to determine the sampling distribution of the effect size estimates when using S_{pooled}, as proposed here.

Correcting Effect Sizes

As pointed out by several authors (see e.g., Morris & DeShon, 1997), a critical issue in meta-analysis is to derive comparable estimates from different studies using different statistical procedures. In certain cases, this concern makes necessary the calculation of a corrected effect size instead of the usual raw estimates.[9] There are two types of corrections that are relevant in the context of a two-factor design with repeated measures on one factor.

First, suppose that a meta-analyst wants to determine if a given counseling method produced change over time and has two studies with relevant information. The first report is a 2 x 2 design like the one in Table 2, with gender as a between-subjects variable and time as the within-subjects factor. In the second study, however, the authors did not separate males and females, and the design is a simple within-subjects design with two measures of attitude.

Suppose further that the meta-analyst knows the means, the number of participants and the F−ratios of the 2 x 2 design. The researcher can compute S_{pooled} for the first study, average the means across genders, and consequently, obtain an estimate of the effect size:

$$g = \frac{M_2 - M_1}{S_{pooled}}. \tag{16}$$

Unfortunately, however, this estimate may not be comparable to the estimated effect size for the second study. If men and women scored differently on the attitude scale, splitting the participants according to gender is likely to reduce the standard deviation in the first, 2 x 2 design, compared to the one factor within-subjects design of the second study (see Cortina & Nouri, 2000, Morris & DeShon, 1997, on the issue of corrected effect sizes). To overcome this difficulty, one can correct the pooled standard deviation obtained from the 2 x 2 design. As proposed by Glass, McGaw and Smith (1981), in the general case of a a x b design:

$$S_{corrected} = \sqrt{\frac{SSA + SSAB + SSS(A) + SSB.S(A)}{df_A + df_{AB} + df_{S(A)} + df_{B.S(A)}}}. \tag{17}$$

Given the reconstruction of the various sums of squares we described before (see Equations 5-11), the computation of the corrected standard deviation is a simple extension of our previous calculations.[10]

The second corrected effect size associated with the design in Table 2 corresponds to the difference between males and females, independent of time of measurement. In this case, the solution proposed by Nouri and Greenberg (1995) is to sum the attitudes at Time 1 and 2 and compare:

$$\overline{Y}_{1\cdot\cdot} = \sum_{j,k} Y_{1jk} \tag{18}$$

[9]In other cases, however, the effect sizes should not be corrected, depending on whether or not the off-factor varies in the population of interest (see Cortina, J. M., & Nouri, H., 2000). In the following example, gender is the off-factor, and it is therefore appropriate to correct the effect size.

[10]As pointed out by Morris and DeShon (1997) in the case of a between-subjects design, the corrected standard deviation is not a new estimate. It is the S_{pooled} that the experimenter would have found had he used a one-factor within-subjects design to analyze the raw data.

with

$$\overline{Y}_{2\cdot\cdot} = \sum_{j,k} Y_{2jk}. \tag{19}$$

According to Ghiselli (1964), the appropriate standard deviation for estimating this effect size is:

$$S_{corrected} = \sqrt{2(1 + r_{pooled})S^2_{pooled}}. \tag{20}$$

However, when the $F-$ratios are available, it may not be necessary to derive $S_{corrected}$. The reason is that F_A is equal to the F-value obtained in the two-cell between-subjects design. Therefore, the usual formula,

$$g = \sqrt{\frac{2F_A}{n}}, \tag{21}$$

applies.

Extending the Method to Unbalanced Designs

In the previous sections, we assumed an equal number of participants in each group. This assumption, however, is seldom met in the literature. Therefore, it is necessary to extend our procedures to unbalanced designs analyzed with unweighted means analysis of variance procedures.[11] In these procedures, the researcher treats the means of each cell as single observations in each treatment. The estimated variance is then multiplied by the coefficient:

$$\alpha = \frac{1}{ab} \sum_{i,j} \frac{1}{n_{ij}}, \tag{22}$$

where a is the number of groups, b is the number of repeated measures, and n_{ij} is the number of participants in group i for the measure j. This new variance can be used to calculate all the F-ratios in the design (for further details, see Neter et al., 1990). Conversely, when researchers want to obtain effect sizes from a report of means and $F-$ratios from an unbalanced design, they can compute the sums of squares and an estimate of the variance of the sums of squares, and then divide this estimate by the coefficient α to obtain S_{pooled}. Computing the correlation r involves the same steps.

