Published in final edited form as:
Psychol Methods. 2009 March ; 14(1): 43–53. doi:10.1037/a0014699.

Effect Sizes for Growth-Modeling Analysis for Controlled Clinical Trials in the Same Metric as for Classical Analysis

Alan Feingold
Oregon Social Learning Center

Abstract

The use of growth-modeling analysis (GMA)--including Hierarchical Linear Models, Latent Growth Models, and General Estimating Equations--to evaluate interventions in psychology, psychiatry, and prevention science has grown rapidly over the last decade. However, an effect size associated with the difference between the trajectories of the intervention and control groups that captures the treatment effect is rarely reported. This article first reviews two classes of formulas for effect sizes associated with classical repeated-measures designs that use the standard deviation of either change scores or raw scores for the denominator. It then broadens the scope to subsume GMA, and demonstrates that the independent groups, within-subjects, pretest-posttest control-group, and GMA designs all estimate the same effect size when the standard deviation of raw scores is uniformly used. Finally, it is shown that the correct effect size for treatment efficacy in GMA--the difference between the estimated means of the two groups at end of study (determined from the coefficient for the slope difference and length of study) divided by the baseline standard deviation--is not reported in clinical trials.

Keywords

effect size; growth modeling; hierarchical linear models; clinical trials

Social scientists contribute to human well-being through the development of interventions to prevent and to treat a wide range of psychological, educational, and behavioral problems. Intervention research is a broad interdisciplinary field, with contributors from clinical psychology, psychiatry, applied developmental psychology, prevention science, and education. An important objective of these investigators is to conduct clinical trials to examine the efficacy of psychosocial and psychopharmacological treatments. Thus, controlled clinical trials are needed to answer two questions: Is a particular intervention effective? How powerful are its effects? The former is addressed in data analysis through tests of statistical significance, and the latter through calculation of effect sizes (McGrath & Meyer, 2006).

Growth-Modeling Designs for Controlled Clinical Trials

Traditionally, data from controlled clinical trials have been examined with classical statistical techniques, such as analysis of variance (ANOVA), which use Ordinary Least Squares (OLS) and the General Linear Model (GLM). Over the last decade, however, *growth-modeling analysis* (GMA)--based on Generalized Least Squares (GLS) and the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977)--has emerged as a competing

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

statistical framework for use in the evaluation of intervention efficacy. GMA compares temporal trajectories (growth curves) between the treatment and comparison groups, with the difference in the slopes for linear trend a common test of efficacy (Atkins, 2005). A difference in trajectories indicates a difference between groups at the end of the study (Raudenbush & Liu, 2001).

The GMA family includes: (1) Hierarchical Linear Models (HLM; Raudenbush & Bryk, 2002), also known as Mixed-effect Regression Models (MRM; Gibbons, Hedeker, Elkin, Waternaux, Kraemer, Greenhouse, et al., 1993; Hedeker, & Gibbons, 2006), or multilevel models (Hox, 2002), (2) Latent Growth Models (LGM; Muthen & Muthen, 2007; Singer & Willett, 2003), also known as Latent Curve Analysis (LCA; Curran & Muthen, 1999; Muthen & Curran, 1997), and (3) General Estimating Equations (GEE; Liang & Zeger, 1986). Prior to the development of GMA, growth curves were modeled through OLS techniques. A classical equivalent of a GMA is the mixed ANOVA with orthogonal polynomial contrasts that extends the 2 x 2 design to fit trends for three or more repeated measures varying over time (Winer, 1971). Such trend analysis can examine whether the linear and quadratic sources of variation for the repeatedly-measured outcomes (the within-subjects factor) are moderated by the treatment (between-subjects) factor. Alternatively, a multivariate analysis of variance (MANOVA) can be used that handles the repeated measures as multiple dependent variables to model temporal trends to test the same effects--but with somewhat less restrictive statistical assumptions--examined in an ANOVA (Hedeker & Gibbons, 2006; Gueorguieva & Krystal, 2004).

However, ANOVA and MANVOA both require complete data with time-structured measures (temporal spacing between occasions must be identical and available for all subjects), and cannot easily (1) handle time-varying covariates, (2) model between-subjects variations in growth trajectories, (3) assess the significance of unconditional and residual variations in growth parameters (i.e., random effects) across subjects, or (4) deal with dependence of observations when, for example, repeated measures of outcomes are nested within patients who are nested within clinics. In clinical trials, missing data is an expected consequence of inevitable attrition (including loss of data as a result of inability to locate participants in intent-to-treat approaches), and the intervals between assessment times often differ by subjects as a result of scheduling conflicts. GMA uses data from all participants rather than from "completers," and is seen as providing less biased treatment effects.

