Published in final edited form as:
J Consult Clin Psychol. 2015 February ; 83(1): 157–168. doi:10.1037/a0037721.

Confidence Interval Estimation for Standardized Effect Sizes in Multilevel and Latent Growth Modeling

Alan Feingold
Oregon Social Learning Center

Abstract

Objective—Multilevel and latent growth models are frequently used interchangeably to examine differences between groups in trajectories of outcomes from controlled clinical trials. The unstandardized coefficient for the effect from group to slope (the treatment effect) from such models can be converted to a standardized mean difference (Cohen's *d*) between the treatment and control groups at end of study. This article addresses the confidence interval (CI) for this effect size.

Method—Two sets of equations for estimating the CI for the treatment effect size in multilevel models were derived and their usage was illustrated with data from the National Youth Study. Validity of the CIs was examined with a Monte Carlo simulation study that manipulated effect potency and sample size.

Results—The equivalence of the two new CI estimation methods was demonstrated and the Monte Carlo study found that bias in the CI for the effect size were not appreciably larger than bias in the CI for the widely used unstandardized coefficient.

Conclusions—Investigators reporting this increasingly popular effect size can estimate its CI with equations presented in this article.

Keywords

clinical trials; effect sizes; confidence intervals; multilevel analysis; latent growth models; hierarchical linear models

Studies of intervention efficacy have historically used classical analysis, especially analysis of variance (ANOVA) and multiple regression analysis (Cohen, Cohen, West, & Aiken, 2003), to compare changes between treatment and control groups. However, the well-worn methods based on the general linear model and ordinary least squares are being supplanted by *growth modeling analysis* (GMA), which typically uses the EM algorithm (Dempster, Laird, & Rubin, 1977) and maximum likelihood estimation. The GMA family includes multilevel modeling/hierarchical linear models (MLM/HLM; Hedeker & Gibbons, 2006; Raudenbush & Bryk, 2002) and latent growth modeling (LGM; Bollen & Curran, 2006; Singer & Willett, 2003), which are often used interchangeably to evaluate interventions by

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

comparing trajectories of outcomes between the treatment and control groups over the course of a randomized clinical trial (Feingold, 2009; Gueorguieva & Krystal, 2004).

The differences between MLM and LGM are more conceptual than mathematical (Curran, 2003; Preacher, Wichman, MacCallum, & Briggs, 2008) and non-parallel slopes in a correctly specified linear GMA obtained with either latent variable approach indicate that the model predicts that the two groups that were expected to be comparable at study onset (baseline) because of randomization differ on the outcome at the end of the study. The treatment effect from a GMA is conventionally defined as the difference in rate of growth between the groups (i.e., the effect from group to slope) and power assessments for planned GMA studies examine the power to detect this difference (e.g., Muthén & Muthén, 2002; Raudenbush & Liu, 2001).

Although the importance of effect sizes has long been recognized (Cumming, 2013; Grissom & Kim, 2012; Olejnik & Algina, 2000), the first studies that used GMA typically reported only null hypothesis significance tests because of a lack of consensus regarding both the conceptualization and calculation of effect sizes from a GMA (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008; APA, 2009). Feingold (2009) reviewed 43 controlled clinical trials that used GMA and were published in *Journal of Consulting and Clinical Psychology* (*JCCP*) from 2003 through 2006 inclusive and found that only 30% of them reported model-based effect sizes for the intervention effects. Moreover, the equations used to calculate effect sizes were not conceptually or mathematically equivalent across studies.

However, Feingold (2009) used the term *effect size* to refer to a standardized statistic. Although effect sizes need not be standardized (Baguley, 2009; Kelley & Rausch, 2011), standardized effect sizes are often invaluable-- particularly in research synthesis--because of the arbitrary metrics used in psychology (Blanton & Jaccard, 2006). If, for instance, two scales that operationalize the identical construct differ in number of items, the scale with the larger number of items will generally have a larger variance in the same sample (Nunnally, 1978). In addition, the mean difference between two groups will typically be greater on the scale with the larger variance when the groups are compared on both measures. Such measurement artifacts make research synthesis of findings from independent studies problematical because different studies generally operationalize the same construct with scales that vary in numbers and types of items. Thus, *d* was a central statistic for Glass (1976) when he introduced meta-analysis and used it to examine the efficacy of psychotherapy (Glass, McGaw, & Smith, 1981).

Standardized Effect Sizes for Multilevel and Latent Growth Models

Drawing on ideas in Raudenbush and Liu (2001), Feingold (2009) formulated an equation for an effect size for the difference between the treatment and control groups in linear slopes from a GMA (MLM or LGM) that transforms the coefficient for the slope difference into a standardized mean difference (Cohen's *d*) between groups at the end of the study,

$$d = (b * \text{duration})/SD, \quad (1)$$

where (a) *b* is the unstandardized coefficient for the effect of group (treatment vs. control, dummy coded) that is the difference in rate of change between conditions on a continuous outcome per unit of time (e.g., per week, given weekly assessments that are coded to differ by one point between them, such as 0, 1, 2, and 3), (b) duration (number of time points minus one when time codes differ by one point) is the length of the study based on units associated with the regression coefficient (e.g., number of weeks if *b* represents the difference in rate of change per week), and (c) *SD* is the pooled within-group standard deviation of the outcome measure (*Y*).

Equation 1 may appear to be radically different than the formula for Cohen's (1988) classical *d*, defined as the difference between the means of two independent groups at the end of the study divided by the pooled within-group *SD*,

$$d=(M_{\mathrm{T}}-M_{\mathrm{C}})/SD, \quad (2)$$

where M_{T} is the mean of the treatment group and M_{C} is the mean of the control group. However, the difference between the two equations, which has a denominator that should estimate the same parameter, is more apparent than real. The means of two independent groups measured at a single time can be compared with a multiple regression analysis, which is mathematically equivalent to both the *t* test for independent groups and the between-subjects ANOVA with two groups (Cohen et al., 2003) with which the classical *d* is most often associated. In regression analysis, group would be a dichotomous independent variable captured by codes differing by one point (e.g., 0 and 1, or -.5 and .5) and the unstandardized regression coefficient (*b*) would then equal the mean difference between the two groups at the end of the study. Thus, given a regression analysis of data from two independent groups,

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

which is mathematically equivalent to the *d* calculated with Equation 2.

