Published in final edited form as:
Rev Gen Psychol. 2013 March ; 17(1): 111–121. doi:10.1037/a0030048.

A Regression Framework for Effect Size Assessments in Longitudinal Modeling of Group Differences

Alan Feingold
Oregon Social Learning Center

Abstract

The use of growth modeling analysis (GMA)--particularly multilevel analysis and latent growth modeling--to test the significance of intervention effects has increased exponentially in prevention science, clinical psychology, and psychiatry over the past 15 years. Model-based effect sizes for differences in means between two independent groups in GMA can be expressed in the same metric (Cohen's *d*) commonly used in classical analysis and meta-analysis. This article first reviews conceptual issues regarding calculation of *d* for findings from GMA and then introduces an integrative framework for effect size assessments that subsumes GMA. The new approach uses the structure of the linear regression model, from which effect sizes for findings from diverse cross-sectional and longitudinal analyses can be calculated with familiar statistics, such as the regression coefficient, the standard deviation of the dependent measure, and study duration.

Keywords

effect sizes; regression; multilevel analysis; clinical trials

Traditionally, psychological data have been examined with classical statistical techniques, such as analysis of variance (ANOVA) and multiple-regression analysis, which use ordinary least squares (OLS) for estimation and are unified under the general linear model (GLM). Accordingly, effect sizes have been developed largely to determine the practical significance of treatment effects and associations in conventional analyses and for use in meta-analysis (Shadish & Haddock, 2009). The well-known equations for calculations of such effect sizes can be found in numerous sources (e.g., Grissom & Kim, 2005; Hedges, 2009; Lipsey & Wilson, 2001; see also special section in *Child Development Perspectives*, Supplee, 2008).

The effect size used is often determined by the nature of the independent and dependent variables. In randomized clinical trials and other experiments, the independent variable is typically categorical (groups or conditions) and the dependent variable is continuous (e.g., scores). Thus, Cohen's (1988) *d* (the standardized mean difference between two groups) is generally the effect size of choice for conveying the magnitude of experimental effects. However, when an experiment has a binary outcome (e.g., success vs. failure) and the data are analyzed by chi-square or logistic regression analyses, the odds ratio (*OR*) is a frequently used effect size that expresses the group difference in probabilities (Fleiss & Berlin, 2009). When the independent and dependent variables are both continuous, the effect size *r* (the ordinary Pearson correlation coefficient), or associated measures based on *percent of variance explained* (e.g., Fairchild, MacKinnon, Taborga, & Taylor, 2009), is frequently used.

Correspondence concerning this article should be addressed to: Alan Feingold, Oregon Social Learning Center, 10 Shelton McMurphey Boulevard, OR 97401-4928. alanf@oslc.org.

Growth Modeling Analysis of Data from Controlled Clinical Trials

Over the last 15 years, *growth modeling analysis* (GMA), which most often uses maximum-likelihood estimation and the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977), has emerged as a compelling alternative statistical framework to classical analysis, particularly for analyzing longitudinal data from controlled clinical trials. GMA—a family of modeling approaches that includes types of multilevel analysis/hierarchical linear models (HLM; Raudenbush & Bryk, 2002; Hedeker & Gibbons, 2006) and covariance structural models/latent curve analysis (Meredith & Tisak, 1990; Singer & Willett, 2003)—compares temporal trajectories (growth curves) between conditions to determine treatment effects. GMA can be conceptualized as a more flexible version of a split-plot ANOVA that uses polynomial contrasts to examine whether trajectories for the repeatedly-measured outcome (the within-subjects factor) are moderated by the treatment (between-subjects) factor (Winer, 1971).

GMA uses repeated measures data from all participants rather than from study "completers," and affords more accurate parameter estimates, especially when data are missing (Atkins, 2005; Little & Rubin, 2002; Schafer & Graham, 2002). In addition, ANOVA has restrictive statistical assumptions (e.g., compound symmetry, homogeneity of variance over time) that are rarely met in practice but are relaxed for GMA (Gibbons, Hedeker, Elkin, Waternaux, Kraemer, Greenhouse, et al., 1993; Gueorguieva & Krystal, 2004). Finally, recent developments in GMA, including latent class growth analysis (e.g., Nagin, 2005) and growth mixture modeling (GMM; e.g., Muthén, Brown, Masyn, Jo, Khoo, et al., 2003), allow for extraction and use of latent classes in the analysis. GMM, for example, can be used to examine whether an intervention is more efficacious for some types of participants than others, which would be useful information when clinicians match individuals to interventions (Babor & Del Boca, 2003).

Because of the many advantages of GMA over classical longitudinal methods, Aiken, West, and Millsap (2008) recently deemed it one of the most important methodological innovations of the twenty-first century, and Kuljanin, Braun, and DeShon (2011) averred that these methods "are currently the dominant approaches to the analysis of longitudinal data in psychology" (p. 249). Thus, GMA is now widely used in program evaluations. Reviews of clinical trials published in *Prevention Science*, *Journal of Consulting and Clinical Psychology* (*JCCP*), and *Archives of General Psychiatry* found an increasing number of studies using GMA in all three journals (Feingold, 2009; Gueorguieva & Krystal, 2004).

Ideally, methods for significance testing and effect size calculations for group differences observed in GMA would have evolved concurrently to facilitate comparisons between findings from traditional and GMA studies. Moreover, it would be useful for meta-analysts, who frequently combine results to produce syntheses of intervention outcomes that typically have far more impact than results of individual trials, to have the effect size from a GMA expressed as the standardized mean difference (d) that is most often used in quantitative synthesis (Lipsey & Wilson, 2001).

However, Feingold (2009) reviewed 43 clinical trials regarding interventions for treatment and prevention published in *JCCP* that had used GMA and found p values to be ubiquitous but noted only 13 reports of effect sizes calculated from model-based coefficients. Moreover, the formulas used in those 13 studies were not conceptually or mathematically equivalent, and none of them expressed the effect size in the same metric deployed in classical analysis, precluding their use in meta-analysis. These results are consistent with the recent conclusion that "in certain situations (e.g., multilevel designs), no consensus exists on how to conceptualize and/or calculate effect size measures" (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008, p. 850).

There are several factors that may have affected attention paid to effect sizes for GMA findings. Equations for calculating effect sizes for classical analyses became ubiquitous a generation ago but only for cross-sectional analyses conducted with independent groups. Formulas for determining effect sizes for classical repeated measures designs--even when only two time points are used--have been little known or applied (Dunlap, Cortina, Vaslow, & Burke, 1996; Feingold, 2009; Morris & DeShon, 2002).

The effect size *d* is calculated by dividing the difference between the means of two independent groups by the pooled within-group standard deviation (Cohen, 1988). The denominator in the equation is thus an estimate of the standard deviation of the outcome measure in the population. However, when participants have been measured twice (e.g., before and after a manipulation), primary investigators and meta-analysts alike have often computed *d* for the difference in means between the two times (or from the interaction between time and group when a comparison group was included) by using the standard deviation of the pretest-posttest change scores as the denominator rather than the more appropriate standard deviation of the outcome (Morris & DeShon, 2002). An effect size based on change score variations from a within-subjects design does not convey the magnitude of an effect because it is confounded with the pretest-posttest correlation (Dunlap et al., 1996), which generalizes to repeated-measures designs that include a control group.