Proliferation of Studies Using GMA

GMA is beginning to supplant classical methods as the primary statistical tool to analyze data collected from controlled clinical trials in psychiatry, prevention science, and clinical psychology.

Psychiatry (psychopharmacological treatment)

Gueorguieva and Krystal (2004) examined articles published in *Archives of General Psychiatry* over a 12-year period. Whereas no study had used GMA in 1989, GMA designs were used in 30% of clinical trials published in 2001. Examples of applications of GMA in psychiatry can be found in research on the effects of tricyclic antidepressant medication on depressive symptoms (Bock, 1983; see also re-analyses by Hedeker & Gibbons, 2006) and of desipramine on drug use (Feingold, Oliveto, Schottenfeld, & Kosten, 2002).

Prevention science (psychosocial intervention)

The prevention research literature also indicates an increase in the popularity of GMA. A manual search of articles published in *Prevention Science* from its launch in 2000 through 2006

found 7 GMA-examined controlled trials in 2005−2006, the same number published in the Journal's first 5 years. Half of these 14 prevention studies used an LGM framework, with HLM and GEE approaches each used in about a quarter of the studies (4 HLM studies and 3 GEE studies). Examples of applications of GMA in the prevention field can be found in studies of the effects of a parenting intervention on child behaviors (DeGarmo, Patterson, & Forgatch, 2004) and of effects of a school-based intervention on adolescent substance use initiation (Trudeau, Spoth, Lillehoj, Redmond, & Wickrama, 2003).

Clinical psychology (psychosocial treatment)

A manual search was conducted to identify controlled clinical trial studies that used GMA and were published from 1997−2006 inclusive in the APA's Journal of Consulting and Clinical Psychology (JCCP), long considered the most prestigious outlet for intervention research (Feingold, 1989; Simpson, McGrath, & Yamada, 2003). In 1997−1998, only three studies used GMA, but the numbers increased progressively to 13, 16, 16, and 27 for the years 1999−2000, 2001−2002, 2003−2004, and 2005−2006, respectively. A comprehensive review of the 43 *JCCP* trials published from 2003 through 2006 inclusive found that the vast majority (35 or 81%) of the articles reported use of HLM, with LGM used in 5 studies and GEE in 3 studies. Thus, whereas prevention researchers who used GMA overwhelmingly favored the LGM framework, the psychosocial treatment evaluators strongly preferred HLM. Examples of the use of GMA in clinical psychology can be found in studies of the effects of couples therapy on marital distress (Atkins, Berns, George, Doss, Gattis, & Christensen, 2005) and of cognitive therapy on symptoms of posttraumatic stress disorder (Monson, Schnurr, Resick, Friedman, Young-Xu, & Stevens, 2006).

Importance of Effect Sizes in Evaluation of Interventions

Effect sizes associated with hypothesis (significance) tests of efficacy are important for several reasons. First, a summary of the effect sizes found for a particular intervention in independent trials is important to establish the generality of the effect. Second, treatments examined in different studies must be compared, which is often accomplished by modeling variations in effect sizes across trials through meta-analysis (Hedges & Olkin, 1985; Lipsey & Wilson, 1993, 2001). Finally, effect sizes are needed for the cost-benefit analyses that are often conducted by policy makers (Petitti, 2000). A treatment that is effective but has a small impact on problem behavior will not be useful if it is expensive to implement. Because of the importance of effect sizes, software has been created to calculate them (e.g., Shadish, Robinson, & Lu, 1999), but only for findings obtained with classical designs.

Nonetheless, only 13 of the 43 clinical trial GMA studies published in *JCCP* in 2003−2006 reported effect sizes calculated using model-based coefficients (e.g., for the slope differences) generated by GMA programs to convey the size of treatment effects. This was surprising because APA mandates the inclusion of effect sizes in their journal articles (American Psychological Association, 2001). More troublesome, the 11 retrieved studies that reported both effect sizes and the methods used to obtain them (i.e., two of the 13 studies reported effect sizes without specifying the formula used) applied formulas that were not mathematically equivalent. Because effect sizes are the building blocks in a meta-analysis, and meta-analysis is commonly used examine treatment efficacy (e.g., Weisz, Jensen-Doss, & Hawley, 2006), the concern that effect sizes associated with GMA significance tests are not calculated consistently merits close scrutiny.