In addition, the product forming the numerator in the Equation 1 is the model-based estimate of the difference in means between the two groups at the end of the study, which has the same expected value as that of the difference between the means of the two groups at the final time point (given random assignment to conditions and a correctly specified linear GMA). Moreover, the GMA effect size is independent of the intercorrelations among the repeated measures over time, which meta-analysts have shown to be an important criterion for the *d* from classical analysis of repeated-measures data to be expressed in the same metric as the *d* calculated from independent (completely randomized) groups designs (Becker, 1988; Dunlap, Cortina, Vaslow, & Burke, 1996; Morris, 2008; Morris & DeShon, 2002). The GMA *d*, as calculated with Equation 1, is now frequently found in studies of intervention efficacy, particularly in *JCCP* (e.g., Aderka, Gillihan, McLean, & Foa, 2013; Arch, Eifert, Davies, Vilardaga, Rose, & Craske, 2012; Chaffin, Funderburk, Bard, Valle, & Gurwitch, 2011; Kerr, DeGarmo, Leve, & Chamberlain, 2014; Ljótsson et al., 2013; Safren, O'Cleirigh, Bullis, Otto, Stein, & Pollack, 2012; Twohig, Hayes, Plumb, Pruitt, Collins, Hazlett-Stevens, & Woidneck, 2010).

Confidence Intervals for Effect Sizes for Multilevel and Latent Growth Models

Current publishing practices in psychology require reports of confidence intervals (CIs) for effect sizes (APA, 2009; Kelley & Preacher, 2012). Indeed, *JCCP* recently published a review on CIs for clinical findings to facilitate compliance with this mandate (Odgaard & Fowler, 2010) but its authors noted that it could not address CIs for effect sizes from GMA because methods for their estimation were not yet available. For example, equations for the CI for the GMA *d* did not accompany the introduction of the effect size (Feingold, 2009).

The traditional approach to the calculation of the CI of a statistic entails the multiplication of the statistic's standard error (*SE*) by a critical *t* or *z* value (e.g., 1.96 for the 95% CI) and then adding and subtracting the product to the point estimate to yield lower and upper confidence limits (CLs). This approach assumes a normal distribution of the effect size and a symmetrical CI, i.e., the difference between the lower CL (LCL) and the point estimate equals the difference between the upper CL (UCL) and the point estimate. An example of this approach is found in estimation of the 95% CI for classical *d* (i.e., *d* calculated with Equation 2 or 3) presented in Borenstein, Hedges, Higgins, and Rothstein's (2009) text on meta-analysis,

$$\mathrm{LCL}_d = d - 1.96(SE) \quad (4)$$

and

$$\mathrm{UCL}_d = d + 1.96(SE), \quad (5)$$

where *SE* is the square root of the estimate of the variance (*v*) of *d* that can be obtained with Equation 6,

$$v = (N/n_1 n_2) + (d^2/2N), \quad (6)$$

n_1 is the sample size of group 1, n_2 is the sample size of group 2, and *N* is the total sample size (i.e., $n_1 + n_2$).

Borenstein et al. (2009) noted that Equations 4 and 5 provide only an approximate CI for *d* but contend that the CI estimates obtained with them are accurate enough to warrant their adaptations for use in meta-analysis. Some methodologists, however, favor use of more precise CIs for effect sizes (e.g., Steiger, 2004). CI estimation may also use *bootstrap* or *bias-corrected bootstrap* methods (Efron & Tibshirani, 1993; Goldstein, 2011). Although Mplus (Muthén & Muthén, 2012), for example, uses the *SE* as the default method for CI estimation, the program can also produce bootstrap (and corrected bootstrap) CIs using case sampling with replacement. Specialized programs have been developed to estimate the CI for the unstandardized difference between groups in slopes from GMA (Kelley & Rausch, 2011). The nonparametric approaches, which produce asymmetrical CIs, use noncentrality parameters associated with the *t* distribution to obtain CIs for standardized mean differences (see review by Odgaard & Fowler, 2010, which includes equations in the appendix).

However, GMA is typically conducted with large samples and treatment effects are typically modest, especially in large samples. These are the conditions under which CIs for effect sizes are most accurately assessed with commonly used approximation procedures based on their *SE*s (Algina & Kesselman, 2003; Hedges & Olkin, 1985; Odgaard & Fowler, 2010). This article presents and validates two new sets of mathematically equivalent equations for the estimation of the CI for the GMA *d*, the first of which can subsume CI estimation methods for the GMA *d* that do not rely on the *SE* of *b* (SE_b).

Transformation of CI for *b* to CI for GMA *d*

The CI for the unstandardized coefficient (*b*)for the slope difference can be obtained by extant GMA software. The respective CI for the GMA *d* can be estimated simply by substituting the reported LCL and UCL of *b* for *b* in the numerator of Equation 1, thus converting the CI for *b* to the CI for *d*,

$$\mathrm{LCL}_d = (\mathrm{LCL}_b * \mathrm{duration})/SD, \quad (7)$$

and

$$\mathrm{UCL}_d = (\mathrm{UCL}_b * \mathrm{duration})/SD. \quad (8)$$

An apparent disadvantage of this approach is that the *SE* of the GMA *d* is not generated or used directly but is nonetheless needed for examination of bias in the CI for *d*, and for *d* to be included in a meta-analysis. However, given the calculation of a symmetrical CI for *d*, the *SE* of *d* can be determined from its UCL by rearranging the terms in Equation 5 to solve for *SE*,

$$SE = (\mathrm{UCL}_d - d)/1.96, \quad (9)$$

and $v = SE^2$.

Variance-calculation alternative for CI estimation for GMA *d*

An alternative approach for CI estimation uses the *SE* of *d* and a critical *z* value (i.e., Equations 4 and 5). The terms in Equation 1 can be rearranged to define the GMA *d* as the product of the coefficient for the group difference in slopes (*b*) and a standardizing factor (duration/*SD)*. Statistical theory dictates that the multiplication of a random variable (e.g., the sampling distribution of *b* for the effect of group on slope) by a constant produces a random variable that is a linear transformation of the original variable and has a variance that is the product of the variance of the original variable multiplied by the square of the constant (Hodges & Lehmann, 2005). Because the variance of *b* outputted by statistical software is the square of SE_b, v^* (the sample variance of *d* when σ is known) for the GMA *d* can be calculated from,

$$v* = SE_b^2 * (\mathrm{duration}/\sigma)^2. \quad (10)$$

Because σ is ordinarily unknown, however, it has to be estimated from *SD*,

$$v = SE_b^2 * (\text{duration}/SD)^2. \quad (11)$$

The CI for the GMA effect size can then be estimated with Equations 4 and 5 and the *SE* calculated as the square root of *v* obtained with Equation 11. (Whereas σ is often used in GMA methods' expositions to denote the Level-1 residual variance, σ in Equation 10 is the parameter for the pooled within-group standard deviation of *Y*.)