Correcting effect sizes for unbalanced designs follows the logic described before. If one needs to average the results of different groups to obtain the within-subjects effect sizes, it is necessary to compute the standard deviation for each repeated measure. As shown in Nouri and Greenberg (1995), the corrected standard deviation for the measure j is given by

$$S_j = \sqrt{\frac{1}{n_j - 1}\left((n_j - b)S^2_{pooled} + \sum_i n_{ij}(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot})^2\right)}, \tag{23}$$

[11] When confronted with unbalanced designs, modern software usually use least square methods, which differ from the unweighted means solutions for variables with more than two levels. However, standard deviations are usually reported in recent publications. Therefore, our method applies mainly to reports published during the sixties and seventies, when the unweighted means methods were frequently used. The reader should of course be careful when applying our method to unbalanced designs with more than two levels in one of the variables.

where n_j is the number of participants for the measure j, and the other terms were already defined.

Extending the Method to Higher-Order Designs

In the preceding sections, we focused our attention on 2 x 2 designs. However, our method generalizes to more complex designs as well. For an a x b design, $MSS(A)$ is an estimate of $\sigma^2 + b\sigma_p^2$, whereas $MSB.S(A)$ is still an estimate of σ^2. Thus, our new estimate of the variance is:

$$S_{pooled} = \sqrt{\frac{MSS(A) + (b-1)MSB.S(A)}{b}}. \quad (24)$$

Accordingly, our estimate of the correlation is:

$$r = \frac{MSS(A) - MSB.S(A)}{MSS(A) + (b-1)MSB.S(A)}. \quad (25)$$

For designs with one within-subjects variable and two or more between-subjects variables, the method remains essentially the same (for a presentation of this type of design, see Myers, 1979).[12] The within-subjects error variance is always an estimate of σ^2, while the between-subjects error variance is an estimate of $\sigma^2 + t\sigma_p^2$, where t is equal to the number of repeated measures for each participant. Consider, for example, an a x b x t design, where the two first factors (A and B) are between-subjects variables and the last one (T) is a within-subjects variable. In this case, $MST.S(AB)$ is an estimate of $\sigma^2 + t\sigma_p^2$. Thus, the estimates of the variance and correlation are, respectively:

$$S_{pooled} = \sqrt{\frac{MSS(AB) + (t-1)MST.S(AB)}{t}} \quad (26)$$

and

$$r = \frac{MSS(AB) - MST.S(AB)}{MSS(AB) + (t-1)MST.S(AB)}. \quad (27)$$

As before, we verified these procedures on the data of the Project RESPECT (Kamb et al., 1996). This time, however, we used all relevant participants in the sample (N = 624), introduced type of intervention (educational messages vs. brief counseling) as a new between-subjects variable, and used three measures of attitudes (3-, 6- and 9-month follow-ups) as our within-subjects factor. Thus, we had a three-factor design (2 x 2 x 3) with repeated measures on the last factor, and with unequal numbers of participants in the different groups (125, 158, 176 and 165 participants for the four groups, respectively). The descriptive statistics and F−ratios for this data set are provided in Tables 4 and 5. Pooling the 12 values provided in Table 4, we found S_{pooled} = 1.52. We also conducted bivariate analysis and found that the pooled correlation for this sample was r_{pooled} = .74.

[12] The method is slightly different, however, for designs with more than one within-subjects variable. Even though we do not present the modifications for this type of design in this paper, programs to compute S_{pooled} for designs with two levels of within-subjects variables are available at www.psych.ufl.edu/˜albarrac/effectsizes.htm.