Effect Sizes for ANOVA and Growth Modeling Analysis

The classical design that is the true analogue of GMA is a repeated-measures ANOVA with polynomial contrasts that examines changes in outcomes over time, and whether these trajectories vary across conditions. That effect sizes for GMA hypothesis tests are not routinely reported is not surprising because the corresponding effect sizes for ANOVA and multiple regression analysis also have rarely been addressed. When repeated-measures ANOVA is used, the standard operating procedure is to report the *F* for the interaction contrast (or sometimes just the omnibus *F* for the interaction) and graph the means for each group over the time factor to display the difference in trends between the treatment and control conditions. GMA users engage in essentially the same reporting practices, except that the temporal trend is expressed as a line or curve of the means--as estimated from the GMA equation--for each condition. Moreover, major GMA programs, such as HLM (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2004), output only an unstandardized coefficient for the group difference in growth rate (e.g., the effect of a time-invariant covariate, such as a treatment factor, on a random slope), and users' manuals for such software--including both HLM (Raudenbush et al., 2004) and Mplus (Muthén & Muthén, 2010)--do not explicitly mention the effect size associated with it.

Need for a New Framework for Effect Size Assessments

Raudenbush and Liu (2001) described an approach for calculating an effect size for the difference between two groups in linear trends from a GMA that uses differences in growth rates between the groups and the standard deviation of raw scores to determine baseline-adjusted effect magnitude at the end of the study in the familiar *d* metric. Feingold (2009) presented a formal equation for an effect size for GMA based on their ideas but it expresses only the difference between two groups in linear growth of a continuous outcome. In many trials, however, three or more groups may be used; trajectories may not be linear; randomization to conditions may occur at the cluster instead of at the individual level (i.e., clinic or school rather than patient or student); subject-factor covariates (e.g., gender or risk status) may be included in the model; the effects of unobserved heterogeneity may be examined; and outcomes may be categorical (e.g., binary) rather than continuous.

Therefore, there is a pressing need for an approach that can guide the calculation of effect sizes for multilevel findings from the wide range of longitudinal designs used to compare means of two or more groups. Ideally, a conceptual basis for effect size assessments would be provided via an integrative framework that can be adapted to different research cross-sectional and prospective designs, thus allowing GMA effect sizes to be comprehended in the context of a general model. Moreover, effect sizes obtained from between-subjects ANOVA, repeated-measures ANOVA, and GMA should all estimate the same parameter (e.g., mean difference between groups at end of study), as design factors should not moderate estimates of effect potency (Olejnik & Algina, 2003) and use of a common metric would allow findings from different kinds of studies to be combined in a meta-analysis.

The Current Work

This article first formulates a regression framework that will be useful to empirical researchers and meta-analysts in computing effect sizes for findings obtained with varied research designs. Next, conceptualizations and calculations of effect sizes are presented. Finally, methodological issues related to these methods—including selection of the standard deviation and handling of groups formed by non-random assignment or measured subject characteristics—are discussed.

Most of this article is concerned with effect sizes for mean differences in analyses conducted with a categorical independent variable and a continuous dependent variable (characteristic of most experiments) that are generally in the metric of *d*. However, the generalization of the procedures to analyses with categorical outcomes—where the *OR* is the effect size—is described, and issues regarding effect sizes when the predictor variable is continuous are also discussed.

A GLMM Regression Framework for Effect Size Assessments

Predictor variables (covariates) used in an analysis conducted with OLS estimation and the GLM may have different distributions (e. g., normal, multinomial, and Poisson) but normality is assumed for the dependent variable (Cohen, Cohen, West, & Aiken, 2003). *Generalized linear models* (Nelder & Wedderburn, 1972) is an extension of GLM that integrates OLS regression with logistic and Poisson regression analyses and uses maximum-likelihood estimation in analysis of both normally-distributed and categorical dependent variables, with similar structural models (Agresti, 2002; Hosmer & Lemeshow, 2000). Generalized linear models also subsumes GMA with fixed effects, including latent class growth analysis, and its formulation has recently been hailed as one of statistics' most important contributions to psychology (Wright, 2009).

An extension of generalized linear models, called *generalized linear mixed models* (GLMM; McCulloch & Searle, 2001) or *hierarchical generalized linear models* (Raudenbush & Bryk, 2002), can accommodate GMA with both fixed and random effects for dependent variables having normal, Bernoulli (binomial), Poisson, or multinomial distributions and thus subsumes GLM and generalized linear models. HLM is a special case of GLMM with normally distributed outcomes using an identity link function (Raudenbush & Bryk, 2002). Although these frameworks unify a wide range of statistical procedures (see Table 1), researchers who have discussed these models have typically been more concerned with null hypothesis significance testing than with effect size estimation. Accordingly, this article discusses conceptualizations and calculations of effect sizes associated with hypothesis tests from different GLMM regression models.

Continuous Outcomes with No Moderation of Intervention Effect

Two Independent Groups

In the simplest (posttest-only) experimental design, participants are randomly assigned to either an experimental or a control condition; an intervention is administered only to the experimental group; a continuous measure is administered to both groups at the end of the study; and the means of the two groups on the outcome are compared. The effect size for this cross-sectional between-subjects design is the classical standardized mean difference, most often calculated with Equation 1,

$$d=(M_{\mathrm{T}}-M_{\mathrm{C}})/SD, \quad (1)$$

where M_{T} is the mean of the treatment group, M_{C} is the mean of the control group, and the *SD* is the pooled within-group standard deviation.

By contrast, the integrative GLMM framework uses the structural model of a regression equation for calculating the effect size,

$$Y=a+b\mathrm{Group}+e, \quad (2)$$

where Y is the dependent variable, a is a constant (intercept), b is an unstandardized regression coefficient, group is a dichotomous independent variable, and e is the error—the difference between the observed Y and the Y that is predicted by the regression model (a + bGroup).

Variable codes differing by one unit (e.g., −1/2 and 1/2 but *not* −1 and 1) can be ascribed to participants in the experimental and control groups (the group variable) for use in the regression of Y on group. The b then equals the difference between the group means and the standard deviation of e is the pooled within-group standard deviation (*SD*). Thus, for the randomized two-groups design,

$$d=b/SD, \quad (3)$$

in a GLMM formulation, where *SD* is the standard deviation of the raw outcome scores within groups (i.e., the square root of the *MSE* in regression and ANOVA, where *MSE* is the error term in between-subjects designs).

Independent-Groups Pretest-Posttest Design

The simplest longitudinal experimental design is an extension of the two-independent-groups design that adds a baseline assessment (or pretest) for both the intervention and the control groups (Morris, 2008). The effect size conveys the difference in change scores between the two groups (Becker, 1988; Feingold, 2009; Morris & DeShon, 2002),

$$d=M_{\mathrm{CHANGE-T}}/SD_{\mathrm{PRE-T}}-M_{\mathrm{CHANGE-C}}/SD_{\mathrm{PRE-C}}, \quad (4)$$

where $M_{\mathrm{CHANGE-T}}$ is the mean of the change score (difference between pretest and posttest means) for the treatment group, $M_{\mathrm{CHANGE-C}}$ is the mean of the change scores for the control group, $SD_{\mathrm{PRE-T}}$ is the pretest *SD* for the treatment group, and $SD_{\mathrm{PRE-C}}$ is the pretest *SD* for the control group (for a worked example, see Feingold, 2009).