Note that *v* can be a good estimator of *v** only when the sample size is large and *SD* is an accurate estimate of σ. With small samples, *SD* of the outcome would vary notably across GMA samples drawn from the same population. Thus, duration/*SD* is not a constant over the sampling distribution of *b* but a random variable with a variance that decreases (and approaches zero) as sample size per replication increases.

Illustration of CI Estimation for the GMA *d* with Two Equivalent Methods

Descriptive statistics from MLM study

Raudenbush (1995; Raudenbush & Liu, 2001) examined longitudinal data from the National Youth Study (Elliot, Huizinga, & Menard, 1989) to illustrate the use of MLM. The participants were 122 boys and 117 girls whose attitude towards deviance was measured five times over a four-year period at ages 11, 12, 13, 14, and 15. Time was mean centered by Raudenbush (1995) with codes of -2, -1, 0, 1, and 2, respectively. Gender was a dichotomous time-invariant covariate (men vs. women) like treatment condition (intervention vs. control) in a randomized clinical trial.

The MLM found $b = .0112$ ($SE_b = .0100$). The within-group *SD* of the attitudinal measure, based on observed data at age 15, was .30 (Raudenbush & Lui, 2001).

The GMA *d* is calculated with Equation 1,

$$d = (.0112 * 4)/.30 = .1493.$$

The CLs for *b* for the slope difference are:

$$\text{LCL}_b = .0112 - (1.96 * .0100) = -.0084$$

and

$$\text{UCL}_b = .0112 + (1.96 * .0100) = .0308.$$

Transformation of CI for GMA *b* to CI for GMA *d*

The CLs for the effect size are calculated with the Equations 7 and 8,

$$\text{LCL}_d = (-.0084 * 4)/.30 = -.112$$

and

$$\mathrm{UCL}_d = (.0308 * 4)/.30 = .411.$$

SE is calculated from the CI using Equation 9,

$$SE = (.4107 - .1493)/1.96 = .133$$

and

$$v = SE^2 = .1334^2 = .018.$$

Variance-calculation alternative for CI estimation for GMA *d*

The *v* is calculated with Equation 11,

$$v = .0100^2 * (4/.30)^2 = .018$$

and

$$SE = .018^{1/2} = .133.$$

Next, the CLs for *d* are calculated using Equations 4 and 5,

$$\mathrm{LCL}_d = .1493 - (1.96 * .133) = -.112$$

and

$$\mathrm{UCL}_d = .1493 + (1.96 * .133) = .411.$$

Most important, the *d* and *v* obtained with each set of equations were calculated from published statistics rather than from analysis of raw data, thereby demonstrating the utility of the equations in meta-analytic applications (for use of *d* and *v* in meta-analysis, see Borenstein et al., 2009; Cooper, Hedges, & Valentine, 2009; Lipsey & Wilson, 2001; Shadish & Haddock, 2009).

The Current Work

The proposed equations for the CI for GMA *d* yield only approximations because the CI for the unstandardized coefficient itself is an estimate in smaller samples, and the transformation of CI for *b* to CI for *d* imparts additional imprecision when σ is not known and estimated by *SD*. Errors in statistics obtained with approximations are manifested as bias that can be assessed with Monte Carlo simulation studies (e.g., Cheung. 2009; Hedges,

Pustejovskya, & Shadish, 2012; Lau & Cheung, 2012; MacKinnon, Lockwood, & Williams, 2004).

To obtain preliminary validation for the new equations, Monte Carlo analyses were conducted to examine biases in the *SE* and CI in both unstandardized coefficients and respective effect sizes as function of sample size and treatment potency, both of which were expected to affect CI bias. Subtracting the bias from the effect sizes from those for the respective unstandardized coefficients indicate bias added from the transformation process. Finally, although equations for CI estimation are presented for GMAs comparing two groups on a continuous outcome, extensions to randomized clinical trials conducted with multiple groups, multiple sites, additional covariates, binary outcomes, and cluster randomization are discussed.

Method

Muthén and Muthén (2002) illustrated the use of Mplus to conduct Monte Carlo simulations to evaluate biases in point estimates and *SE*s of coefficients for small (.10) and medium (.20) slope differences between two groups in linear LGMs with random intercepts and random slopes. Each model included a single dichotomous time-invariant covariate (e.g., group) as the (Level 2) predictor of variations in intercepts and slopes with a continuous outcome. Because the first phase of the current work could be accomplished with these Monte Carlo models, the Mplus input statements--which were reported in their entirety in Muthén and Muthén (2002)--were adapted for use in assessing biases in both the unstandardized coefficients and respective effect sizes. However, whereas their tutorial used a single *N* for each model, this study manipulated sample size, specifying *N*s of 50, 100, 150, 250, and 500 for each of the two coefficients for slope differences. The 10 models used 4-equidistant time points coded to differ by one unit between measurement occasions, and residual variances specified were also taken from Muthén and Muthén: .25 for the intercepts, .09 for the slopes, and .50 for model residuals (i.e., Level-1 variances, homogeneous across time). The current simulations also used 10,000 replications per analysis.

Biases in the point estimates and *SE*s for the unstandardized coefficients and effect sizes were examined following the steps outlined in Muthén and Muthén (2002), which also provided guidelines for acceptable degrees of bias: 5% or less for point estimates and 10% or less for *SE*s. Mplus evaluates CI accuracy with the *coverage coefficient*: the proportion of the replications for which the CI contains the true parameter value (Muthén & Muthén, 2012). Perfect coverage for the 95% CI is .95, and acceptable values were said to be in the . 91 to .98 range. Bias in each CI can thus be estimated by subtracting the coverage coefficient from .95, with acceptable values in the −.03 to .04 range.