Table 4. Descriptive Statistics for Attitudes at Time 3, 4 and 5

	Attitude 3		Attitude 4		Attitude 5	
Type of Intervention	*M*	*SD*	*M*	*SD*	*M*	*SD*
Educational messages						
Females (n = 125)	.870	1.670	.955	1.636	.907	1.695
Males (n = 158)	1.627	1.450	1.720	1.347	1.616	1.422
Brief Counseling						
Females (n = 176)	.823	1.610	.820	1.590	.895	1.641
Males (n = 165)	1.697	1.359	1.731	1.387	1.565	1.503

Table 5. Analysis of Variance for Attitude Change at Time 3, 4 and 5

Source	df	F
Between-subjects		
Intervention (A)	1	.06
Gender (B)	1	49.01
A x B	1	.11
Within-subjects		
Time (T)	2	1.06
T x A	2	.34
T x B	2	1.59
T x A x B	2	.61

Given the $F-$ratios provided in Table 5, several possibilities are offered to the meta-analyst for computing the missing standard deviation and correlation. In fact, any choice of one between-subjects $F-$ratio and one within-subjects $F-$ratio (a total of 12 possible combinations) would lead to an estimate of S_{pooled} and r.[13] In the following, we show our computations using F_B and F_{AT}, but any other choice would yield the same results.

First, one must calculate the coefficient α introduced in equation (22). We have:

$$\alpha = \frac{1}{4}\left(\frac{1}{125} + \frac{1}{158} + \frac{1}{176} + \frac{1}{165}\right) = .00652 \qquad (28)$$

Now, we need to compute the sums of squares associated with the given (or chosen) $F-$ratios:

$$MSB = \frac{6}{\alpha} * \sum_j (\overline{Y}_{\cdot j \cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot\cdot})^2 = \frac{6}{.00652} * (.391^2 + .391^2) = 280.7 \qquad (29)$$

$$\begin{aligned} MSAT &= \frac{1}{\alpha} \sum_{i,k} (\overline{Y}_{i \cdot k \cdot} - \overline{Y}_{i \cdot\cdot\cdot} - \overline{Y}_{\cdot\cdot k \cdot} + \overline{Y}_{\cdot\cdot\cdot\cdot})^2 \\ &= \frac{1}{.00652} * (.0195^2 + .0173^2 + .0021^2 + .0194^2 + .0174^2 + .0021^2) \\ &= .209 \end{aligned} \qquad (30)$$

[13]Note that, in this type of design, the meta-analyst still needs only two $F-$values, one corresponding to a within-subjects effect and the other to a between-subjects effect. However, several $F-$values of each type might be available, and these values may yield different estimates of the standard deviation and correlation. All these values, however, should be equal, with the exception of rounding error. If they are not, the discrepancy is probably due to misreport.

Using the $F-$ratios, we derive the between-subjects and within-subjects error variances:

$$MSS(AB) = \frac{MSB}{F_B} = \frac{280.7}{49.01} = 5.727 \quad (31)$$

$$MST.S(AB) = \frac{MSAT}{F_{AT}} = \frac{.209}{.34} = .615 \quad (32)$$

Finally, using equations (26) and (27), we compute:

$$S_{pooled} = \sqrt{\frac{5.727 + 2 * 0.615}{3}} = 1.52 \quad (33)$$

$$r = \frac{5.727 - 0.615}{5.727 + 2 * 0.615} = .73 \quad (34)$$

Consistent with the result of Appendix A, our estimate S_{pooled} is equal to the one derived from the raw data, while $r = .73$ is very close to the pooled standard deviation $r_{pooled} = .74$ computed from the raw data.

Discussion

Meta-analysis enables researchers to average effect sizes and draw inferences about the generalizability of a given phenomenon. For example, one can study the effects of persuasive communications over time and reach conclusions concerning change or stability as a result of the interventions (see e.g., Bukoski, 1997). Up to now, however, the available techniques to calculate effect sizes from designs with both between-subjects and within-subjects factors were limited, because they required information on the correlation between the repeated measures in the design, as well as the mean, number of participants and standard deviation in each cell (see e.g., Becker, 1988; Johnson & Eagly, 2000; Hedges & Olkin, 1985; Nouri & Greenberg, 1995). In contrast, the methods we described in this paper suggest ways of calculating effect sizes on the sole basis of the means, number of participants and two $F-$ratios.

Under most conditions, the estimates we described generate values that are identical to the ones generated from the primary data of a study (see Appendix A and B). Thus, as we showed, the standard deviation we calculated for the Project RESPECT yielded 1.24 regardless of whether we used the primary data or the report of means, number of participants and two $F-$ratios (see Tables 2 and 3). Furthermore, the estimate of the correlation between the two longitudinal measures we considered also yielded similar estimates ($r = .47 \approx .50$).

