

Combining Effect Sizes Across Different Factorial Designs:

A Perspective Based on Generalizability Theory

Scott B. Morris

Illinois Institute of Technology

Richard P. DeShon

Michigan State University

Paper presented at the 17<sup>th</sup> Annual Conference of the Society for Industrial and Organizational Psychology, Toronto, Ontario, Canada (April, 2002).

Combining Effect Sizes Across Different Factorial Designs:  
A Perspective Based on Generalizability Theory

In recent years, measures of effect size have become central to many areas of psychological research. Many argue that reporting effect sizes is essential due to the inherent problems with statistical significance testing (Wilkinson et al., 1999). In addition, effect size estimates serve as the basic data for meta-analysis, and therefore are critical for modern literature reviews. Accurate effect size estimates are also needed for effective power analysis, which plays an important role in the design of new research.

Effect size estimates will be useful only to the extent that they can be interpreted beyond the scope of an individual study. Unfortunately, studies that utilize different research designs will often produce effect sizes that estimate different parameters, and therefore are not directly comparable. Several aspects of the study design will impact the comparability of the effect sizes, such as the choice of an outcome measure, the strength of experimental manipulations, and the control for confounding factors. In this paper, we focus on only one of these issues -- the impact of the study design on the metric of the effect size estimate.

In order for effect size estimates to be comparable across designs, they must be expressed in a common metric. For example, if different studies use different measures for the dependent variable, the mean difference between groups will not be directly comparable, unless the effect size is defined in a common metric.

This requirement has led to the widespread use of standardized measures of effect size, such as the standardized mean difference ( $d$ ) or the correlation coefficient ( $r$ ), which are not influenced by the scaling of the dependent variable (Hedges & Olkin, 1985).

Unfortunately, using the standardized mean difference does not guarantee metric comparability. Differences in the design of the study may also lead to different standard deviations, which can result in different effect size estimates. Studies with more or less homogeneous samples, or different degrees of experimental control for extraneous factors will generally not produce estimates of effect size that are comparable. Similarly, studies with different factorial designs or that include different covariates may not estimate the same effect size.

In this article, we will focus on the problem of combining results across different factorial ANOVA designs. For example, consider two studies on the effectiveness of a training program designed to improve attitudes toward gender diversity. In Study A, employees are randomly assigned to training and no-training conditions. Study B uses a factorial design, where the training condition is crossed with the participants' gender.

If gender has an impact on attitudes, then the error variance in Study A will be greater than the error variance in Study B. The error term in Study A estimates the variance within a training condition, and therefore includes variance due to gender as well as other factors. The error term in study B estimates the variance within a training condition and within a gender group. Therefore, the variance in attitudes due to gender will add to the within-level variance in Study A, but not to the within-cell variance in Study B. Effect size estimates from these two studies would be defined using these different variance estimates, and therefore would not estimate the same population parameter.

The error terms for both studies are correct, but they reflect different definitions of the population. In Study A, the results are generalized to a population where gender is allowed to vary naturally. In Study B, the results are generalized to a population where gender is constant. The problem is that the person using the research results may want to generalize effect size estimates to a population that is different from the population intended by the original researcher. This issue is particularly important for meta-analysis, where effect sizes from a number of studies are brought together to estimate a common parameter.

Procedures for correcting effect sizes for the effect of factorial design have long been available (Glass, McGaw & Smith, 1981; Morris & DeShon, 1997). For the example described above, the variance due to gender could be added into the estimate of the standard deviation for Study B. This would result in standard deviations from both studies that include variance due to gender, and the resulting effect size estimates would be comparable.

Cortina and Nouri (2000) pointed out that the Glass et al. (1981) correction is only appropriate when the non-focal factor is expected to vary in the population to which the researcher wants to generalize results. When the non-focal factor was artificially manipulated by the experimenter, and would not vary naturally in the population, the correction is unnecessary. Olejnik and Algina (2000) made a similar distinction between individual difference factors and manipulated factors.

Unfortunately, these distinctions do not fully reflect the complexity of the situation. In determining the appropriate definition of the effect size, one must consider not only the nature of the variables in the study, but also how these variables relate to the population to which inferences will be made. Because different researchers may adopt different definitions of the relevant population, it is possible to define the effect size in more than one way. Furthermore, when conducting a meta-analysis, one must consider how the variables are treated in the entire set of studies to be synthesized. Even after applying the Glass et al. (1981) correction, effect sizes may still not be comparable across all designs. Concepts from Generalizability Theory (GT) offer a useful framework for defining the best estimate of effect size and for determining whether comparable estimates are possible from a set of studies.