Calculating Pooled σ of *Y*4 from the Monte Carlo Population Model

Before Monte Carlo analyses could be conducted to examine biases in effect sizes, δ and σ associated with Monte Carlo model parameters for the unstandardized coefficients and residual variances had to be defined, with σ determined first because it was needed to obtain δ. The residual variance of .25 specified for the intercept (the within-group σ^2 of a latent "true score" of *Y* at the occasion for which time is coded 0) was added to the Level-1

residual variance of .50 to obtain the within-group σ^2 of .75. Thus, σ is the square root of .75 = .866, which is the pooled within-group σ when time = 0. Because the Mplus Monte Carlo procedure generates growth data for this model with within-group σ^2s for *Y* that are generally heterogeneous across time, the within-group σ^2 at the final time point (i.e., for *Y*4) was needed to obtain a GMA *d* equivalent to classical *d*, as the latter is determined using outcome dispersion measured at the final (only) time. Therefore, the current Monte Carlo studies used end-centering of the time variable (Muthén & Muthén, 2000) by assigning codes of -3, -2, -1, and 0 for time for T1 (*Y*1), T2 (*Y*2), T3 (*Y*3), and T4 (*Y*4), respectively (see Appendix A for one of the input statements used in the Monte Carlo study). Note that different parameterizations (centerings of time) do not have any impact on effects involving slopes (Muthén & Muthén, 2000) but determine whether the within-group σ^2 is for *Y*1 or *Y*4 (unless the within-group σ^2 s are homogeneous over time and σ^2 of *Y*1 = σ^2 of *Y*4).

Calculating δ for *Y*4 from Monte Carlo Parameter Specifications

Given specification in the statements for each Monte Carlo model of 4 time points differing by 1 unit, duration = 3 for all models. Therefore, using a population equivalent of Equation 1,

$$\delta = (\beta * \text{duration})/\sigma, \quad (12)$$

where β is the parameter for the coefficient for the slope difference. Thus,

$$\delta = (.10 * 3)/.866 = .3464,$$

for the five models with β = .10, and

$$\delta = (.20 * 3)/.866 = .6928,$$

for the five models with β = .20.

Validation of δ and σ in the Monte Carlo Models

To verify that the calculated parameters for δ and σ were correct for the population from which the Monte Carlo analyses drew replications, a Monte Carlo analysis was run with the same commands used in the bias evaluation phase for a small slope difference, except for specifications of a single replication and a sample size sufficiently large (*N* = 100,000) that the manufactured data set could be considered the population from which Monte Carlo replications were drawn. Thus, the initial simulation study was an ordinary LGM analysis of single artificial data set--which was requested and stored as a separate output file--that generated statistics that could be interpreted as parameters. Thus, GMA δ =.3464, or .35 after conventional averaging of the effect size to two decimal places. The single-replication Monte Carlo study yielded a *b* for the slope difference of .1012; the residual variance of the intercept was .2539; and the Level-1 variances were .5002, .5010, .4980, and .4959 for *Y*1, *Y*2, *Y*3, and *Y*4 (*M* = .4988), respectively. Thus, for this dataset, GMA *d* = (.1012*3)/(.2539 + .4988)$^{1/2}$ = .3499 = .35.

Next, a between-subjects ANOVA was used to compare the means of the two groups in the Monte-Carlo generated raw data at the final point (*Y*4), and to obtain the *MSE* that is the square of the within-groups *SD* used to compute classical *d*. Classical *d* = (.3043 - −.0019)/.751$^{1/2}$ = .3532 =.35. Note that the *MSE* of .751 (calculated from homogeneous variances of .750 and .752 for the two groups) from the ANOVA of end-of-study generated observations was virtually identical to the .750 parameter calculated by summing the Monte-Carlo specifications for the residual variance of the intercept (.25) and the model (Level-1) residual variance (.50). Most important, after the rounding of effect sizes to two decimal places, GMA δ = GMA *d* = classical *d* =.35 in the manufactured data set.

Monte Carlo Analysis of GMA *d*s for Treatment Effects

To examine biases in the *SE* and CI for the GMA *d*, each of the 10 input statements used to generate the Monte Carlo analyses included a command to output an ASCII file containing the results of 27 model parameters from the 10,000 replications generated by each model. Next, these ASCII files were converted to SPSS data files for analyses with newly created SPSS macros for calculating, for each replication, the (a) GMA *d* with Equation 1, (b) the estimated CI for GMA *d* with Equations 7 and 8 (the CI transformation approach), and (c) *SE* (square root of *v*) of GMA *d* with Equations 9 (see Appendix B for SPSS syntax used for the 5 Monte Carlo analysis of effect sizes for the smaller treatment effect). As in the single-replication Monte Carlos study, the *SD* used to compute GMA *d* in each replication was the square root of the sum of the residual variance of the intercept and the mean of the four observed Level-1 residual variances, which is the model-based estimate of σ for Y4, as *b**duration is the model-based estimate of the population mean difference between groups at that time point.

Coverage for the GMA *d* from each method was calculated with the procedure used in Maas and Hox (2005): a new variable was created in which a score of "1" indicated that a CI for a GMA *d* from a given replication included the effect size specified in the simulations and a score of "0" indicated the estimate was not included in the CI. Then the cumulative frequency distributions of these variables were determined. Coverage was calculated as the percentage of "1"s in the dichotomous distribution, which is the same method Mplus used to calculate coverage for *b*.

Results

Bias in the Unstandardized Coefficient (*b*)

The findings from the 10 Monte Carlo analyses of the unstandardized coefficients (*b*s) obtained from Mplus are reported in Table 1, with the sample size and the *b* specified for each simulation given in the first two columns.

Estimates—Column 3 (Estimates Avg) reports the mean of the 10,000 *b*s for each Monte Carlo analysis. The differences between the means in this column and the respective parameters of .10 or .20 (reported in column 2) indicate the absolute biases in *b* (see column 7). As noted in Muthén and Muthén (2002), relative bias is calculated by dividing the absolute bias by the respective parameter (column 2) and multiplying by 100 to convert the

proportion to a percentage (see column 8). Absolute bias in *b* never exceeded .0006, and relative biases were always less than 1.0%.

Standard errors and CIs—Column 4 in Table 1 reports the Estimates *SD*, which is the standard deviation of the 10,000 *b*s generated by each Monte Carlo simulation, and Column 5 (*SE* Avg) reports the mean of the 10,000 SE_bs. Absolute bias in SE_b was estimated by subtracting the statistics in column 5 from the respective Estimates *SD* in column 4. These differences (reported in column 9) are divided by the Estimates *SD* and multiplied by 100 to determine the relative bias in SE_b (reported in the penultimate column). CI bias was calculated from subtracting the coverage values (column 6) from .95 (and reported in the last column).