Given randomization at study onset, $SD_{\mathrm{PRE-T}}$ and $SD_{\mathrm{PRE-C}}$ are both estimates of the same parameter and can be replaced by the pooled standard deviation, *SD*, in Equation 4 (Morris, 2008). In the case of groups not formed by random assignment, Equation 4 should only be used when there is homogeneity of variance between the two groups. Such homogeneity of variance is assumed by the statistical model and Equation 4 could yield dubious effect sizes when the assumption is violated.

From a GLMM framework, the *d* can be calculated with an expansion of Equation 2 to add a term for the pretest variable,

$$Y=a+b_1\mathrm{Pretest}+b_2\mathrm{Group}+e, \quad (5)$$

where b_1 is the coefficient for a continuous pretest score and b_2 is the coefficient for a dichotomous treatment variable. Thus, b_2 is the difference between the groups at end of treatment adjusted for the mean difference at onset of study. Because random assignment ensures that the correlation between pretest and group is expected to be zero, the expected value of b_2 is not changed by the inclusion of a pretest covariate in the model. However, the standard deviation of *e* from a model that does *not* include the pretest (i.e., Equation 2) should be used as the denominator when using Equation 5 to calculate the numerator. When a pretest (or any other covariate) is included in a regression, the standard deviation of *e* is not the standard deviation of the outcome but of a residualized outcome.

Three or More Independent Groups

Researchers often use more than two groups (conditions) in their studies. A priori comparisons or contrasts among means are often used when participants are randomly assigned to more than two groups and effect sizes are more meaningful for such planned comparisons than for omnibus comparisons (Rosenthal, Rosnow, & Rubin, 2000).

In a randomized clinical trial with three groups, for example, one contrast might compare the means of two different treatments and a second contrast could compare the mean of the participants receiving any treatment with the mean of the control group. Because contrasts are essentially comparisons between two weighted means, a *d* can be calculated for each planned comparison in a given analysis. When the contrast compares two groups (e.g., two different treatment conditions), the effect size can be calculated using Equation 1, and the within-group standard deviation is typically pooled from all groups in the study. When a comparison is conducted that incorporates means from more than two groups, the effect size can be computed by averaging means for two or more groups for at least one term in the numerator. In the preceding example, the effect size for the contrast that compares the average of the two treatment groups (TA and TB) with that of the control group from a traditional ANOVA framework would be,

$$d=(\{M_{\mathrm{TA}}+M_{\mathrm{TB}})/2\}-M_{\mathrm{C}})/SD. \quad (6)$$

From the GLMM framework, however, the effect size for each contrast would be obtained by modification of the regression model for two groups (Equation 2) to include variables for the two contrasts, i.e.,

$$Y=a+b_1\mathrm{Group1}+b_2\mathrm{Group2}+e, \quad (7)$$

If, for example, Group1 is the code for the contrast comparing the two treatment groups, the groups could be assigned values of −1/2 for TA, 0 for the control condition, and 1/2 for TB. Then b_1 in Equation 7 would equal the difference between the means of the two treatment conditions. The effect size for this comparison would be calculated with Equation 3, with b_1 substituted for *b*. If Group2 is the variable for the contrast between treated and untreated participants, the groups would be assigned values of 1/3 for TA, −2/3 for the control condition, and 1/3 for TB. Then b_2 would be the difference between the mean of the control group and the average of the means of the two treatment groups. (The difference between the contrast weights for the groups compared must differ by one unit so that b equals the difference in means between groups compared by the contrast.) The d for this contrast would also be calculated with Equation 2 but with b_2 substituted for b.

Repeated Measures with Three or More Time Points for Linear Model

The extension of the independent-groups pretest-posttest design to handle three or more levels of the repeated measures (time) factor is a split-plot ANOVA with a contrast that compares linear growth on the dependent variable between the intervention and control groups. Although there has been little discussion in the psychological or methodological literature of an effect size for this interaction contrast in ANOVA, Feingold (2009) recently described an effect size for a test of the corresponding hypothesis in GMA, which generalizes to the effect size conceptualization for its classical counterpart because both types of analyses are variants of GLMM.

In ANOVA (and regresson), the difference between the means of the two groups at the end of the study can be estimated from the model by calculating the difference in the rate of change between groups per unit of time through modeling of the group means as a function of group and time and then multiplying that value by study duration. This model-derived product is divided by the pooled within-group SD of the outcome measure (Y) to produce an effect size in the same d metric used in the completely randomized and independent-groups pretest-posttest designs.

For example, in an experiment that assesses the treatment and control groups at baseline and then weekly for each of the following three weeks, the means for the control group might be 2, 3, 4, and 5 at baseline, week 1, week 2 and week 3, respectively, indicating a steady improvement in the absence of treatment. The corresponding means for the intervention group might show greater growth, such as 2, 4, 6, and 8. If time (expressed in weeks) is defined as a mean-centered variable with values of , say, −3/2, −1/2, 1/2, and 3/2 and the four repeated measures means are regressed separately on the four times for each of the two groups, the resulting bs are 1.0 and 2.0 for the treatment and control conditions, respectively. Thus, the control group gained 1 point per week on the dependent measure and the intervention group gained 2 points per week over the course of the trial. The difference between these regression coefficients is 1.0, which could also be obtained from the b for a Group × Time interaction in a single multiple regression that used all 8 cell means (i.e., with the condition means from the two groups combined) as Y values and the 3 codes for the main and interactive effects of group and time as predictor variables,

$$Y=a+b_1\mathrm{Group}+b_2\mathrm{Time}+b_3\mathrm{Group}^{*}\mathrm{Time}+e, \quad (8)$$

where Y is a mean of one of two groups at each of four times, group is the condition coding (−1/2 for control and 1/2 for treatment), time is the time coding, Group*Time is the Group × Time cross-products, and e is zero (because the linear trend of the means is perfect for both conditions in the illustrative data). Then the effect size would be calculated using b_3 from Equation 8 in Equation 9,

$$d=(b_3{}^{*}\mathrm{duration})/SD. \quad (9)$$

When the number of weeks is 3 (as with 3 posttest scores in a study with weekly assessments) and the within-groups standard deviation is, say, 10, d = (1)(3)/10 = .30.

GMA researchers working within a multilevel/HLM framework often use the structural model associated with this analytic question in its *linear mixed models* formulation (e.g., Raudenbush & Bryk, 2002), where the coefficients for the fixed effects are identical to those for OYS regression (Equation 8). However, in HLM, the *e* term at the end of Equation 8 is partitioned into three sources for (a) within-subjects (level-1) variation, (b) level-2 variation in random intercepts among subjects, and (c) level-2 variation in random slopes among subjects. The errors (random effects), and the differences between them in OLS linear regression and GMA models, can be ignored in effect size calculations because only the fixed effects (regression coefficients) are used in those calculations. (Thus, in the following presentation of GMA equations, *e* will be used to refer to the sum of these three sources of errors.) Therefore, the effect size for the treatment effect from a GMA can be calculated with Equation 9 (Feingold, 2009). (For readers more familiar with the 2-level formulation of HLM, the coefficient for effect of group on slope is the same as the Group × Time interaction in both the linear mixed model framework for HLM and in the Equation 8 in OLS regression).