### When are Effect Sizes Comparable?

Hedges & Olkin (1985) note that when measures are linearly equitably, standardized effect sizes will be directly comparable, even when they are obtained from studies using outcome measures on different scales.

Consider two populations, representing the two conditions being compared in an experiment. It is assumed that the populations are normally distributed with equal variance  $\sigma^2$  and means  $\mu_1$  and  $\mu_2$ . The standardized mean difference ( $\delta$ ) is defined

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}. \quad (1)$$

The outcome measure can be rescaled by a linear transformation,  $y' = Ay + B$ , where A and B are arbitrary constants. The standardized mean difference between groups on the rescaled outcome would be

$$\delta' = \frac{(A\mu_1 + B) - (A\mu_2 + B)}{A\sigma} = \frac{\mu_1 - \mu_2}{\sigma}. \quad (2)$$

Thus, the linear transformation has no effect on the standardized mean difference.

A key assumption of this approach is that the standard deviation from each study estimates the same population parameter, except for the multiplicative constant A. That is, the only difference in standard deviations is due to the different scaling of the dependent variable. However, in practice, other factors can also impact the standard deviation obtained in different studies. For example, studies that use more homogeneous populations or more stringent experimental control of nuisance variables will tend to have smaller standard deviations. Because the standard deviations from these studies do not estimate the same population parameter, the standardized mean differences will not estimate the same parameter, and therefore should not be combined in a meta-analysis. Similarly, the error terms used in different experimental designs can reflect different definitions of the population variance.

### Effect Sizes from Factorial Designs

Consider a study with two factors (e.g., training and gender). The effect of training could be analyzed using either a single factor ANOVA (ignoring gender), or using a training x gender ANOVA. When the non-focal factor (i.e., gender) has a non-zero effect on the dependent variable, the two designs will result in different error terms.

The difference between single factor and factorial ANOVA can be seen by examining the mean square error (MSE). In a single-factor ANOVA, MSE is equivalent to the pooled variance within levels of the factor. Thus, the square root of MSE can be substituted uses as an estimate of the population variance in when computing the effect size estimate. For a single-factor ANOVA, the mean square error can be defined as

$$\underline{MSE} = \frac{\underline{SS}_T - \underline{SS}_a}{\underline{df}_T - \underline{df}_a}, \quad (3)$$

where  $\underline{SS}_T$  and  $\underline{df}_T$  represent the total sum of squares and degrees of freedom, and  $\underline{SS}_a$  and  $\underline{df}_a$  represent the effect of interest.

In a fixed-effects between-groups ANOVA with two independent factors, the MSE reflects the within-cell variance, that is, the variance within a particular level of both factors.

$$\underline{MSE} = \frac{\underline{SS}_T - \underline{SS}_a - \underline{SS}_b - \underline{SS}_{ab}}{\underline{df}_T - \underline{df}_a - \underline{df}_b - \underline{df}_{ab}}, \quad (4)$$

where  $\underline{a}$  represents the main effect of interest,  $\underline{b}$  represents the main effect of the non-focal variable, and  $\underline{ab}$  represents the interaction term. Because the sums of squares for the additional effects are removed from the error term, the MSE for a factorial design will typically be smaller than the MSE for a single-factor ANOVA. When these values are used in to estimate  $\sigma$ , the smaller MSE in the factorial ANOVA will result in a larger effect size estimate.

In the example described above, Study A compared training groups, and gender was unmeasured. Therefore, the only possible estimate of the standard deviation would include variance due to gender. In order to be comparable, the effect size estimate in Study B should also be computed using a SD that includes variance due to gender.