The absolute bias in SE_b was .0042 in the smallest sample size ($N = 50$) but was no greater than .0012 at *N*s of 100 or 150. Almost perfect estimates of the coefficients were observed when $N > 250$. Relative bias was largest (3.3%) in the smallest sample size but was well below 10%. The relative biases were between 1.0% and 2.0% at *N*s of 100 and 150, and about 1.0% or less when *N* was at least 250. Thus, both absolute and relative biases in the SE_b decreased with sample size, and were particularly small when *N* was at least 250. Coverage for the CI was always in the .937–.947 range, and was nearly perfect (.945–.947) at *N*s of 250–500, with CI biases of .003 to .013 (see last column in Table 1).

Bias in the Effect Size (GMA *d*)

Estimates—Table 2 reports the parallel Monte Carlo analyses for the GMA *d*s conducted with SPSS using the Monte Carlo results files exported by Mplus. The first column in Table 2 specifies the simulation sample size and column 2 reports the effect size parameter (δ) corresponding to unstandardized coefficients in Table 1. Column 3 (Estimates *SD*) reports the mean of the 10,000 GMA *d*s for each simulation analysis. The absolute biases in *d* (differences between respective values in columns 2 and 3 in Table 2) are reported in column 7, with relative biases given in column 8. The absolute biases were consistently larger for GMA *d*s than for respective *b*s in part because the parameters for the former were much larger than the latter.

Relative biases in GMA *d* and *b* were comparable (and less than 1.0%) when *N* was at least 250. However, although relative bias in *b* was always less than 1.0%, the relative bias in *d* was 2.3–2.6% at $N = 50$ and 1.7–2.0% at $N = 100$–150. Thus, biases in GMA *d* decreased with sample size but never exceeded 5%.

Standard errors and CIs—Column 4 in Table 2 reports the standard deviation of the 10,000 GMA *d*s generated for each simulation analysis, and column 5 reports the means of the 10,000 *SE*s of the GMA *d*s. The absolute biases in *SE*s are the differences between respective values in columns 4 and 5 in Table 2, which are reported in column 9; relative biases are reported in the penultimate column. Whereas relative biases in SE_b were in the .7% to 3.3% range, the corresponding biases in the *SE* of GMA *d* were slightly greater, ranging from 1.2% to 4.7% (with medium effects consistently evincing greater relative bias than small effects for GMA *d*).

The coverage proportions for GMA *d* are reported in column 6 of Table 2 and were virtually identical to respective values for *b*, with differences in the .000–.003 range across effect sizes and sample sizes. CI biases were in the .003 to .014 range (see last column in Table 2), which is nearly the same as the range reported for the biases in the CI for SE_b (Table 1). Thus, the CIs were about as accurate for GMA *d*s as for the respective unstandardized coefficients from which they were derived--and coverage was consistently close to .95, particularly when $N > 250$.

Equivalence of Two CI Estimation Approaches in the Monte Carlo Study

That the variance-calculation method (using Equations 4, 5, and 11) yields a CI for GMA *d* that is mathematically equivalent to the *SE* obtained with the direct CI transformation approach (Equations 7 and 8) was demonstrated previously with summary statistics from a GMA of sex differences in antisocial attitude in the NYS. The equivalence of the two methods was also demonstrated in the Monte Carlo study, with the added benefit of verifying the accuracy of the findings from the simulations. For example, in the Monte Carlo run with 500 subjects, SE_b was .0390 (see column 5 in last row in Table 1). The *SD* of .865, calculated by taking the square root of the mean of the 10,000 sums of intercept residual variance and the model residual variance from each replication, was a very good estimation of σ (.866) calculated from residuals. Thus, using the equations for the variance-calculation approach, GMA $v = (.0390)^2\ (3/.866)^2 = .0183$, and $SE = .0183^{1/2} = .1351$. The mean of the *SE* from the 10,000 replications in the Monte Carlo analysis with model with 500 subjects was .1353 (see column 5 in last row of Table 2), which is the same (after sampling and round-off errors).

Discussion

This article introduced two sets of equivalent equations that estimate the CI of the GMA *d* by either transforming the CI for the unstandardized coefficient (*b*) for the treatment effect to the CI for GMA *d*, or by calculating the variance for the GMA *d* and applying traditional formulas for CI calculation. A key advantage to these all-purpose methods is they can be used in conjunction with outputs from different software programs for GMA. Options specific to individual programs (or different versions within statistical packages) may be configurable to produce the GMA *d* and its CI without the need for post hoc analysis of customary statistics. However, the use of program-specific options would not necessarily be simpler to implement or produce different (i.e., more accurate) CIs than would be obtained with the presented equations.

The Monte Carlo simulation analyses found essentially no bias in the point estimates of the coefficient for the slope difference (*b*). However, large samples (i.e., $N > 150$) were needed for bias in the point estimates of GMA *d* to also be negligible, although the bias was always below the 5% problematical level.

The biases in the estimated CI for *d* were determined by: (a) small-sample bias in the CI for *b*, and (b) bias from use of an estimate rather than a parameter for the standard deviation of *Y* in the CI transformation equation. When σ of *Y* is known, the sampling distribution of GMA *d* is a linear transformation of the sampling distribution of *b* because each *b* in its

sampling distribution is multiplied by a constant (duration/σ) to yield the respective GMA *d*. The relative bias in the CI would then be identical for *b* and *d*. However, when σ must be estimated from the data, each *b* in its distribution is multiplied by a slightly different value because the *SD* in the replication yielding each *b* is not the same across the replications defining the sampling distribution of *b*. As the *N* for each replication increases, the variance of the *SD* across replications decreases, reaching zero when sample size is infinite, and *SD* = σ.

However, irrespective of sample size examined, use of *SD* instead of σ was found to have minimal detrimental consequences for the CI estimation for the GMA *d* because bias was about the same for the CI for *b* and the CI for *d* for the treatment effect. The best explanation for this finding is that standard deviation of *Y* was accurately estimated in the replications. Thus, the bias in CI of GMA *d* were largely the result of bias in the CI unstandardized coefficient when calculated from SE_b when sample size was not very large. The findings from the Monte Carlo analyses that used $N > 250$ of coverage values of .945–.947 for the CI for *b* and .944–.947 for the CI for GMA *d* suggest the CIs based on *SE*s are valid for both standardized and unstandardized coefficients for treatment effects from GMAs conducted with reasonably large samples.

Extensions of the CI Estimation Equations to Complex GMA

The new equations are nominally of use only when comparing two independent groups in growth rate on a continuous outcome, which is the simplest form of GMA of data from randomized clinical trials. Most GMAs of clinical data, however, are more complex than the textbook case. For example, three or more groups, or two or more sites, may be used; 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-characteristic covariates (e.g., gender or risk status) may be included in the model; the effects of unobserved heterogeneity may be examined with latent class or cluster analysis; and outcomes may be categorical (e.g., binary) rather than continuous.