An important question is whether the techniques we described are robust across conditions. With respect to unbalance, when primary researchers only report the total number of participants in their study, we recommend that the meta-analyst assume a balanced design and simply divide the overall N in the design by the number of cells. Of course, this rule of thumb might introduce a bias in the effect size estimates. However, note that the number of participants in each cell is only relevant for calculating the coefficient α defined in equation (22). The estimates of S_{pooled} and r are obtained by multiplying estimates based on the means and the F-ratios by the square root of α. In the case of a design with two groups and two repeated measures, we show in Appendix C that assuming a balanced design leads to underestimating S_{pooled} and r by approximately $2e^2$, where e is a measure of the deviation from a balanced design, and $2e^2$ is an error in proportion to the estimates of S_{pooled} and r.[14] For example, mistakenly assuming $n = 20$ in a study with two groups of 15 and 25 participants respectively, would introduce a bias of approximately 3% in the estimates of S_{pooled} and r. Note that, even when the difference in the number of participants is important (15 vs. 25), the resulting bias is small, regardless of the means and F-ratios. For this reason, assuming a balanced design when only the total number of participants is available is unlikely to introduce a substantive bias in our estimates. However, if one has no information concerning the total number of participants in the study, then our method does not allow for an estimate of either S_{pooled} or r.

Concerning correction of effect sizes, the procedures we presented share the same properties of similar corrections suggested by Glass and his colleagues (1981; see also Morris & DeShon, 1997). Thus, it is not that one effect size is more correct than the other, but that researchers should use effect sizes that are on the same metric (Morris & DeShon). Furthermore, the correction is only valid to the extent that $F-$ratios are unbiased; consequently, heterogeneity of variance and non-normalty can produce biased $F-$ratios (Wilcox, 1993). In any event, interest in meta-analysis has increased greatly in the last few years and researchers need methodologies that allow them to deal with the reality of insufficient data reports. The procedures we reported in this paper may help to alleviate the lack of technical resources for investigators who wish to calculate effect sizes in the context of designs with between-subjects and within-subjects factors.

Two possible extensions to the present work should be considered as directions for future research. First, as we have proved in Appendix B, the estimate r derived with the method presented in this paper and the classical estimate r_{pooled} are identical only when the various variances and correlations are equal in the different cells and groups of the sample. It would be important to estimate how discrepancies in sample variances and correlations affect the estimates r and r_{pooled}. Secondly, on the basis of the information available in psychological reports, one cannot be sure that the primary researchers have taken steps to ensure the validity of conducting an analysis of variance. Therefore, estimating the extent to which a violation of the assumptions underlying such designs (i.e., homogeneity of variance, compound symmetry and normality) affects our effect size estimates would contribute to our current set of meta-analytic tools.

References

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

Bukoski, W. J. (1997). *Meta-analysis of drug abuse prevention programs.* National Institute on Drug Abuse: Research Monograph 170.

Cooper, H., & Hedges, L. V. (1994). Research synthesis as a scientific enterprise. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261-281). New York: Russell Sage Foundation.

[14] Similar, although more complicated, formulas might be derived for more complex designs. In any case, the bias is independent of the means and $F-$ratios, and the error can always be calculated directly by comparing the real value of α with its value when assuming a balanced design.

Cortina, J.M., & Nouri, H. (2000). *Effect sizes for ANOVA designs.* Thousand Oaks, CA: Sage.

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

Ghiselli, E. E. (1964). *Theory of psychological measurement.* New York: McGraw-Hill.

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

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

Holmes, M. H. (1995). *Introduction to Perturbation Methods.* New York: Springer.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Cumulating research findings across studies.* Beverly Hills, CA: Sage.

Johnson, B. T. (1993). *DSTAT 1.10: Software for the meta-analytic review of research literatures.* Hillsdale, NJ: Erlbaum.

Johnson, B. T., & Eagly, A. H. (2000). Quantitative synthesis of social psychological research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social psychology* (pp. 496-528). London: Cambridge University Press.

Johnson, H. H., & Watkins, T. A. (1971). The effects of message repetitions on immediate and delayed attitude change. *Psychonomic Science, 22,* 101-103.

Kamb, M. L., Dillon, B. A., Fishbein, M., Willis, K. L., & Project RESPECT Study Group. (1996). Quality assurance of HIV prevention counseling in a multi-center randomized controlled trial. *Public Health Reports, 111,* 99-107.