Glass et al. (1981) showed that the within-level standard deviation from a one-way design could be computed from factorial ANOVA results. The within-level standard deviation that would have occurred from a single-factor design is

$$\hat{\underline{S}}_p = \sqrt{\frac{\underline{SS}_b + \underline{SS}_{ab} + \underline{SS}_e}{\underline{df}_b + \underline{df}_{ab} + \underline{df}_e}}, \quad (5)$$

where  $SS_e$  and  $df_e$  are the within-cell sum of squares and degrees of freedom. If only  $F$ -values and  $df$  are available, Morris and DeShon (1997) provided an alternate formula for computing the corrected effect size,

$$d_c = d \sqrt{\frac{df_e + df_b + df_{ab}}{df_e + df_b F_b + df_{ab} F_{ab}}}, \quad (6)$$

where  $d_c$  is the effect size computed using the standard deviation in Equation 5, and  $d$  is the effect size computed based on the MSE.

However, this correction factor is not always appropriate. Cortina and Nouri (2000) and Olejnik & Algina (2000) both argue that the correct standard deviation to be used in computing the standardized mean difference will depend on the nature of the non-focal variable. They recommend use of the correction when the non-focal variable reflects individual differences that vary in the population, but not when the non-focal variable is artificially manipulated by the experimenter.

In order to determine the conditions under which the standardized mean difference effect sizes will be comparable across studies, it is necessary to first understand the sources of variance that impact test scores. Generalizability theory (G-Theory) provides a useful framework for exploring this issue.

#### Effect Sizes and Generalizability Theory

G-Theory was developed to model the sources of variance that impact observed scores obtained through a measurement process (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). The theory is largely associated with measurement issues and the estimation of reliability but, in reality, it has far broader applicability. When using G-theory to model the generalizability of inferences, a researcher must first identify the factors that influence observed scores on the variable of interest. When considered together, the factors that combine to simultaneously and interactively influence observed scores define a multidimensional inference space (termed the universe of admissible observations). The researcher may wish to generalize the research inferences across all levels of each factor expected to influence observed scores. For instance, in a drug response study the researcher may wish to generalize inferences across all possible levels for each factor such as drug dosage levels, gender, age, health status, weight, SES, genetic predispositions, and various environmental factors. If this is the case, then the researcher must ensure that the data are collected in a manner that allows each of these variables to influence responses (DeShon, 2002).

Most frequently, however, a researcher is interested in generalizing inferences across a more limited set of conditions than might be represented in the universe of admissible observations. In G-theory, this more restricted inference space is termed the Universe of Generalization. In this more restricted inference space, some factors will vary randomly, but others will be assumed to have fixed

values that would not vary across replications of the research. So, for instance, a particular researcher may not be interested in the effect of genetic variation on the response to the drug and instead may limit attention to only one (or a small subset) genetic profile. The variance components that are free to vary are then used to estimate a generalizability coefficient, which indicates the extent to which scores will generalize to other conditions in the Universe of Generalization.

Similar issues are faced when combining effect size estimates from studies with different designs. Each study will be subject to different sources of variation. Some studies will control for a particular nuisance variable, others will manipulate this same variable in a factorial design, and still others will ignore this variable. The treatment of variables in these different designs implicitly defines different inference spaces that reflect the inferential desires of the researcher who originally conducted the research. To cope with the ambiguity resulting from the difference in research designs and corresponding inference spaces, the meta-analyst must first identify the relevant factors (similar to the Universe of Admissible Observations), and how they have been addressed in each study. This does not mean that all possible contributing factors must be identified. Variables that have been ignored in all studies, or which are controlled in all studies can be excluded from the inference space (e.g., the Universe of Admissible Observations). This inference space defines a population of individuals, across which inferences may be generalized.

To combine the standardized mean difference across studies, each must be standardized relative to the same population or inference space. The results of each primary study will be defined in terms of the restricted inference space adopted by the primary study's authors. It is highly unlikely that this inference space will be the same across studies and the inference space in the primary study may not match the meta-analyst's desired inference space. Therefore, the results of each study must be transformed into a common metric, as suggested by the Glass et al. (1981) correction formula. The problem is that these correction formulas force a particular inference space, which will not always be appropriate or desirable. The appropriate correction depends on both the study's design and how that design relates to the meta-analyst's desired inference space.

#### Choice of an Inference Space

Consider again the training example given above, where the focal variable is training vs. no-training, and a potential nuisance factor is participant gender. Table 1 illustrates the different ways in which the non-focal variable might have been treated in the original study, and how this relate to two possible ways the meta-analyst might define the inference space.