Feingold (2013) introduced an integrative regression framework based on generalized linear mixed models/hierarchical generalized linear models (McCulloch, & Searle, 2001; Raudenbush & Bryk, 2002) for the conceptualization and calculation of GMA *d* for more complex GMAs but did not address CI estimation. However, the *v*s and CIs of the GMA *d*s derived from unstandardized coefficients for treatment effects obtained in the full range of GMA designs subsumed by the regression framework can be estimated with relatively simple extensions of the new equations.

Binary outcomes—Randomized clinical trials often use binary outcome variables. In evaluations of treatments for addiction, for example, abstinence is a commonly used dichotomous dependent variable (e.g., Feingold, Oliveto, Schottenfeld, & Kosten, 2002). In addition, continuous outcome measures may be dichotomized prior to analysis because their distributions are not normally distributed (e.g., when there is a preponderance of zeros), thus violating the assumption of a normality of the dependent variable required for GMA (see

first part in 2-part growth modeling for examples of such dichotomization; Olsen & Schafer, 2001).

With binary outcome data, the structural model for the GMA shifts from a linear regression to a logistic regression framework that models probabilities rather than scores (Hosmer & Lemeshow, 2000). The effect size is then an odds ratio (OR; Agresti, 2002; Feingold, Tiberio, & Capaldi, 2014) conveying the difference between the two groups in proportions (e.g., of clients who achieve or maintain abstinence), and there are guidelines for its practical significance (Chen, Cohen, & Chen, 2010; Rosenthal, 1996).

Effect sizes for group differences in binary outcomes in GMA, although not in the *d* metric, are computed almost exactly as for continuous ones. However, with such categorical data, GMA--like logistic regression--models logits (transformed probabilities) instead of raw scores (Raudenbush & Bryk, 2002). The regression coefficient for the treatment effect (*b*) thus represents the difference between the groups in rate of change per unit of time in logits. The *b* for is multiplied by duration to yield the model-derived estimates for the difference between groups in means of the logits at the end of the study, which can be exponentiated to yield the GMA OR (Feingold, 2013; Hosmer & Lemeshow, 2000). The estimation of CI for the GMA OR entails a slightly different transformation process than that used for GMA *d*. Each of the CLs of *b* is multiplied by duration, and the CI for GMA OR is estimated by exponeniating each of these two products.

More than two groups—A priori comparisons or contrasts (Rosenthal, Rosnow, Rubin, 2000) can be used to obtain separate coefficients in a GMA for the end of study difference in slopes for each planned comparison. In a randomized clinical trial with three groups, for example, one contrast might compare the trajectories of patients receiving two different treatments and a second contrast could compare the trajectory of participants receiving any treatment with that of the control group. Thus, instead of a single coefficient for the group effect, multiple coefficients for treatment effects would be obtained in each model (one per contrast). A separate GMA *d* (along with a contrast-specific CI and *v*) can be estimated from the unstandardized coefficient associated with each contrast.

Cluster-randomized designs—In some randomized clinical trials, groups or dyads of individuals (clinics, therapy groups) are randomly assigned to conditions and every participant 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* (Hedges, 2007, 2009), which may include repeated measures collected over the course of study that can be examined with 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 MLM 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). The calculation of GMA *d*, and its *v* and CI, would use the same equations but the coefficient associated with the treatment effect (and its *SE*) would be obtained at Level 3 rather than Level 2 (for more details, see Feingold, in 2013; see Hedges, 2007, 2009, for calculation of the appropriate *SD* for cluster-randomized designs).

Multi-site studies—Multi-site studies are similar to cluster-randomized studies in that multiple independent data sets are used in both designs but participants are randomized to conditions within clusters in the former. Data from multi-site GMA studies would also be examined with a 3-Level GMA model but the treatment effect would be observed at Level 2. The treatment effect is the mean regression coefficient (i.e., averaged across sites) and its transformed GMA *d* is the average effect size. This design is conceptually similar to meta-analysis, except that meta-analysts typically work with independently conducted studies and convert findings to standardized effect sizes before rather than after synthesizing them because of typical variations in measurement of outcomes across independent studies.

Moderation of treatment effects—The treatment effect may be moderated by observed or latent factors. When interactions of treatment with moderating categorical variables (e.g., gender, race) are observed, GMA *d*s (and their associated *v*s and CIs) that are the equivalent of simple effects in ANOVA can be determined (see Feingold, 2013, for equations).

Limitation and Directions for Future Research

The Monte Carlo study manipulated only factors (sample size and effect size) that were predicted to affect bias in the CI for the GMA *d*. Future simulation studies might examine generalizability of the accuracy of the new equations by considering other potential sources of bias, including nonnormality, violation of missing data assumptions (Little & Rubin, 2002), unbalanced designs, heterogeneity of within-person variances across time, heterogeneity of variance across treatment conditions (Feingold, 1995; Grissom, 2000), and robustness to model misspecification--all of which may interact with effect size and sample size in influencing accuracy of CI estimation for the GMA *d*.

Approaches for the CI estimation for effect sizes for nonlinear models also need to be developed and validated. For example, Feingold (2013) suggested the calculations of effect sizes at different time points based on GMA models with linear and quadratic slopes to communicate the practical important of treatment effects at different phases of a study. Yet, CI estimation for such GMA *d*s would require the *SE*s of the coefficients for both linear and quadratic terms.

Finally, although the equivalence of the two methods of CI estimation was demonstrated, a formal mathematical proof of the equivalence of the two sets of equations would be of interest, particularly to methodologists.

Summary and Conclusions

Monte Carlo studies demonstrated that estimates of an increasingly used effect size (GMA *d*) for the difference between two independent groups at the end of a study examined with GMA (LGM or MLM/HLM) exhibited ignorable bias that did not differ practically from the bias observed in the commonly used unstandardized coefficient (*b*)--from which the GMA *d* is a simple and approximately linear transformation. Bias in the CI for *d* was essentially nonexistent in the relatively large samples typically used in GMA and thus the new equations provide interval estimates for the GMA *d* that can be reported by investigators

conducting GMA to address increasing demands by reviewers, editors, and publishers for CIs for effect sizes.

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. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health

References

Aderka IM, Gillihan SJ, McLean CP, Foa EB. The relationship between posttraumatic and depressive symptoms during prolonged exposure with and without cognitive restructuring for the treatment of posttraumatic stress disorder. Journal of Consulting and Clinical Psychology. 2013; 81:375–382. [PubMed: 23339538]

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

Algina J, Kesselman HJ. Approximate confidence intervals for effect sizes. Educational and Psychological Measurement. 2003; 68:233–244.