Kamb, M. L., Fishbein, M., Douglas, J., Jr., Rhodes, F., Rogers, J., Bolan, G., Zenilman, J., Hoxworth, T., Malotte, K., Iatesta, M., Kent, C., Lentz, A., Graziano, S., Byers, R., Peterman, T., & The Project RESPECT Study Group. (1998). Efficacy of risk reduction counseling to prevent human immunodeficiency and sexually transmitted diseases: A randomized controlled trial. *Journal of the American Medical Association, 280,* 1161-1167.

Kumkale, T., AlbarracÝn, D., & Seignourel, P. (2001). *Meta-analysis of the sleeper effect in persuasion.* Gainesville, FL: Manuscript in preparation.

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., & DeShon, R. P. (1997). Correcting effect sizes computed from factorial analysis of variance for use in a meta-analysis. *Psychological Methods, 2,* 192-199.

Myers, J. L. (1979). *Fundamentals of experimental designs.* Boston: Allyn-Bacon.

Neter, J., Wasserman, W., & Kutner, M. H. (1990). *Applied linear statistical models.* Homewood, IL: Irwin.

Nouri, H., & Greenberg, R. H. (1995). Meta-analytic procedures for estimation of effect sizes in experiments using complex analysis of variance. *Journal of Management, 21,* 801-812.

Olkin, I. (1990). History and goals. In K. W. Wachter & M. L. Straf (Eds.), *The future of meta-analysis* (pp. 3-10). New York: Russell Sage Foundation.

Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *British Medical Journal, 3,* 1243-1246.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261-281). New York: Russell Sage Foundation.

Watts, W. A., & McGuire, W. J. (1964). Persistence of induced opinion change and retention of the inducing message contents. *Journal of Abnormal and Social Psychology, 68,* 233-241.

Weber, S. J. (1971). Source primacy-recency effects and the sleeper effect. *Proceedings of the Annual Convention of the American Psychological Association, 6,* 385-386.

Wilcox, R. (1993). Robustness in ANOVA. In L. Edwards (Eds.), *Applied analysis of variance in behavioral science* (pp. 354-374). New York: Dekker.

Winer, B. J. (1971). *Statistical principles in experimental design.* New York: McGraw-Hill

Appendix A
Equivalence with the Classical Estimate

When presenting our method for estimating the variance, we claimed that in the case of a 2 x 2 design with repeated measures on one factor,

$$S^2_{pooled} = \frac{MSS(A) + MSB.S(A)}{2} \tag{35}$$

is the usual estimate of the pooled variance. Actually, this result is true for any kind of design with both between-subjects and within-subjects factors, with the modification

$$S^2_{pooled} = \frac{MSS(A) + (T-1)MSB.S(A)}{T}, \tag{36}$$

where T is the number of repeated measures for each participant. Here, we prove this result for an a x b design, where b is the number of repeated measures, and the design is assumed to be balanced. The demonstration is essentially the same for unbalanced or more complex designs.

For an a x b design, using notations from Neter et al. (1990), the classical estimate of the variance is given by

$$\Delta = \frac{1}{ab(n-1)} \sum_i \sum_j \sum_k (Y_{ijk} - \overline{Y}_{ij\cdot})^2. \tag{37}$$

Then,

$$\begin{aligned} MSS(A) &= \frac{b}{a(n-1)} \sum_i \sum_k (\overline{Y}_{i\cdot k} - \overline{Y}_{i\cdot\cdot})^2 \\ &= \frac{1}{a(n-1)} \sum_i \sum_j \sum_k (\overline{Y}_{i\cdot k} - \overline{Y}_{i\cdot\cdot})^2. \end{aligned} \tag{38}$$