Effect-Size Calculation Controversy for Repeated-Measures Designs

The correct approach to the calculation of the effect size for a within-subject (repeated-measures) or mixed design is controversial because there are two possible denominators that

can be used in the formula: (1) the standard deviation of the pretest-posttest *change scores* that reflect within-group variations in improvement over the course of the trial (Gibbons, Hedeker, & Davis, 1993; Mullen & Rosenthal, 1985; Rosenthal, 1991) or (2) the standard deviation of the *raw scores* (often based on the pretest or baseline data) that estimate variations in the outcome measure in the population (Becker, 1988). Advocates of the use of the standard deviation of raw scores contend that effect sizes based on change scores are "biased" estimates of the true effect sizes, which are estimated correctly when raw scores are used (Dunlap, Cortina, Vaslow, & Burke, 1996; Hunter & Schmidt, 2004; Lipsey & Wilson, 2001). Morris and DeShon (2002) offered a compromise view that the appropriate denominator depends on the research domain. For example, an effect size calculated with the standard deviation of change scores might be justifiable for studies on, say, learning but not for studies (including clinical trials) that compare two or more independent groups.

The controversy is exclusive to repeated-measures and mixed designs because such designs include a source of variance not found in between-subjects designs--namely, the correlation between the paired scores (Dunlap et al., 1996; Morris & DeShon, 2002). When the standard deviation of change scores is used for the denominator, treatment potency and the strength of the pretest-posttest correlation are confounded. However, this correlation must contribute to effect sizes used for power analysis because it is one of the determinants of power (Cohen, 1988). Thus, whereas effect sizes calculated from standard deviations of change scores are appropriate for use in power analysis, standard deviations of raw scores should be used to determine intervention potency. These concerns are equally applicable to effect sizes for GMA, which also analyze repeatedly-assessed outcomes.

Overview of Current Study

First, state-of-the-art procedures for calculation of effect sizes for classical experimental designs are demonstrated with the use of an artificial dataset. Next, it is shown that these formulas can be adapted to calculate effect sizes from GMA outputs (regardless of whether the GMA is conducted with HLM, LGM, or GEE) that estimate the same parameters as do traditional formulas when applied to data collected from classical designs. Finally, the effect-size formulas used in contemporary controlled clinical trials that used GMA are applied to the mock data to identify sources of effect size discrepancies, and the implications for both power analysis and meta-analysis are discussed.

Clinical Trials Designs and Mock Data Analysis

Mock Longitudinal Dataset

Table 1 reports hypothetical data from a trial in which 20 subjects were randomly assigned to either an intervention (with subjects coded .50 for Tx, or treatment) or a control condition (with subjects coded −.50 for Tx), with an equal number (n = 10) in each group. At the baseline assessment, T1, the scores for the dependent variable were the same in the treatment and control groups, with each distribution perfectly symmetrical. Thus, means and standard deviations were identical for the two groups at T1. Three additional waves of data (T2-T4), with equal spacing between intervals, were generated for these subjects. Control subjects' raw scores increased by one point between assessment times, whereas the scores for subjects administered the intervention increased by two units per wave over the same period. Thus, in both groups, scores increased as a purely linear function of time, but the rate of improvement was steeper for the treatment group, indicating that the intervention was efficacious. In addition, the within-group standard deviations were homogeneous across the eight cells formed by crossing condition with time. These data are used to demonstrate the comparability of effect sizes for treatment potency (and *inconsistency* of corresponding effect sizes needed for power analysis) among different designs.

Research Designs for Evaluating Interventions

Independent groups (completely randomized) design—Subjects are randomly assigned to treatment and control groups; the intervention is administered only to the treatment group; and the means of the two groups are compared at the end of the study (Morris & DeShon, 2002). The mock data in Table 1 can be examined by a between-subjects ANOVA (or a *t* test for independent groups) if the data from the first three waves (T1-T3) are excluded from the analysis and the ANOVA compares the means of the two groups at T4, $F(1, 18) = 33.75$, $p < .001$.

Cohen (1988) developed the effect size *d* to convey the magnitude of the difference in central tendency between two groups for this design, which is calculated by dividing the mean difference between groups by the pooled within-group standard deviation (*SD*, the square root of the MS_e from the ANOVA).

$$d = (M_T - M_C)/SD \quad (1)$$

$= (11 - 8)/1.15 = 2.60$,[1] where M_T is the end-of-study mean of the treatment group and M_C is the corresponding mean of the control group.