American Psychological Association. Publication manual of the American Psychological Association. 6th ed.. Washington, DC: Author; 2009.

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]

Arch JJ, Eifert GH, Davies C, Vilardaga JCP, Rose RD, Craske MG. Randomized clinical trial of cognitive behavioral therapy (CBT) versus acceptance and commitment therapy (ACT) for mixed anxiety disorders. Journal of Consulting and Clinical Psychology. 2012; 80:750–765. [PubMed: 22563639]

Baguley T. Standardized or simple effect size: What should be reported? British Journal of Psychology. 2009; 100:603–617. [PubMed: 19017432]

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

Blanton H, Jaccard J. Arbitrary metrics in psychology. American Psychologist. 2006; 61:27–41. [PubMed: 16435974]

Bollen, KA.; Curran, PJ. Latent curve models: A structural equation perspective. Hoboken, NJ: Wiley; 2006.

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

Chaffin M, Funderburk B, Bard D, Valle LA, Gurwitch R. A combined motivation and parent-child interaction therapy package reduces child welfare recidivism in a randomized dismantling field trial. Journal of Consulting and Clinical Psychology. 2011; 79:84–95. [PubMed: 21171738]

Chen H, Cohen P, Chen S. How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. Communications in Statistics—Simulation and Computation. 2010; 39:860–864.

Cheung MW. Comparison of methods for constructing confidence intervals of standardized indirect effects. Behavior Research Methods. 2009; 41:425–438. [PubMed: 19363183]

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

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

Cooper, H.; Hedges, LV.; Valentine, JC., editors. The handbook of research synthesis. 2nd ed.. New York: Russell Sage; 2009.

Curran PJ. Have multilevel models been structural equation models all along? Multivariate Behavioral Research. 2003; 38:529–569.

Cumming, G. Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis. New York: Routledge; 2013.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series 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.

Efron, B.; Tibshirani, R. An introduction to the bootstrap. Boca Raton, FL: Chapman & Hall; 1993.

Elliott, DS.; Huizinga, D.; Menard, S. Multiple problem youth: Delinquency, substance use, and mental health problems. New York: Springer-Verlag; 1989.

Feingold A. The additive effects of differences in central tendency and variability are important in comparisons between groups. American Psychologist. 1995; 50:5–13.

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. A regression framework for effect size assessments in longitudinal modeling of group differences. Review of General Psychology. 2013; 17:111–121. [PubMed: 23956615]

Feingold A, Oliveto A, Schottenfeld R, Kosten TR. Utility of crossover designs in clinical trials: Efficacy of desipramine vs. placebo in opioid-dependent cocaine abusers. The American Journal on Addictions. 2002; 11:111–123. [PubMed: 12028741]

Feingold A, Tiberio S, Capaldi DM. New approaches for examining associations with latent categorical variables: Applications to substance abuse and aggression. Psychology of Addictive Behaviors. 2014; 28:257–267. [PubMed: 23772759]

Glass GV. Primary, secondary, and meta-analysis of research. Educational Researcher. 1976; 5(10):3–8.

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

Goldstein, H. Multilevel statistical models. 4th ed.. Hobokin, NJ: Wiley; 2011.

Grissom RJ. Homogeneity of variance in clinical data. Journal of Consulting and Clinical Psychology. 2000; 68:155–165. [PubMed: 10710850]

Grissom, RJ.; Kim, JJ. Effect sizes for research: Univariate and multivariate applications. 2nd ed.. New York: Routledge; 2012.

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.

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.

Hedges, LV.; Olkin, L. Statistical methods for meta-analysis. Orlando, FL: Academic Press; 1985.

Hedges LV, Pustejovskya JE, Shadish WR. A standardized mean difference effect size for single case designs. Research Synthesis Methods. 2012; 3:224–239.

Hodges, JL.; Lehmann, EL. Basic concepts of probability and statistics. 2nd ed.. Philadelphia, PA: Society for Industrial and Applied Mathematics; 2005.

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

Kelley K, Preacher KJ. On effect size. Psychological Methods. 2012; 17:137–152. [PubMed: 22545595]

Kelley K, Rausch JR. Sample size planning for longitudinal models: Accuracy in parameter estimation for polynomial change parameters. Psychological Methods. 2011; 16:391–405. [PubMed: 21744968]

Kerr DCR, DeGarmo DS, Leve LD, Chamberlain P. Juvenile justice girls' depressive symptoms and suicidal ideation 9 years after multidimensional treatment foster care. Journal of Consulting and Clinical Psychology. 2014 Advance online publication.

Lau RS, Cheung GW. Estimating and comparing specific mediation models in complex latent variable models. Organizational Research Methods. 2012; 15:3–16.

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

Ljótsson B, Hesser H, Andersson E, Lindfors P, Hursti T, Rück C, Hedman E. Mechanisms of change in an exposure-based treatment for irritable bowel syndrome. Journal of Consulting and Clinical Psychology. 2013 Advance online publication.

Maas CJM, Hox JJ. Sufficient sample sizes for multilevel modeling. Methodology. 2005; 1:86–92.

MacKinnon DP, Lockwood CM, Williams J. Confidence limits for the indirect effect: Distribution of the product and resampling methods. Multivariate Behavioral Behavioral Research. 2004; 39:99–128.

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

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 BO, Muthén LK. The development of heavy drinking and alcohol-related problems from ages 18 to 37 in a U.S. national sample. Journal of Studies on Alcohol. 2000; 61:290–300. [PubMed: 10757140]

Muthén LK, 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. 7th ed. Los Angeles, CA: Muthén and Muthén; 2012.

Nunnally, J. Psychometric methods. New York: McGraw-Hill; 1978.

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.

Olejnik S, Algina J. Measures of effect size for comparative studies: Applications, interpretations, and limitations. Contemporary Educational Psychology. 2000; 25:241–286. [PubMed: 10873373]

Olsen MK, Schafer JL. A two-part random-effects model for semicontinuous longitudinal data. Journal of the American Statistical Association. 2001; 96:730–745.

Preacher, KJ.; Wichman, AL.; MacCallum, RC.; Briggs, NE. Latent growth modeling. Los Angeles, CA: Sage; 2008.

Raudenbush, SW. Hierarchical linear models to study the effects of social context on development. In: Gottman, JM., editor. The analysis of change. Mahwah, NJ: Erlbaum; 1995. p. 165-201.(1995).