Hence,

$$\begin{aligned} & a(n-1)(MSS(A) + (b-1)MSB.S(A)) \\ &= \sum_{i,j,k} \Big((\overline{Y}_{i\cdot k} - \overline{Y}_{i\cdot\cdot})^2 + (Y_{ijk} - \overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot k} + \overline{Y}_{i\cdot\cdot})^2 \Big) \\ &= \sum_{i,j,k} \Big(2(\overline{Y}_{i\cdot k} - \overline{Y}_{i\cdot\cdot})^2 + (Y_{ijk} - \overline{Y}_{ij\cdot})^2 - 2(Y_{ijk} - \overline{Y}_{ij\cdot})(\overline{Y}_{i\cdot k} - \overline{Y}_{i\cdot\cdot}) \Big) \\ &= \sum_{i,j,k} \Big((Y_{ijk} - \overline{Y}_{ij\cdot})^2 + 2(\overline{Y}_{i\cdot k} - \overline{Y}_{i\cdot\cdot})(\overline{Y}_{i\cdot k} - \overline{Y}_{i\cdot\cdot} - Y_{ijk} + \overline{Y}_{ij\cdot}) \Big). \end{aligned} \tag{39}$$

Now, note that

$$\begin{aligned} & \sum_j (\overline{Y}_{i\cdot k} - \overline{Y}_{i\cdot\cdot})(\overline{Y}_{i\cdot k} - \overline{Y}_{i\cdot\cdot} - Y_{ijk} + \overline{Y}_{ij\cdot}) \\ &= (\overline{Y}_{i\cdot k} - \overline{Y}_{i\cdot\cdot}) \left(b\overline{Y}_{i\cdot k} - b\overline{Y}_{i\cdot\cdot} - \sum_j Y_{ijk} + \sum_j \overline{Y}_{ij\cdot} \right) \\ &= 0, \end{aligned} \tag{40}$$

which proves the result.

Appendix B

Difference Between the Estimates r and r_{pooled}

It is important to understand why the two estimates r and r_{pooled} are in general different. For that purpose, we relate their expressions to the variances and covariances in the sample in the case of a 2 x 2 design. The results, however, inmediately generalize to higher order designs.

For every i and j, COV_i denotes the covariance in the sample for group i, and V_{ij} denotes the variance in the sample for group i and measure j:

$$COV_i = \frac{1}{n-1}\sum_k (Y_{i1k} - \overline{Y}_{i1\cdot})(Y_{i2k} - \overline{Y}_{i2\cdot}) \tag{41}$$

$$V_{ij} = \frac{1}{n-1}\sum_k (Y_{ijk} - \overline{Y}_{ij\cdot})^2 \tag{42}$$

In order to evaluate r, as defined in equation (15), we will first find an expression for the numerator $MSS(A) - MSB.S(A)$. We have

$$\begin{aligned} & MSS(A) - MSB.S(A) \\ = & \frac{1}{2(n-1)}\sum_{ijk}\left((\overline{Y}_{i\cdot k} - \overline{Y}_{i\cdot\cdot})^2 - (Y_{ijk} - \overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot k} + \overline{Y}_{i\cdot\cdot})^2\right) \\ = & \frac{1}{2(n-1)}\sum_{ijk}(Y_{ijk} - \overline{Y}_{ij\cdot})(2\overline{Y}_{i\cdot k} - Y_{ijk} + \overline{Y}_{ij\cdot} - 2\overline{Y}_{i\cdot\cdot}) \end{aligned} \tag{43}$$

Using the fact that

$$\overline{Y}_{i\cdot k} = \frac{Y_{i1k} + Y_{i2k}}{2} \tag{44}$$

$$\overline{Y}_{i\cdot\cdot} = \frac{\overline{Y}_{i1\cdot} + \overline{Y}_{i2\cdot}}{2}, \tag{45}$$

equation (43) becomes

$$\begin{aligned} MSS(A) - MSB.S(A) &= \frac{1}{n-1}\sum_{ik}(Y_{i1k} - \overline{Y}_{i1\cdot})(Y_{i2k} - \overline{Y}_{i2\cdot}) \\ &= \frac{1}{n-1}(COV_1 + COV_2) \end{aligned} \tag{46}$$