Within-subjects (one-group pretest-posttest) design—A single group of subjects is administered a pretest, the intervention, and a posttest. The pretest and posttest means are then compared via the *F* test from a repeated-measures ANOVA (or *t* test for dependent groups, where *t* is the square root of the *F* from the ANOVA).

The mock data in Table 1 can be examined in this design by using only the T1 and T4 data from the 10 subjects in the treatment group. For this subsample, the pretest mean is 5 and the posttest mean is 11, indicating a 6-unit increase in the outcome variable for treated subjects, $F(1, 9) = 540.00$, $p < .001$.

The *F* value from the within-subjects analysis is much larger than that previously reported for the independent groups analysis because (1) there is a high correlation ($r = .75$) between pretest and posttest scores (and the variance associated with this covariation can be removed from the error term to yield a more powerful statistical test), and (2) there is a maturation effect bias (Shadish, Cook, & Campbell, 2002). In the unused data from the control group, the means increased from 5 at pretest to 8 at posttest. Thus, an estimated half of the improvement in the treatment group from baseline to posttest was due to maturation and the rest to the intervention.

The two effect sizes for the treatment effect obtained from the within-subjects analysis of the mock data in Table 1 (with corrections for maturation bias) are:

$$d_{WS-CHANGE} = (M_{POST} - M_{PRE} - \text{Correction})/SD_{CHANGE} = (11 - 5 - 3)/.82 = 3.67, \text{and} \quad (2)$$

$$d_{WS-RAW} = (M_{POST} - M_{PRE} - \text{Correction})/SD_{RAW} = (11 - 5 - 3)/1.15 = 2.60. \quad (3)$$

[1]Strictly speaking, with such a small sample size, a correction should be applied to the effect size (Hedges & Olkin, 1985). However, for pedagogic reasons, the correction will be ignored.

The $d_{WS\text{-}CHANGE}$ is larger than $d_{WS\text{-}RAW}$, which indicates that the within-subjects design is more powerful than the between-subjects design, but $d_{WS\text{-}RAW}$ conveys treatment potency--and is identical to the *d* reported previously for the randomized design.

Independent-groups pretest-posttest design (IGPP)—Subjects are administered a pretest and then randomly assigned to intervention or control groups; the intervention is administered only to subjects in the treatment condition; both groups are assessed again at the end of the study. The analysis determines whether the magnitude of the change in the intervention group differs significantly from that in the control group.

The mock data in Table 1 can be examined by IGPP by using only the T1 and T4 waves from each group. The last column reports the T4 – T1 change (gain) scores for all subjects. The *F* ratio compares the means of the change scores between the treatment and control groups at T4. In other words, the same analytic methods are applied in IGPP as in the independent groups design except that the change scores are used as the dependent variable instead of the T4 scores, $F(1, 18) = 67.50$, $p < .001$. [2]

As in the within-subjects design, two effect sizes that can be calculated from an IGPP design, one using the standard deviation of the change scores in the denominator and the other using the standard deviation of the raw scores (Morris & DeShon, 2002; Morris, 2008):

$$d_{IGPP-CHANGE} = (M_{CHANGE-T}/SD_{CHANGE-T}) - (M_{CHANGE-C}/SD_{CHANGE-C}) \quad (4)$$

= (6/.82) − (3/.82) = 3.67, where $SD_{CHANGE\text{-}T}$ is the mean change score for the treatment group and $SD_{CHANGE\text{-}C}$ is the mean change score for the control group. (If homogeneity of variance across conditions is assumed, each of these terms can be replaced by SD_{CHANGE}.)

$$d_{IGPP-RAW} = \left(M_{CHANGE-T}/SD_{RAW(PRE-T)}\right) - \left(M_{CHANGE-C}/SD_{RAW(PRE-C)}\right) \quad (5)$$

= (6/1.15) − (3/1.15) = 2.60, where $M_{CHANGE\text{-}T}$ is the mean of the change scores for the treatment group, $M_{CHANGE\text{-}C}$ is the mean of change scores for the control group, $SD_{RAW(PRE\text{-}T)}$ is the pretest *SD* for the treatment group, and $SD_{RAW(PRE\text{-}C)}$ is the pretest *SD* for the control group. (If homogeneity of variance is assumed, each of the last two terms can be replaced by SD_{RAW}.) Note that the values for both effect sizes are identical to the corresponding maturation-corrected within-subjects effect sizes (Eqs. 2 and 3), and the effect size calculated with the raw score standard deviation is also the same as in the independent groups design (Eq. 1).