Repeated Measures with Three or More Time Points for a Non-Linear Model

If the outcome trajectory is linear for each group, the effect size reflecting the difference between the means of the two groups at end of the study also conveys the difference between them at every point during the study. For example, if the effect size for an 8-week trial is .80 and the trajectories of *d* are linear, the effect sizes are .20, .40, .60, and .80 at weeks 2, 4, 6, and 8 weeks, respectively. However, the treatment group may improve rapidly relative to controls early in the study but the differential improvement rate may vanish over the course of the trial. Then, for example, the effect sizes for a study of the same duration with the same end-of-study mean difference might be 1.20, .90, .80, and .80 at weeks 2, 4, 6, and 8, respectively. Alternatively, it might take time for the treatment to “kick in.” Then the corresponding effect sizes might be .10, .20, .40, and .80 over the same period. Note that the difference between the groups at week 8 is identical in the three examples but there are important differences in trajectories of effect sizes ascribable to non-linearity.

Because an ANOVA with polynomial trends (equivalent to a multiple regression analysis) examines growth rates, model-based estimates of the means at each time for each group can be derived from a regression analysis that treats the observed means as the dependent variable. Thus, an effect size may be calculated for each time by calculating the difference between the model-estimated means of the groups at that time and dividing it by the standard deviation. Thus, effect sizes can be calculated for ANOVA by modeling the group means from both groups with linear regression as a function of treatment (a dichotomous variable), time (linear trend coefficients) and the Group × Time interaction (Equation 9). A non-linear model requires at least two additional coefficients, one for a quadratic trend of time (e.g., with contrast weights of 1/2, −1/2, −1/2, and 1/2) and one for the interaction of the quadratic trend with group, i.e.,

$$Y=a+b_1\text{Group}+b_2\text{Linear}+b_3\text{Linear}^*\text{Group}+b_4\text{Quadratic}+b_5\text{Quadratic}^*\text{Group}+e, \quad (10)$$

where *Y* is the mean of one of the groups at one of the four times, Linear represents linear change in time, and Quadratic represents corresponding contrasts for non-linear change. Both linear and non-linear change may vary across group.

In a 3-week study, for example, the means may be 5, 6, 7, and 8 at baseline, week 1, week 2, and week 3 for the control group and 5, 8, 10, and 11 for the treatment group, both respectively (*SD* = 10). The model-based mean for each group at each time can be calculated from the regression equation given in Equation 10 using the observed means and variable codings to obtain the coefficients. (In the current example, these model-based means are the same for each group as the respective observed means because the linear and quadratic trends coefficients together perfectly explain the variability among the repeated measures.)

Non-linearity of the growth in the treatment effect observed in both repeated-measures ANOVA and GMA can be examined using this quadratic regression model, from which a score can be predicted for each group at each time. The model-estimated mean difference between the groups at each time can then be calculated separately for each and divided by *SD* to yield the *d* for the treatment effect at that time, and plots of these *d*s against time would fit a straight line only for a correctly specified linear ANOVA/regression/GMA model.

Continuous Outcomes with Moderation of Intervention Effect

Two Independent Groups with Observed Subject Factor Moderation

The rationale for matching of individuals to interventions is that treatment efficacy may be moderated by subject characteristics (Babor & Del Boca, 2003). For example, one sex benefiting more from the intervention than the other.

In a 2 (Group) × 2 (Sex) design, the *d* for the main effect of treatment (group) would traditionally be calculated from the difference in the respective marginal means (i.e., averaged over gender) divided by the pooled within-group *SD* from the four Group × Sex subsamples. The *d* for the interaction effect (expressing the magnitude of the moderation of the treatment factor by sex) would be obtained from the difference between the diagonal means calculated from the 2 × 2 matrix of cell means, which would be divided by *SD*. If the interaction is statistically significant, the main effect of the treatment should be disregarded and the simple effect *d*s calculated for the effects of the intervention as a function of gender by using the within-gender mean differences as numerators in the calculations of those *d*s.

In the alternative GLMM framework approach, two regression coefficients would be added to Equation 2, one for the main effect of sex and a second for the interaction of treatment and sex,

$$Y=a+b_1\text{Group}+b_2\text{Sex}+b_3\text{Group}^*\text{Sex}+e, \quad (11)$$

where treatment group and sex are each coded −1/2 and 1/2 for the two respective levels and the coding for the interaction (Group*Sex) is the cross-products of the codes for the two factors. For the main effect of treatment, $d = b_1/SD$. If b_3 is significant, *d* would then be calculated separately for each sex (simple effect *d*s) using the following pair of equations:

$$d_{\text{M}}=(b_1-.5b_3)/SD.$$

and

$$d_{\text{W}}=(b_1+.5b_3)/SD, \quad (12)$$

where d_{M} is the treatment effect size for men and d_{W} is the treatment effect size for women.

Two Independent Groups with Latent Class Moderation

Cross-sectional statistical models that agglomerate similar participants--such as cluster analysis, latent class analysis (LCA), and latent profile analysis--are useful for incorporating effects associated with unobserved heterogeneity among participants in the sample (Collins & Lanza, 2010; Muthén & Muthén, 2010; Steinley & Brusco, 2011). As with observed groups, latent cluster membership may moderate the effect of the randomized factor. For example, cluster analysis has found that there are two classes of alcoholics, whose members may well respond differently to treatment (Feingold, Ball, Kranzler, & Rounsaville, 1996).

Thus, the analysis would need to examine the interaction between cluster membership and the treatment factor on the outcome in an appropriate solution following a *class enumeration analysis* (see Collins & Lanza, 2010) to determine the number of latent classes that should be extracted. If the Group × Class interaction is statistically significant, the regression of outcome on group can be allowed to vary across classes. Thus, a separate coefficient for the intervention effect could be generated for each class, which can be used to calculate the simple effect *d*s using Equation 3.

Repeated Measures with Observed Subject Factor Moderation

If an observed categorical subject factor, such a gender, moderates the difference in slopes between the treatment and control groups, there is a three-way interaction of the measured factor, the treatment factor, and the polynomial change factor. A term for the three-way interaction is included in Equation 13, which combines the cross-sectional model for subject factor moderation (Equation 11) with the longitudinal model for group differences in trajectories (Equation 8), i.e.,

$$Y=a+b_1\mathrm{Group}+b_2\mathrm{Sex}+b_3\mathrm{Time}+b_4\mathrm{Group}^*\mathrm{Sex}+b_5\mathrm{Group}^*\mathrm{Time}+b_6\mathrm{Time}^*\mathrm{Sex}+b_7\mathrm{Group}^*\mathrm{Sex}^*\mathrm{Time}+e. \quad (13)$$

If the three-way interaction is significant, simple effect *d*s for treatment can be obtained from Equation 13 using the codes of 0 and 1 in the model to generate within-gender coefficients. This is conceptually equivalent to conducting a separate GMA for men and women and using Equation 9 and calculating a *d* for treatment for each sex using the pooled within-group *SD* in the computation of both *d*s.