Combining this result with equation (35), we derive

$$r = \frac{COV_1 + COV_2}{2S^2_{pooled}} \tag{47}$$

On the other hand, if r_1 and r_2 denote the estimates of the correlations between the two repeated measures for group 1 and group 2, respectively, we have

$$r_1 = \frac{COV_1}{\sqrt{V_{11}V_{12}}} \tag{48}$$

$$r_2 = \frac{COV_2}{\sqrt{V_{21}V_{22}}} \tag{49}$$

To pool these two correlations, one must first calculate $z_1 = f(r_1)$ and $z_2 = f(r_2)$, where the function f is defined by

$$f(x) = \frac{1}{2}\ln\left(\frac{1+x}{1-x}\right) \tag{50}$$

Then, one must compute z, the average z_1 and z_2, and take $r_{pooled} = f^{-1}(z)$ (see Johnson, 1993). Therefore, we have

$$r_{pooled} = \frac{\sqrt{\frac{(1+r_1)(1+r_2)}{(1-r_1)(1-r_2)}} - 1}{\sqrt{\frac{(1+r_1)(1+r_2)}{(1-r_1)(1-r_2)}} + 1} \tag{51}$$

It is apparent from equations (47), (48), (49) and (51), that the expressions of r and r_{pooled} differ in general. Of course, if $r_1 = r_2$, then $r_{pooled} = r_1 = r_2$. Therefore, a sufficient condition for having $r = r_{pooled}$ is given by

$$V_{11} = V_{12} = V_{21} = V_{22} = S^2_{pooled} \tag{52}$$

and

$$COV_1 = COV_2 \tag{53}$$

Of course, these conditions are never met exactly in any given data set. However, we can expect that, when the covariances as well as the variances in the sample do not differ by much, the estimates r and r_{pooled} will be reasonably close.

Appendix C

Bias Associated with Assuming a Balanced Design

In many studies, primary researchers report the total number of participants in the experiment, but fail to indicate the number of participants in each cell of the design. When confronted with this situation, we recommend that the meta-analyst assume a balanced design. In the following, we estimate the bias introduced by this rule of thumb, in the case of a design with two groups and two repeated measures (2 x 2 design). Note that, although this paper is primarily concerned with designs including both between-subjects and within-subjects factors, the following estimate applies for any kind of design, as long as the method of unweighted means analysis of variance was used in the original study.

Consider a 2 x 2 design with repeated measures on one factor, and with n_1 participants in Group 1 and n_2 participants in Group 2. We denote $N = n_1 + n_2$ and:

$$e = \frac{n_1}{N} - \frac{1}{2},$$

which is a measure of deviation from a balanced design. We also denote by α_u the coefficient corresponding to an unbalanced design with n_1 and n_2 participants in groups 1 and 2, respectively, and by α_b the coefficient corresponding to a balanced design with a total of N participants. In the situation described above, α_u is the real coefficient, corresponding to the original study, while α_b is the best guess of the meta-analyst, given the total number of participant , N.

Using equation (22), we find that:

$$\begin{aligned} \alpha_u &= \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \\ &= \frac{1}{2N}\left(\frac{1}{0.5+e} + \frac{1}{0.5-e}\right) \\ &= \frac{1}{N}\left(\frac{1}{1+2e} + \frac{1}{1-2e}\right) \end{aligned} \tag{54}$$

Taking $e = 0$, we see that

$$\alpha_b = \frac{2}{N} \tag{55}$$

Now, to estimate S_{pooled} (or r), one has to multiply a quantity independent of α by the square root of α. Therefore, the error introduced by mistakenly assuming a balanced design is given, in proportion, by:

$$\begin{aligned} Bias &= \frac{\sqrt{\alpha_b} - \sqrt{\alpha_u}}{\sqrt{\alpha_u}} \\ &= \frac{\sqrt{2} - \sqrt{\frac{1}{1+2e} + \frac{1}{1-2e}}}{\sqrt{\frac{1}{1+2e} + \frac{1}{1-2e}}} \\ &= \sqrt{1-4e^2} - 1 \end{aligned} \tag{56}$$

Now, by asymptotic expansion of the square root function (see Holmes, 1995), one finds that:

$$Bias = (1 - 2e^2 + o(e^2)) - 1 = -2e^2 + o(e^2) \tag{57}$$

Therefore, when the error e is small, the bias introduced by assuming a balanced design is a decrease (in proportion) of approximately $2e^2$ in the estimates of S_{pooled} and r. For example, in a design with 15 participants in group 1 and 25 participants in group 2, we have:

$$e = \frac{15}{40} - \frac{1}{2} = -.125 \tag{58}$$

Therefore, the bias, as calculated from equation (36), is a decrease of 3.175% in the estimates, while our approximation (37) gives a decrease of approximately 3.125%.