GMA design—The IGPP can be conceptualized as a simple GMA design that uses an OLS framework. First, growth is estimated for each subject by calculating the change (pretest-posttest difference) score. Second, the mean of the growth scores in the treatment group is contrasted with that of the control group. In HLM, by comparison, three or more assessment times are required to yield a growth score for each subject using empirical Bayes estimates of change that consider both the longitudinal data for each subject and the data from other subjects in the analysis, and can handle missing data when every subject is not observed at all scheduled times (Raudenbush & Bryk, 2002). The empirical Bayes estimate of the person-specific slope indexes the rate of growth for each subject, which is analogous to the pretest-posttest difference

[2]These data could also be analyzed equivalently by a mixed, or split-plot, ANOVA, in which the T1 and T4 scores constitute the two levels of a repeated-measures factor and the two groups form the two levels of the between-subjects factor. The treatment effect is tested by the *F* for the interaction of the two factors, which is identical to the *F* from the one-way ANOVA performed on the change scores.

score for each subject in the IGPP. As with IGPP, at stage two, the means of the growth scores are compared between groups, although GMA uses GLS rather than OLS in analysis (Raudenbush & Bryk, 2002).

Unlike in the previous designs, a GMA can use all four waves of mock data reported in Table 1, and an HLM analysis was conducted in which the time factor was mean centered by using linear weights (−3, −1, 1, and 3 for T1 through T4, respectively) for a four-level design obtained from a table of orthogonal polynomials (Snedecor & Cochran, 1967) for the within-subjects (Level 1 in HLM terminology) facet of the analysis. As shown in Table 1, Tx was a dummy-coded variable (control = −.50, treatment = .50), and its effect was obtained in the between-subjects (Level 2) facet of the analysis. The treatment effect is manifested in the unstandardized coefficient reflecting the degree to which treatment group assignment moderates the slope effect associated with improvement over time (Raudenbush & Liu, 2001). Because the Level 2 treatment predictor is a dummy variable with codes differing by one unit, the unstandardized coefficient (β_{11}) is the difference between the means of the slopes of the treatment and the control group, $\beta_{11} = .50$, $t(18) = 8.80$, $p < .001$.

Raudenbush and Liu (2001) presented two formulas for effect sizes for GMA analyses associated with treatment effects, both of which correspond directly to the effect sizes described previously for IGPP (i.e., $d_{\text{IGPP-CHANGE}}$ and $d_{\text{IGPP-RAW}}$, Eqs. 4 and 5):

$$d_{\text{GMA-CHANGE}} = \beta_{11}/(\tau)^{1/2} = .50/(.00029)^{1/2} = 29.36, \text{and} \quad (6)$$

$$d_{\text{GMA-RAW}} = \beta_{11}(\text{time})/SD_{\text{RAW}} = .50(6)/1.15 = 2.60. \quad (7)$$

Note that β_{11} is the difference between the groups in mean growth rates, and is thus analogous to $M_{\text{CHANGE-T}} - M_{\text{CHANGE-C}}$ that is used as the numerator in the formulas for computing the effect sizes in IGPP (Eqs. 4 and 5). The square root of τ (the denominator in $d_{\text{GMA-CHANGE}}$) is the estimate of the within-group variability of the "true" scores of the slopes, which is analogous to SD_{CHANGE} in the formula for $d_{\text{IGPP-CHANGE}}$ (Eq. 4). The value of β_{11} (average growth rate) must be multiplied by time to obtain the difference between the model-estimated means of the two groups at the end of the study (adjusted for baseline differences between groups). (Because the linear weights for time used in the analysis differed by two units between waves, the value for time in the numerator of $d_{\text{GMA-RAW}}$ is 6.) The effect size of 2.60 calculated with the standard deviation of raw scores equals the effect size found when classical designs were used to analyze the mock data. Most important, and consistent with the current formulation, Raudenbush and Liu (2001) used the formula for $d_{\text{GMA-CHANGE}}$ for power calculations and $d_{\text{GMA-RAW}}$ to convey effect magnitude.

Effect-Size Formulas Reported in the Clinical Treatment Literature

Table 2 lists the 11 clinical trial articles published in *JCCP* in 2003−2006 inclusive that calculated effect sizes from GMA outputs and also reported the formula used for the calculations. The formulas are also tabled, and the last column reports the effect size for the treatment effect in the mock dataset when calculated from each formula.