Repeated Measures with Latent Class Moderation

As with LCA for cross-sectional analysis, a class enumeration analysis must be conducted to determine the number of classes that should be extracted in the GMM. The effect of Group on slope should be allowed to vary across classes if there is a Group × Class interaction. Thus, a separate coefficient for the difference between slopes can be obtained for each class. The within-class effect sizes can be calculated with Equation 9. If the interaction is not significant, the regression coefficient for the Group × Time interaction can be constrained to be the same across classes and the effect size for intervention efficacy calculated from that averaged coefficient.

Cluster Randomized Designs

Two Independent Groups

In some experiments and randomized clinical trials, groups or dyads of individuals (e.g., clinics, therapy groups, or sibling pairs) are randomly assigned to conditions and everyone in the group (called a *cluster*) either receives or does not receive the treatment. The lack of independence among subjects within clusters requires use of a *cluster-randomized design* for significance testing and CI calculations but not for effect size estimations because clustering affects standard errors of the parameters but not their estimates (Hedges, 2007, 2009). Thus, the mean difference between groups in a cluster-randomized design is the same as if individuals rather than clusters had been assigned to conditions in a fully randomized study.

However, there are two types of standard deviations that may be used in the effect size calculations for results from a cluster-randomized design. For determining treatment magnitude within clusters (e.g., clinics), the within-group standard deviation should be based on variation pooled within clusters nested within groups. However, to calculate an effect size for the population from which the total sample is drawn, the within-groups standard deviation should be used (with clustering among subjects ignored). The use of the latter standard deviation yields a zero-order effect size whereas the use of the former yields a partial effect size in which cluster is controlled, and theory rather than statistics should dictate the choice between them. (For a full explication of issues related to effect sizes for cluster-randomized designs, including the selection of the appropriate standard deviation, see Hedges, 2007, 2009).

Repeated Measures Designs

A cluster-randomized design may include repeated measures collected over the course of study that can be examined by GMA. The data then take on a hierarchical structure in which the repeated measures are nested within individuals who are nested within clusters, which can be examined with a *3-Level HLM model* (Raudenbush & Bryk, 2002). In a cluster-randomized repeated-measures design, the intervention is administered to Level-3 units (clusters) in a 3-Level model rather than to Level-2 units (the individuals) in a 2-Level HLM model that is appropriate when individuals are randomly assigned to conditions. At Level 2 of the former, slopes are estimated for each cluster based on the trajectories of individuals in that cluster, and each cluster is assigned a code for its treatment category (e.g., −1/2 = control, 1/2 = treatment). At Level 3, the slopes for the clusters are modeled as a function of the condition to which they were assigned.

The coefficient associated with the effect of treatment on the differences in cluster slopes at level 3 reflects the same difference in growth rate between the experimental and control subjects as b_3 in the 2-level design used when individuals are assigned to conditions. Thus, the effect size can be calculated with Equation 9 using the Group × Time coefficient (i.e., b_3) from Level 3 of the model.

In addition, the 3-Level GMA model is used when participants are randomly assigned to conditions within a cluster (e.g., community). However, in such cases, the effect sizes is based on the Group × Time regression coefficient at Level 2 rather than at Level 3 but calculated in the same way using the Level 2 Group × Time coefficient. Moreover, the previously noted concerns about use of the correct standard deviation apply equally to cluster-randomized repeated-measures designs.

Binary Outcomes

Two Independent Groups

Although a standardized mean difference is an appropriate effect size only in analysis with continuous outcomes, experimental and clinical research often examines binary responses. For instance, students enrolled in a substance abuse prevention program may or may not remain abstinent from drugs. When the outcome is dichotomous, the structural model shifts from a linear regression to a binary logistic regression formulation that models probabilities rather than scores (Hosmer & Lemeshow, 2000). The effect size is then based on the difference between the two groups in proportions, from which probabilities of subjects falling into each category (e.g., staying abstinent) can be estimated.

In probability theory, the *odds* of an event occurring is the probability that the event occurs divided by the probability that the event does not occur (Agresti, 2002). The odds of an event (e.g., attaining a goal) can be calculated separately for each group using the observed proportions and the ratio of the odds between the two groups is the *OR*. For example, if 50% of treated individuals and 80% of controls used drugs within 6-months after the end of a prevention trial, the odds of a treatment recipient relapsing are .50/.50 = 1.00; for a control participant, the odds are .80/.20 = 4. The *OR* for relapse for controls relative to treatment recipients is 4. The *OR* is a commonly used effect size in meta-analysis of group differences in binary outcomes (Borenstein, et al., 2009; Haddock, Rindskopf, & Shadish, 1998).

The structure of the logistic regression model is similar to that of the linear regression model but there are two important differences. First, the logistic regression model does not contain the error term (*e*) found in Equation 2 because the variances of proportions, unlike of means, are known exactly and do not have to be estimated as separate parameters. Second, the outcome (*Y*) is not a raw score but a transformed probability value known as a *logit*, which is the log of the odds of the event occurrence that is modeled in logistic regression (Agresti, 2002). Thus, the *b* from the logistic equivalent of linear regression is not a difference between groups in score means but in logits. The *b* must be exponentiated (i.e., transformed using e^{b}) to yield the associated *OR*, although the conversion is performed by default in logistic regression programs. For binary outcomes in two groups, the *OR* associated with the *b* for the treatment effect can be obtained directly from a logistic regression program or calculated from the frequencies in the 2 × 2 contingency table.

Growth Modeling Analysis

Effect sizes can be calculated for binary outcomes in GMA almost exactly as with continuous ones. However, with binary outcomes, GMA--like logistic regression--models logits instead of raw scores (Raudenbush & Bryk, 2002). The regression coefficient for the Group × Time interaction would thus represent the difference between the two groups in changes over time in logit units, which must be multiplied by study duration to yield the model-derived estimates for the differences between the groups in means of the logits at the end of the study and then exponentiated to an *OR* for interpretation.

Selection of the Standard Deviation

The proposed GLMM framework for effect size assessments focuses on the calculation of the regression coefficients that equal an observed or model-estimated mean difference between two independent groups at end of a study. However, the mean difference must then be divided by the appropriate standard deviation to yield a meaningful standardized mean difference, *d*. Thus, the examples presented implicitly assumed that such standard deviations were used with the regression coefficients to obtain interpretable values of *d*. With the exception of the cluster randomized design, the selection and calculation of the appropriate standard deviation was not addressed, except to note that the standard deviation of the outcomes rather than of difference scores should be used to ensure comparability of effect sizes across different experimental designs—and to allow for Cohen's (1998) suggestions of the values for *d* that constitute small, medium, and large effects to be used across those designs.

The standard deviation chosen for use as the denominator effect size calculations in should be driven by practical and theoretically meaningful considerations applicable to the study. For example, in a short-term (e.g., 12-week) randomized trial of an intervention administered to adults, the standard deviation of the outcome for untreated patients is likely to be constant across assessment waves. In such a case, the baseline standard deviation of all participants could be used (Feingold, 2009), with no need to consider treatment conditions because random assignment ensure that both groups are expected to be equal at the onset of the study. The advantage of the baseline standard deviation is that it can calculated from all subjects (whereas attrition may occur at later waves) and it cannot be influenced by treatment effects (as could be the case with estimates of variability based on multilevel models or by using the within-group standard deviation of observed scores at end of study).