The first column of Table 1 indicates how a nuisance variable was treated in the original study design. In some studies, the variable will be uncontrolled, that is, it will vary randomly in the sample. The natural variance in this variable therefore becomes part of the error term. In other situations, the

variable might be constant. This would occur if the variable is controlled in the research design, or if it reflects a variable that does not vary naturally in the population. Another possibility would be a study that includes the nuisance variable as a factor in a factorial design. Here, two different effect sizes can be estimated, depending on whether the Glass correction is applied.

For each situation, Table 1 indicates the error term used to define the effect size and the sources of variance estimated by this error term. The variance estimate determines whether the effect size will be in the appropriate metric for the inference space defined by the researcher. Two possible inference spaces are considered, one where the nuisance variable varies randomly in the population (space 1), and one where the nuisance variable is fixed (space 2). The last two columns indicate whether effect sizes from each design are appropriate for each inference space.

In general, uncorrected effect sizes are comparable across designs only when the non-focal variable has been controlled by the experimenter (either held constant or manipulated), or when the non-focal variable has no effect in the population. In each case, the non-focal variable is fixed in the inference space.

Corrected effect sizes, on the other hand can be combined across uncontrolled and factorial designs, but only when an additional assumption has been met. The necessary assumption is that effect of the non-focal variable in the primary study is equal to the effect this variable naturally has in the population. For example, consider a study where the non-focal variable is the participant's gender. Because the difference between men and women in the study will estimate the population difference, they study result can be used to estimate the population variance component, and the correction is appropriate.

In contrast, consider a non-focal variable that is artificially manipulated. Experimental researchers often use extreme manipulations in order to increase the sensitivity of the study to detect small effects. In such cases, the effect of the non-focal treatment may overestimate the population variance component, leading to an inaccurate correction. Similarly, a manipulation that is not as strong as the stimuli in the natural environment will result in an underestimate of the variance component.

In this situation, neither the corrected nor uncorrected effect sizes from the factorial design would be comparable to an effect size from a study where the nuisance variable is uncontrolled. The corrected effect size is likely to overestimate the variance due to the nuisance variable, artificially inflating the standard deviation, and reducing the effect size estimate. The uncorrected effect size, on the other hand, will likely underestimate the variance due to the non-focal variable, resulting in an overestimate of the effect size.

### Conclusion

When combining effect sizes across research designs, it is important to understand the impact of the design on the effect size estimate. Effect sizes computed from ANOVA designs with different factors

will often not be directly comparable, unless appropriate corrections are applied. In order to determine the proper correction, a meta-analyst must first identify the nature of the variables involved and the desired inference space.

For variables that are fixed in the inference space, comparable effect sizes can be estimated from any study design where the variables are treated as fixed. This applies to variables that were manipulated or experimentally controlled by the researcher. In general studies where the variable is uncontrolled will not provide an appropriate effect size estimate for this inference space, unless they can be assumed not to vary in the populations studies. Effect sizes from factorial designs should not be corrected for the influence of these variables.

For variables that are random in the inference space, comparable effect sizes could be estimated only from studies where the variable is uncontrolled. Also, if the variable were examined in a factorial design, the variance due to this variable could be estimated and used to adjust the effect size estimate using the procedure suggested by Glass et al (1981). However, this correction will only be appropriate when the effect of the variable in each study accurately represents the variance that would occur naturally. Because experimental manipulations may be stronger or weaker than the random variance component, the accuracy of these corrections will often be questionable.

In many situations, it will not be possible to define an effect size that is comparable across all of the studies in a meta-analysis. If meta-analytic results are to be trusted, new strategies must be developed to deal with these situations. The primary problem results from the use of the sample standard deviation in the estimate of the effect size. Because different designs can lead to different definitions of the standard deviation, this practice leads to incompatible effect size estimates.

One possible solution would be to utilize an estimate of effect size that is not dependent on the study-level variance estimate. For example, in some research areas, it may be feasible to conduct the meta-analysis on the unstandardized mean difference (Lipsey & Wilson, 2001). Although use of a standardized effect size has become the standard practice in meta-analysis, it is not always necessary. As long as the same outcome measure is utilized in all studies, there is no reason to utilize a standardized measure of effect size. However, although there are some situations where meta-analysis is performed on a common measure, in most cases, the researcher must find a way to equate effect sizes based on a variety of measures.