Table 2 shows that 6 of the 11 studies (Antoni, Lechner, Kazi, Wimberly, Sifre, et al., 2006;Atkins, Eldridge, Baucom, & Christensen, 2005;Brown, Catalano, Fleming, Haggerty, & Abbott, 2005;Christensen, Atkins, Berns, Wheeler, Baucom, & Simpson, 2004;Dimidjian,

Hollon, Dobson, Schmaling, Kohlenberg, et al., 2006;Tolan, Gorman-Smith, & Henry, 2004) calculated a GMA effect size defined herein as $d_{GMA\text{-}CHANGE}$ (Eq. 6). As previously noted, this statistic is appropriate for use in calculations of power analysis for planned HLM studies but not for conveying effect potency because it uses the standard deviation of change scores rather than raw scores. Thus, the value for *d* in the table for these 6 studies is the same effect size reported for the $d_{GMA\text{-}CHANGE}$ in the analysis of the mock data.

Two of the remaining 5 studies (Atkins, Berns, George, Doss, Gattis, & Christensen, 2005; Schultz, Cowan, & Cowan, 2006) reported effect sizes in the *r* metric obtained from a formula that uses the *t* ratios and the *df*--both of which are reported for treatment effects in HLM outputs--that is widely found in texts directed at meta-analysts (e.g., Rosenthal, 1991). Thus, the effect size for the mock data using this formula had to be converted to an algebraically equivalent *d* (McGrath & Meyer, 2006). The resulting *d* is incorrect because the *r* was biased, as it was calculated from statistics that incorporated the intercorrelations among the repeated measures and is thus an effect size based on the standard deviation of change scores. (Because a study by Beevers and Miller, 2005, reported HLM coefficient-based effect sizes in the *r* metric, it is likely they also used this formula, although it was not specified in their article.) An additional study (Sandler, Ayers, Wolchik, Tein, Kwok, et al., 2003) used an effect-size formula that was algebraically equivalent to that used by Atkins et al. (2005) and Schultz et al. (2006) but was expressed in the *d* rather than *r* metric. Thus, no conversion was needed to produce the same effect size found in Atkins et al. and Schultz et al. when applied to the mock data.

The final two studies (Fraser, Galinsky, Smokowski, Day, Terzian, et al., 2005; Rye, Pargament, Pan, Yingling, Shogren, & Ito, 2005) used non-standard formulas, without adequate explanations for their use. Neither of the two formulas used the estimate of the population standard deviation (i.e., standard deviation of raw scores) in the denominator nor produced the effect size of 2.60 found consistently in the analyses of the mock data.

Absence of Effect-size Formulas in the Prevention Literature

The review could not locate a single study in *Prevention Science* that provided formulas for effect size calculations. A comparison of reporting practices for GMA analyses between this journal and *JCCP* is problematical because of the confounding of journal with GMA methodology. The vast majority of GMA studies published in *JCCP* used HLM, and the vast majority of GMA studies published in *Prevention Science* used LGM. Nearly all reported model-based effect sizes observed in the GMA articles reviewed were calculated by investigators who used HLM (and published in *JCCP*).

A key reason for this difference in reporting practices is that the LGM approach is grounded in a structural equation modeling framework. When intervention evaluators use LGM, their reports generally include path analysis-type figures that display the regression coefficients for all the different paths. The diagrams include coefficients for the paths from the intervention to a latent “slope factor” that reflects the treatment effect. The mostly developmental researchers who use LGM to evaluate interventions are primarily interested in formulating a structural model that will not be rejected by a chi-square test, and less concerned with reporting analogues of *d* than with providing the goodness-of-fit indices (e.g., Comparative Fit Index, Root Mean Square Error of Approximation) associated with the tests of their models.

Discussion

Explanations for Effect-Size Discrepancies in JCCP

None of the formulas used to calculate the GMA-derived effect sizes reported in *JCCP* during a recent four-year period yielded the expected effect size of 2.60 when applied to the analyses of mock data. What is the source of the discrepancies in effect sizes? And why do most researchers using GMA (especially prevention researchers who conduct LGM analysis) fail to report model-based effect sizes at all, even when such reportage is mandatory for publication of research in APA journals (and journals like *Prevention Science* that recommend authors follow APA recommendations for article preparation)?