Now consider a possible multi-year prevention study designed to forestall alcohol abuse in adolescents. At the start of the study, when the participants are very young, drinking levels may be very low, resulting in relatively little variability. By the end of the study, when the participants are late adolescents, quantity of alcohol consumed and variation in consumption would be expected to be much higher. The standard deviation of the dependent variable at the end of study would then be more, as use of the smaller baseline standard deviation would result in inflated effect sizes.

Finally, the effects of other subject-level covariates in the models should be considered in the selection of the standard deviation used to calculate the effect size. Take the case of a design consisting of four groups formed by crossing treatment and gender (and where the numerator in the effect size calculation is obtained with Equation 11). If there is a sex difference on the dependent measure, the standard deviation will be greater when the effects of gender are ignored than if the standard deviation is averaged within both gender group, and effect sizes will be larger when using the latter as the denominator. (For further discussion of issues and controversies regarding the standard deviation that should be used in the denominator of the effect size calculation, see Glass, McGaw, & Smith, 1981; Hedges, 2007, 2009).

Issues of Statistical Power in GMA

There are two effect sizes for group differences that can be calculated from repeated-measures analyses (both classical and GMA)--one that uses the standard deviation of the outcome as the denominator (as discussed herein) and the other that uses the standard deviation of change scores (Feingold, 2009; Morris & DeShon, 2002). Power analysis for GMA uses the latter effect size (Feingold, 2009; Muthén & Muthén, 2002; Raudenbush & Liu, 2001), which was not discussed in this article because it is influenced by the correlations among the repeated measures obtained at different time.

Implications for Meta-Analysis

In addition to their value in primary research, effect sizes have been widely used as ingredients in meta-analyses that combine and compare the findings of group differences across studies (Borenstein, et al, 2009; Cooper, 2010; Shadish & Haddock, 2009). Because the *d*s calculated from the GLMM regression framework are all standardized differences between the observed or model-estimated means of two groups (or contrasts comparing sets of two groups), the effect sizes obtained using this approach can be combined irrespective of the design used in the study from which they were calculated (see Feingold, 2009, for more details about using the GMA *d* in meta-analysis). Most important, because linear growth models indicate the treatment effects increase with length of study, duration should be controlled in a meta-analysis, either by grouping studies by their length and performing a separate synthesis for shorter and longer studies, or by combining all studies and including duration as a continuous moderator of effect sizes in the analyses (e.g., Hettema & Hendricks, 2010; Jensen, Cushing, Aylward, Craig, Sorell, & Steele, 2011).

Effect Size Assessments in Non-Experimental Research

Although many meta-analyses have synthesized findings from experiments (including randomized clinical trials), meta-analysis has also been used to examine differences between extant groups (e.g., gender, risk status). For example, meta-analytic investigations have reviewed differences between men and women in cognitive abilities (e.g., Hyde, 2005), personality (e.g., Feingold, 1994), and social behavior (e.g., Eagly, 1995). Although most primary studies of gender differences have been cross-sectional, sex has also been used as a binary time-invariant covariate in GMA (e.g., Huttenlocher, Haight, Bryk, & Seltzer, 1991; Leahey & Guo, 2001).

Randomization Considerations

An important distinction between experimental and non-experimental research is that randomization to conditions is not a characteristic of the latter. In randomized studies, the expected value of the difference between two groups at study onset is zero, making identical the null hypotheses that (1) the groups are equivalent in slopes and (2) the groups have the same means at the end of the study. In contrast, parallel linear trajectories between, say, men and women would coexist with sex differences in means at the end of the study if there were differences between the sexes at its start. Thus, the examinations of group differences in initial status and slopes are both important when the groups are not assumed to be equivalent at the onset of a study. Indeed, when the group factor is not based on randomization, the primary goal of the investigation may be to examine the association of group membership with the person-specific (random) intercepts to explain individual differences in personality or behavior (for an example, see Feingold, Kerr, & Capaldi, 2008). However, to calculate mean differences at end of study between extant groups (which may be due to differences in initial status, growth, or both) from GMA, the coefficient for the effect of group on initial status and the slope should both be used to estimate the end-of-study mean for each group. The difference between these estimated means should be divided by the outcome standard deviation to compare the groups at the end of the study without controlling for baseline differences (Feingold, 2009).

In addition, when randomization is used to form groups, the group factor is uncorrelated with baseline characteristics of subjects, such as SES. Thus, the expected value of the regression coefficient (and the mean difference it equals) for that factor is not changed by the inclusion of one or more baseline variables in the regression equation. Therefore, adjusted mean differences (i.e., mean differences after other variables are controlled) estimate the same parameter as the observed mean differences, and the inclusion of such subjects variable covariates in the model can be ignored when calculating the effect size for the randomized factor. However, this is not true for analysis using naturally-occurring groups. For example, if ethnicity is correlated with SES, the inclusion of SES in a regression model would change (and most likely reduce) the coefficient for ethnicity because of shared variance between the two variables. Thus, it would generally be inadvisable to use such a partial effect size in a meta-analysis of racial differences (e.g., Roberts, Cash, Feingold, & Johnson, 2006).

Continuous Covariates

Another difference between experimental and non-experimental research is that the latter is more likely to include predictors that are continuous (e.g., age) rather than categorical (e.g., condition), for which the effect size is often the correlation coefficient (r). Correlations also can be used in meta-analysis (Schmidt, Lee, & Oh, 2009). In GMA, a continuous non-manipulated level-2 predictor (a time-invariant covariate) can be related to both the random intercept and random slope of the outcome variable in a multilevel model. Although the GLMM framework can subsume associations in which the independent variable is continuous, the interpretation is more complex because the regression coefficient is no longer a mean difference but a difference in rate of change of the outcome--or of intercepts and slopes in GMA--as a function of rate of change on the continuous covariate. Effect sizes involving a continuous predictor are less likely to be expressed in the d metric than in a metric based on proportion of variance explained, particularly for effects in mediation analysis (e.g., Preacher & Kelley, 2011).

Discussion

An equation with a single regression coefficient is sufficient for some effect size calculations in cross-sectional (i.e., between-subjects) data. However, in longitudinal analysis, the structural model always include coefficients for both time and the Group × Time interaction, with the latter being the coefficient for the effect of group on slope in multilevel analysis that represents the difference in rate of change per unit of time (e.g., week) between the two groups. The product of this *b* and length of study is a model-derived estimate of the difference between the baseline-adjusted means of the two groups at the end of the study, which can be divided by the outcome measure's *SD* to estimate the same effect size parameter estimated in a between-subjects study. Elaborations of the regression model are required when, among other situations, there are more than two groups, trajectories are non-linear, or the experimental factor is moderated by observed categorical variables or latent classes. This article introduces a GLMM framework that unifies effect sizes for comparisons between two groups, or contrasts among multiple groups, in a common *d* or *OR* metric across a wide gamut of cross-sectional and longitudinal designs.