Another approach would be to use the population standard deviation, rather than the sample standard deviation to standardize the effect size estimate. This would adjust for difference due to the measures used in different studies, without introducing complications that arise due to the design of the individual studies.

For most professionally developed measures, normative data is available. As long as normative samples are very large, they could provide an estimate of the standard deviation that is relatively unaffected by sampling error, and therefore this estimate can be treated as a population parameter. Each effect size would be defined as the observed mean difference divided by a constant. As a result, procedures for meta-analysis on the unstandardized mean difference could be applied.

This solution creates some additional complications that will have to be addressed. The meta-analyst would have to ensure that the normative samples used for different measures reflect the same population. Otherwise, the standard deviations will not be comparable, nor would the resulting effect size estimates.

A further complication would arise if the normative data were unavailable or based on small samples for some of the measures used in the meta-analysis. This would introduce imprecision into the effect size estimates for these measures, which would not be reflected in the meta-analysis. Furthermore, if more than one effect size was estimated from the same population norm, this would introduce dependency into the effect sizes. Therefore, more complex meta-analysis models may be needed to account for these dependencies.

In this article, we demonstrated the utility of Generalizability Theory as a conceptual framework for obtaining comparable effect size estimates from a variety of research designs. This approach highlights the importance of understanding how non-focal variables were treated in the original research, and of explicitly specifying the role of these variables in the inference space. Our focus was on ANOVA designs with different factors; however, the framework should apply to other types of research designs as well. Similar issues are faced when combining results across regression models with different sets of predictors or across ANCOVA designs with different covariates. The conceptual framework allowed by Generalizability Theory should also provide a valuable tool for interpreting and combining effect size estimates from these research designs.

Table 1

Effect of study design and Glass correction on the population reflected in the standardized mean difference.

Treatment of Factor B in Study Design	Error Term	Variance Estimated	Same Metric as Inference Space?	
			Space 1 $\sigma_e^2 + \sigma_b^2 + \sigma_{ab}^2$	Space 2 $\sigma_e^2$
Uncontrolled	W/A	$\sigma_e^2 + \sigma_b^2 + \sigma_{ab}^2$	Yes	No <sup>a</sup>
Constant (Experimentally Controlled )	W/AB	$\sigma_e^2$	No <sup>a</sup>	Yes
Included in Factorial Design (without Glass correction)	W/AB	$\sigma_e^2$	No <sup>a</sup>	Yes
Included in Factorial Design (with Glass correction)	W/A	$\sigma_e^2 + \Sigma \beta^2 + \Sigma (\alpha\beta)^2$	Only if $\Sigma \beta^2 = \sigma_b^2$ and $\Sigma (\alpha\beta)^2 = \sigma_{ab}^2$	No <sup>a</sup>

Note: W/A = variance within levels of A. W/AB = variance within cells defined by factors A and B.  $\sigma_e^2$  = population variance not due to variables A and B.  $\sigma_b^2$  = variance in population due to variable B (nuisance variable).  $\sigma_{ab}^2$  = variance in population due to AxB interaction.  $\Sigma \beta^2$  = effect of experimental manipulation of variable B (nuisance variable).  $\Sigma (\alpha\beta)^2$  = effect of manipulation of AxB interaction.

<sup>a</sup>When B and AB are zero, all conditions will estimate either inference space.

## References

- Cortina, J. M., & Nouri, H. (2000). Effect size for ANOVA designs. Thousand Oaks, CA: Sage.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.
- DeShon, R. P. (2002). Generalizability Theory. In F. Drasgow & N. Schmitt (Eds.), Measuring and analyzing behavior in organizations. San Francisco, CA: Jossey-Bass
- Glass, G. V., McGaw, B., & Smith, M. L. (1981). Meta-analysis in social research. Newbury Park, CA: Sage.
- Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. Orlando, FL: Academic Press.
- Lipsey, M. W., & Wilson, D. B. (2001). Practical meta-analysis. Thousand Oaks, CA: Sage.
- Morris, S. B., & DeShon, R. P. (1997). Correcting effect sizes computed from factorial analysis of variance for use in meta-analysis. Psychological Methods, *2*, 192-199.
- Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, Interpretations, and limitations. Contemporary Educational Psychology, *25*, 241-286.
- Wilkinson, L., & the Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, *54*, 594-604.