One likely reason why most interventions researchers do not report effect sizes for GMA is that the topic has not yet been covered in the texts and users' manuals for either HLM (e.g., Hedeker & Gibbons, 2006; Hox, 2002; Raudenbush & Bryk, 2002; Raudenbush et al., 2004) or LGM (e.g., Muthen & Muthen, 2007). Moreover, none of the key programs used to conduct GMA provide effect sizes associated with the coefficients they report (and test for statistical significance). Unfortunately, users of GMA programs often report the coefficient for the slope difference as if it were the effect size rather than an unstandardized regression value that equals the mean differences in slopes of arbitrarily-scaled outcome measures.

With limited guidance available on the computation of effect sizes for GMA in the same metric used in traditional designs, program evaluators have adapted procedures described in meta-analysis texts (e.g., Rosenthal, 1991), or reported effect sizes based on group differences in observed data (e.g., end-of-trial means for completers) rather than from estimates based on the GMA models (e.g., Blumenthal, Babyak, Carney, Keefe, Davis, et al., 2006). GMA researchers who cited Raudenbush and Liu's (2001) article on effect sizes for HLM designs all used the less appropriate of the two formulas the authors presented because that article's focus was on the effect size appropriate for use in power analysis for GMA (i.e., $d_{\text{GMA-CHANGE}}$, Eq. 6). The formula they provided for the effect magnitude ($d_{\text{GMA-RAW}}$, Eq. 7) was mentioned only in passing, and expressed verbally rather than mathematically (see the last complete paragraph on p. 393, where Raudenbush and Liu calculated a $d_{\text{GMA-RAW}}$ of .30 for illustrative data from a four-year study where β_{11} was .0219 and the population standard deviation was estimated as 15).

Most important, the literature addressing the differences between the two types of effect sizes in repeated measures has limited the discussion to traditional designs (e.g., Morris & DeShon, 2002). Thus, the present article is the first to discuss the relevance of these issues to GMA. Also, the authors who have advocated the use of raw score standard deviations for effect size calculations have directed their concerns to meta-analysts (Dunlap et al., 1996; Hunter & Schmidt, 2004: Lipsey & Wilson, 2001) when it is *primary researchers* who need to calculate and report effect sizes associated with significance tests for the efficacy of their interventions (Fidler, Cumming, Thomason, Pannuzzo, Smith, et al., 2005). As a result, the important distinctions between the two types of effect sizes are not widely known.

Implications for Power Analysis

Because GMA outputs can be used to estimate the same effect size parameter as results from an independent groups analysis, the GMA effect size for efficacy (i.e., $d_{\text{GMA-RAW}}$, Eq. 7) can be used in the calculation of power for a planned independent groups analysis (as well as for both within-subjects and IGPP analyses if either the pretest-posttest correlation or population standard deviation is also known). However, the effect size calculated using the standard deviation of change scores ($d_{\text{GMA-CHANGE}}$, Eq. 6) that is appropriate for use in the power calculations for HLM should not be used in power calculations for ANOVA. By contrast, effect

sizes calculated from observed means (i.e., from traditional analyses) cannot be used as the required statistics for calculation power for an HLM study using the methods and programs of Raudenbush and his colleagues (Raudenbush & Liu, 2001;Spybrook, Raudenbush, Liu, & Congdon, 2006) because the HLM power analysis requires an effect size based on the standard deviations of the "true" scores of the slopes (i.e., τ, Eq. 6). Because this effect size can only be obtained from an HLM study, only $d_{\text{GMA-CHANGE}}$ from a previous HLM analysis can be used for calculating power for a planned HLM analysis.

Implications for Meta-Analysis

Both types of GMA effect sizes may be fruitfully used in meta-analyses but for different purposes. The *d*s based on the standard deviation of model-based estimates of change ($d_{\text{GMA-CHANGE}}$, Eq. 6) could be cumulated from a series of HLM studies to obtain an estimate of the effect size that would be useful for power analysis in planning future HLM studies. Indeed, a recently developed program to calculate power for HLM analysis requires the user to estimate $d_{\text{GMA-CHANGE}}$ (Spybrook et al., 2006), and meta-analyses of these values from prior HLM analyses would be the ideal means to obtain such estimates. Additionally, moderator variables analysis could be employed when *d*s are heterogeneous to identify study characteristics associated with the effect sizes. This would allow a researcher planning a study with an HLM design to assume the $d_{\text{GMA-CHANGE}}$ that would be expected given the study characteristics of the planned research.