Moreover, the concepts discussed in this paper apply to all types of GMA, as well as to classical repeated measures analyses (ANOVA/multiple regression analysis), for which effect size equations have also been lacking. Note that recent simulation studies have shown that the *OR* is biased in small sample sizes (Nemes, Jonasson, Genell, & Steineck, 2009), although this is unlikely to be a concern when the *OR* is derived from GMA, as such analysis requires large samples for parameter estimation. Although it is possible to convert *OR*s to equivalent *d*s (e.g, Chinn, 2000; Haddock, Rindskopf, & Shadish, 1998; Sánchez-Meca, Marín-Martínez, & Chacón-Moscoso, 2003), such transformations are only recommended when conducting a meta-analysis of studies that vary in the distributions of their outcome variables.

The conceptual model uses the structure from simple linear regression to calculate an effect size from familiar statistics: (a) the regression coefficient, (b) study duration (for longitudinal designs), and (c) the standard deviation of the dependent measure (see Table 2 for overview of regression equations used in the framework). The effect size in a between-subjects analysis is computed by coding the two groups with values differing by one unit so that the regression coefficient is the mean difference between them, which is then divided by the pooled within-group standard deviation (or whatever type of standard deviation is preferred by the investigator) of continuous outcomes. In longitudinal analysis (repeated measures ANOVA, multiple regression, and GMA), the regression coefficient for the group difference in rate of change must be multiplied by study duration and that product is divided by the standard deviation of the outcome measure. Although both equations generate *d*, the key difference is that in multiple regression analysis and ANOVA the group difference at end of the study is based on observed or estimated baseline-adjusted means of study completers, whereas in GMA it is based on adjusted means predicted by a GMA model that uses data from all participants.

Acknowledgments

This work was supported by National Institutes of Health grants: RC1DA028344 from the National Institute on Drug Abuse and R01AA018669 from the National Institute on Alcohol Abuse and Alcoholism.

References

Agresti, A. Categorical data analysis. 2nd ed.. New York: Wiley; 2002.

Aiken LS, West SG, Millsap RE. Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of Ph.D. programs in North America. American Psychologist. 2008; 63:32–50. [PubMed: 18193979]

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. Reporting standards for research in psychology: Why do we need them? What might they be? American Psychologist. 2008; 63:839–851. [PubMed: 19086746]

Atkins DC. Using multilevel models to analyze couple and family treatment data: Basic and advanced issues. Journal of Family Psychology. 2005; 19:98–110. [PubMed: 15796656]

Babor, TF.; Del Boca, FK., editors. Treatment matching in alcoholism. New York: Cambridge University Press; 2003.

Becker BJ. Synthesizing standardized mean change scores. British Journal of Mathematical and Statistical Psychology. 1988; 41:257–278.

Borenstein, M.; Hedges, LV.; Higgins, JPT.; Rothstein, HR. Introduction to meta-analysis. New York: Wiley; 2009.

Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. Statistics in Medicine. 2000; 19:3127–3131. [PubMed: 11113947]

Cohen, J. Statistical power analysis for the behavioral sciences. 2nd ed.. Hillsdale, NJ: Erlbaum; 1988.

Cohen, P.; Cohen, J.; West, SG.; Aiken, LS. Applied multiple regression/correlation analysis for the behavioral analysis. 3rd ed.. Mahwah, NJ: Erlbaum; 2003.

Collins, LM.; Lanza, ST. Latent class and latent transition analysis. Hoboken, NJ: Wiley; 2010.

Cooper, HM. Research synthesis and meta-analysis. 4th ed.. Los Angeles: Sage; 2010.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical SocietySeries B. 1977; 39:1–38.

Dunlap WP, Cortina JM, Vaslow JB, Burke MJ. Meta-analysis of experiments with matched groups or repeated measures designs. Psychological Methods. 1996; 1:170–177.

Eagly AH. The science and politics of comparing women and men. American Psychologist. 1995; 50:145–158.

Fairchild AJ, MacKinnon DP, Taborga MP, Taylor AB. R^2 effect-size measures for mediation analysis. Behavior Research Methods. 2009; 41:486–498. [PubMed: 19363189]

Feingold A. Gender differences in personality: A meta-analysis. Psychological Bulletin. 1994; 116:429–456. [PubMed: 7809307]

Feingold A. Effect sizes for growth-modeling analysis for controlled clinical trials in the same metric as for classical analysis. Psychological Methods. 2009; 14:43–53. [PubMed: 19271847]

Feingold A, Ball SA, Kranzler HR, Rounsaville BJ. Generalizability of the Type A/Type B distinction across different psychoactive substances. American Journal of Drug and Alcohol Abuse. 1996; 22:449–462. [PubMed: 8841691]

Feingold A, Kerr DCR, Capaldi DM. Associations of substance use problems with intimate partner violence in long-term relationships. Journal of Family Psychology. 2008; 22:429–438. [PubMed: 18540771]

Feingold A, Ball SA, Kranzler HR, Rounsaville BJ. Generalizability of the Type A/Type B distinction across different psychoactive substances. American Journal of Drug and Alcohol Abuse. 1996; 22:449–462. [PubMed: 8841691]

Fleiss, JL.; Berlin, JA. Effect sizes for dichotomous data. In: Cooper, H.; Hedges, LV.; Valentine, JC., editors. The handbook of research synthesis. 2nd ed.. New York: Russell Sage; 2009. p. 237-253.

Gibbons RD, Hedeker D, Elkin I, Waternaux CM, Kraemer HC, Greenhouse JB, et al. Some conceptual and statistical issues in analysis of longitudinal psychiatric data. Archives of General Psychiatry. 1993; 50:729–750.

Glass, GV.; McGaw, B.; Smith, ML. Meta-analysis in social research. Thousand Oaks, CA: Sage; 1981.

Grissom, RJ.; Kim, JJ. Effect sizes for research: A broad practical approach. NY: Erlbaum; 2005.

Gueorguieva R, Krystal JH. Move over ANOVA: Progress in analyzing repeated-measures data and its reflection in papers published in the *Archives of General Psychiatry* . Archives of General Psychiatry. 2004; 61:310–317. [PubMed: 14993119]

Haddock CK, Rindskopf D, Shadish WR. Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. Psychological Methods. 1998; 3:339–353.

Hedeker, D.; Gibbons, RD. Longitudinal data analysis. Hoboken, NJ: Wiley; 2006.

Hedges LV. Effect sizes in cluster-randomized designs. Journal of Educational and Behavioral Statistics. 2007; 32:341–370.

Hedges, LV. Effect sizes in nested designs. In: Cooper, H.; Hedges, LV.; Valentine, JC., editors. The handbook of research synthesis. 2nd ed.. New York: Russell Sage; 2009. p. 337-356.

Hettema JE, Hendricks PS. Motivational interviewing for smoking cessation: A meta-analytic review. Journal of Consulting and Clinical Psychology. 2010; 78:868–884. [PubMed: 21114344]

Hosmer, DW.; Lemeshow, S. Applied logistic regression. 2nd ed.. New York: Wiley; 2000.

Huttenlocher J, Haight W, Bryk A, Seltzer M. Early vocabulary growth: Relation to language input and gender. Developmental Psychology. 1991; 27:235–249.