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

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 JA. Qualitative descriptors of strength of association and effect size. Journal of Social Service Research. 1996; 21:37–59.

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

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

Safren SA, O'Cleirigh CM, Bullis JR, Otto MW, Stein MD, Pollack MH. Cognitive behavioral therapy for adherence and depression (CBT-AD) in HIV-infected injection drug users: A randomized controlled trial. Journal of Consulting and Clinical Psychology. 2012; 80:404–415. [PubMed: 22545737]

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.

Steiger H. Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. Psychological Methods. 2004; 9:164–182. [PubMed: 15137887]

Twohig MP, Hayes SC, Plumb JC, Pruitt LD, Collins AB, Hazlett-Stevens H, Woidneck MR. A randomized clinical trial of acceptance and commitment therapy versus progressive relaxation training for obsessive-compulsive disorder. Journal of Consulting and Clinical Psychology. 2010; 78:705–716. [PubMed: 20873905]

Appendix A

Mplus Input Statement for a Monte Carlo Study with .10 for the Slope Difference and *n* = 250

```
MONTECARLO: NAMES ARE y1–y4 x;
CUTPOINTS = x (0);
NOBSERVATIONS = 250;
NREPS = 10000;
SEED = 53487;
CLASSES = C(1);
GENCLASSES = C(1);
SAVE = CIM250.dat;
RESULTS = CIM250.sav;
ANALYSIS: TYPE = MIXTURE;
ESTIMATOR = ML;
MODEL MONTECARLO:
%OVERALL%
[x@0]; x@1;
i BY y1–y4@1;
s BY y1@–3 y2@-2 y3@-1 y4@0;
[y1–y4@0];
[i*0 s*.2];
i*.25;
s*.09;
i WITH s*0;
y1–y4*.5;
i ON x*.3;
s ON x*.1;
%C#1%
[i*0 s*.2];
MODEL:
%OVERALL%
i BY y1–y4@1;
s BY y1@–3 y2@-2 y3@-1 y4@0;
[y1–y4@0];
[i*0 s*.2];
i*.25;
s*.09;
```

```
i WITH s*0;
y1–y4*.5;
i ON x*.3;
s ON x*.1;
%C#1%
[i*0 s*.2];
OUTPUT: TECH9;
```

Appendix B

SPSS Syntax for Transforming Unstandardized Coefficients from Mplus Monte Carlo Output Files to GMA *ds* and Monte Carlo Analysis of GMA *ds*

```
compute b=v9.
compute seb=V20.
compute meanres=(v2+v3+v4+v5)/4.
compute sd=sqrt(meanres+v10).
compute gmad=b*(3/sd).
compute llci=(b − 1.96*seb)*(3/sd).
compute ulci=(b + 1.96*seb)*(3/sd).
compute cov=0.
Execute.
if(llci gt.3464 or ulci lt .3464)cov=1.
execute.
FREQUENCIES
VARIABLES=cov
/ORDER=ANALYSIS.
compute sed=(ulci-gmad)/1.96).
execute.
DESCRIPTIVES
VARIABLES=gmad sed
/STATISTICS=MEAN STDDEV.
```

J Consult Clin Psychol. Author manuscript; available in PMC 2016 February 01.

Table 1

Monte Carlo Analyses of Unstandardized Coefficients for the Group Difference in Slopes for a Linear Latent Growth Model as a Function of Treatment Potency and Sample Size

	Estimates			SE_b		Bias in Estimates, SE_b, and CI				
N	**Pop *b* (β)**	**Avg**	***SD***	**Avg**	**Coverage**	**Est**	**%**	SE_b	**%**	**CI**
50	.1000	.0996	.1258	.1216	.937	−.0004	.40	−.0042	3.34	.013
50	.2000	.1996	.1258	.1216	.937	−.0004	.20	−.0042	3.34	.013
100	.1000	.1006	.0878	.0866	.943	.0006	.60	−.0012	1.37	.007
100	.2000	.2006	.0878	.0866	.943	.0006	.30	−.0012	1.37	.007
150	.1000	.1002	.0720	.0708	.942	.0002	.20	−.0012	1.67	.008
150	.2000	.2002	.0720	.0708	.942	.0002	.10	−.0012	1.67	.008
250	.1000	.0996	.0554	.0550	.947	−.0004	.40	−.0004	.72	.003
250	.2000	.1996	.0554	.0550	.947	−.0004	.20	−.0004	.72	.003
500	.1000	.0998	.0394	.0390	.945	−.0002	.20	−.0004	1.02	.005
500	.2000	.1998	.0394	.0390	.945	−.0002	.10	−.0004	1.02	.005

Note. Pop *b* = population value for unstandardized coefficient for effect of group on slope (β), Estimates Avg = average estimate (*b*) across replications, Estimates *SD* = standard deviation of estimates across replications, SE_b Avg = average SE_b across replications, Coverage = 95% coverage for *b*.

Table 2

Monte Carlo Analyses of the Effect Sizes (GMA *d*s) for the Group Difference in Slopes as a Function of Δ and Sample Size

	Estimates			SE_d		Bias in Estimates, SE_d, and CI				
N	**δ**	**Avg**	***SD***	**Avg**	**Coverage**	**Est**	**%**	SE_d	**%**	**CI**
50	.3464	.3545	.4506	.4325	.936	.0081	2.34	−.0181	4.02	.014
50	.6928	.7108	.4539	.4325	.935	.0180	2.60	−.0214	4.71	.015
100	.3464	.3533	.3095	.3040	.942	.0069	1.99	−.0055	1.78	.008
100	.6928	.7048	.3117	.3040	.943	.0120	1.73	−.0077	2.47	.007
150	.3464	.3504	.2532	.2477	.940	.0040	1.15	−.0055	2.17	.010
150	.6928	.7003	.2551	.2477	.939	.0075	1.08	−.0074	2.90	.011
250	.3464	.3468	.1938	.1915	.947	.0004	.12	−.0023	1.19	.003
250	.6928	.6951	.1952	.1915	.945	.0023	.33	−.0037	1.90	.005
500	.3464	.3466	.1373	.1353	.944	.0002	.06	−.0020	1.46	.006
500	. 6928	.6939	.1383	.1353	.944	.0011	.16	−.0030	2.17	.006

Note. Estimates Avg = average estimate (*d*) across replications, Estimates *SD* = standard deviation of estimates across replications, SE_d Avg = average SE_d across replications, Coverage = 95% coverage for *d*.