Hyde JS. The gender similarities hypothesis. American Psychologist. 2005; 60:581–592. [PubMed: 16173891]

Jensen CD, Cushing CC, Aylward BS, Craig JT, Sorell DM, Steele RG. Effectiveness of motivational interviewing interventions for adolescent substance use behavior change: A meta-analytic review. Journal of Consulting and Clinical Psychology. 2011; 79:433–440. [PubMed: 21728400]

Johnson W, Carothers A, Deary IJ. A role for the X chromosome in sex differences in variability in intelligence? Perspectives on Psychological Science. 2009; 4:598–611.

Kuljanin G, Braun MT, DeShon RP. A cautionary note on modeling growth trends in longitudinal data. Psychological Methods. 2011; 16:249–264. [PubMed: 21517180]

Leahey E, Guo G. Gender differences in mathematical trajectories. Social Forces. 2001; 80:713–732.

Lipsey, MW.; Wilson, DB. Practical meta-analysis. Thousand Oaks, CA: Sage; 2001.

Little, RJA.; Rubin, DB. Statistical analysis with missing data. 2nd ed.. New York: Wiley; 2002.

McCulloch, CE.; Searle, SR. Generalized, linear, and mixed models. New York: Wiley; 2001.

Meredith W, Tisak J. Latent curve analysis. Psychometrika. 1990; 55:107–142.

Morris SB. Estimating effect sizes from pretest-posttest-control group designs. Organizational Research Methods. 2008; 11:364–386.

Morris SB, DeShon RP. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. Psychological Methods. 2002; 7:105–125. [PubMed: 11928886]

Muthén B, Brown CH, Masyn K, Jo B, Khoo S-T, Yang C-C, et al. General growth mixture modeling for randomized preventive interventions. Biostatistics. 2003; 3:459–475. [PubMed: 12933592]

Muthén KL, Muthén BO. How to use a Monte Carlo study to decide on sample size and determine power. Structural Equation Modeling. 2002; 4:599–620.

Muthén, LK.; Muthén, BO. Mplus user’s guide. 6th ed.. Los Angeles, CA: Muthén and Muthén; 2010.

Nagin, DS. Group-based modeling of development. Cambridge: Harvard University Press; 2005.

Nelder JA, Wedderburn RWM. Generalized linear models. Journal of the Royal Statistical Society Series A. 1972; 135:370–384.

Nemes S, Jonasson MJ, Genell A, Steineck G. Bias in odds ratios by logistic regression modelling and sample size. BMC Medical Research Methodology. 2009; 9:56–61. [PubMed: 19635144]

Odgaard EC, Fowler RL. Confidence intervals for effect sizes: Compliance and clinical significance in the *Journal of Consulting and Clinical Psychology* . Journal of Consulting and Clinical Psychology. 2010; 78:287–297. [PubMed: 20515205]

Olejnik S, Algina J. Generalized eta and omega squared statistics: Measures of effect size for some common research designs. Psychological Methods. 2003; 8:434–447. [PubMed: 14664681]

Preacher KJ, Kelley K. Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. Psychological Methods. 2011; 16:93–115. [PubMed: 21500915]

Raudenbush, SW.; Bryk, AS. Hierarchical linear models: Applications and data analysis methods. 2nd ed.. Thousand Oaks, CA: Sage; 2002.

Raudenbush, S.; Bryk, A.; Cheong, Y.; Congdon, R.; du Toit, M. HLM 6: Hierarchical linear and nonlinear modeling. Lincolnwood, IL: Scientific Software International; 2004.

Raudenbush SW, Liu X. Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. Psychological Methods. 2001; 6:387–401. [PubMed: 11778679]

Rosenthal, R.; Rosnow, RL.; Rubin, DB. Contrasts and effect sizes in behavioral research: A correlational approach. Cambridge, England: Cambridge University Press; 2000.

Sánchez-Meca J, Marín-Martínez F, Chacón-Moscoso S. Effect-size indices for dichotomized outcomes in meta-analysis. Psychological Methods. 2003; 8:448–467. [PubMed: 14664682]

Schafer JL, Graham JW. Missing data: Our view of the state of the art. Psychological Methods. 2002; 7:147–177. [PubMed: 12090408]

Shadish, WR.; Haddock, CK. Combining estimates of effect sizes. In: Cooper, H.; Hedges, LV.; Valentine, JC., editors. The handbook of research synthesis. 2nd ed.. New York: Russell Sage; 2009. p. 257-277.

Singer, JD.; Willett, JB. Applied longitudinal data analysis: Modeling change and event occurrence. New York: Oxford; 2003.

Steinley D, Brusco MJ. Choosing the number of clusters in K-means clustering. Psychological Methods. 2011; 16:285–297. [PubMed: 21728423]

Supplee LH. The application of effect sizes in research on children and families [Special section]. Child Development Perspectives. 2008; 3:164–205.

Winer, BJ. Statistical principles in experimental design. 2nd ed.. New York: McGraw-Hill; 1971.

Wright DB. Ten statisticians and their impacts for psychologists. Perspectives on Psychological Science. 2009; 4:587–597.

Table 1

Comparison of GLM, GLMS, and GLMM models

Model	Estimation	Effects	Dependent Variable Distribution	Sample Statistical Analyses
GLM	OLS	Fixed	Continuous	ANOVA, Multiple regression analysia
GLMS	ML	Fixed	Continuous, Categorical, Count	Logistic regression, LCGA
GLMM	ML	Fixed, Random	Continuous, Categorical, Count	GMA/HLM, GMM

Note. GLM = General Linear Model, OLS = Ordinary Least Squares, ANOVA = Analysis of Variance, GLMS = General Linear Models, ML = Maximum Likelihood, LCGA = Latent Class Growth Analysis, GLMM = Generalized Linear Mixed Models, GMA = Growth Modeling Analysis, HLM = Hierarchical Linear Models, GMM = Growth Mixture Modeling.

Table 2

Summary of regression equations derived from GLMM framework for calculations of GMA *d* from different research designs

Equation	Design Characteristics
$Y = a + b\text{Group} + e$ (2)	2 groups, 1 time point (no pretest)
$Y = a + b_1\text{Pretest} + b_2\text{Group} + e$ (5)	2 groups, 2 time points (pretest and posttest)
$Y = a + \mathrm{b}_1\text{Group1} + b_2\text{Group2} + e$ (7)	3 or more groups, 1 time point, orthogonal a priori contrasts
$Y = a + b_1\text{Group} + b_2\text{Time} + b_3\text{Group*Time} + e$ (8)	2 groups, 3 or more time points, linear model
$Y = a + b_1\text{Group} + b_2\text{Linear} + b_3\text{Linear*Group} + b_4\text{Quadratic} + b_5\text{Quadratic*Group} + e$ (10)	2 groups, 3 or more time points, non-linear model
$Y = a + b_1\text{Group} + b_2\text{Sex} + b_3\text{Group*Sex} + e$ (11)	2 groups, 1 time point, subject moderator (e.g., sex)
$Y = a + b_1\text{Group} + b_2\text{Sex} + b_3\text{Time} + b_4\text{Group*Sex} + b_5\text{Group*Time} + b_6\text{Time*Sex} + b_7\text{Group*Sex*Time} + e$ (13)	2 groups, 3 or more time points, subject moderator

Note. Equation numbers are in parentheses following equations.