By comparison, the *d*s calculated with the standard deviation of raw scores ($d_{\text{GMA-RAW}}$, Eq. 7) should be included in a meta-analysis that examines the strength of the treatment effect by pooling outcomes from different GMA studies, or studies using diverse designs that include GMA. A key difficulty with using either effect size in a meta-analysis to which they should appropriately contribute relates to the sampling variance associated with each effect size (Morris & DeShon, 2002). The sampling variances of the effect sizes, which needs to be included in a meta-analysis, have been derived for traditional designs (Becker, 1998;Glass, McGaw & Smith, 1981;Hedges & Olkin, 1985;Morris & DeShon, 2002) but not for GMA.

Until statisticians have derived the sampling variances for GMA effect sizes, psychologists need interim solutions. It would almost certainly be unwise to exclude a randomized clinical trial from a meta-analysis because the sampling variability is not precisely known when an unbiased estimate of the effect size and the sample size are both available. It also would be inappropriate to use in a meta-analysis end-of-treatment statistics from a completers' subsample to calculate an effect size for "independent groups" because any advantages gain by conducting a GMA at the empirical level would be lost at the meta-analytic level.

One possible alternative would be to enter into a standard meta-analysis program the GMA effect size and the requisite sample sizes. The program would then handle the effect size exactly as if it had been obtained from an independent group design with the same sample size. This approach would be superior to using end-of-treatment data to calculate *d* because the effect size used would be based on data from all subjects using end-of-study means estimated by the GMA model.

HLM versus LGM Designs

Although the mock dataset was analyzed by HLM to illustrate calculation of effect sizes, the identical mean difference between slopes can be obtained in LGM, thus allowing for the computation of $d_{\text{GMA-RAW}}$ using outputs from LGM programs (e.g., Mplus). Because of the mathematical model used, however, no within-group variations in change estimates are available for LGM (Muthen & Curran, 1997). Thus, $d_{\text{GMA-CHANGE}}$ cannot be calculated using

LGM outputs, and other methods are needed to compute power analyses for studies using LGM (see Muthen & Curran, 1997).

Choice of Standard Deviation for Use in Calculation of GMA Effect Sizes

Although there is a consensus that the denominator in the effect size should use the estimate of the standard deviation of the outcome in the population, there is a debate as to whether that estimate should be based on the pooled-within group standard deviation, or obtained from the control group alone. This controversy has existed since the dawn of meta-analysis (Glass et al., 1981). Most meta-analytic procedures used today assume homogeneity of variance, and work with variances pooled from the two groups to yield a more stable estimate of within-group variability (e.g., Hedges & Olkin, 1985). This is also a practical approach because the meaningfulness of *d* is difficult to interpret when groups differ simultaneously in central tendency and variability (Feingold, 1995). However, HLM can be used to examine variability differences to test the assumption, and to compare groups when both types of differences are significant (Raudenbush et al., 2004). More important, the concern with estimating variability from the control group alone really applies only to completely randomized designs, in which all estimates of variability are obtained after the administration of the intervention. When baseline data are collected at the onset of a study in which participants have been randomized to groups (as in both IGPP and GMA), all descriptive parameters should be identical in both groups.

Intercept and Non-linear Polynomial Differences between Groups

As with the effect size for IGPP ($d_{\text{IGPP-RAW}}$, Eq. 5), the end-of-study effect size for GMA ($d_{\text{GMA-RAW}}$, Eq. 7) does not compare the means between the groups at the end of the study but the degree to which that mean difference is greater at the end than at the onset. Assuming random assignment at the start of the study, the expected mean difference between the two groups at baseline is zero. Thus, the expected value of the numerator in the calculation of $d_{\text{GMA-RAW}}$ at end of the study is the degree to which the mean is higher for treatment than controls at that time. (However, $d_{\text{GMA-RAW}}$ is corrected for baseline group differences that occur as the result of sampling errors.)

The concepts delineated to explain effect sizes for trajectory comparisons between intervention and control groups can be extended to analyses in which two groups are not formed by randomization, such as gender (e.g., Raudenbush, 1995) or diagnostic status (e.g., Feingold, Kerr, & Capaldi, 2008). Because men and women, for example, may differ at baseline as well as in their growth rates over time, the two types of end-of-study mean differences may not be the same. Including both the intercept and linear slopes in the model to estimate end-of-study means would yield an effect size that would reflect the sex difference at the end of the study rather than the degree to which the sex difference was greater at the end than at the onset. In addition, a quadratic term can be included in the growth model when differences in rates of acceleration are hypothesized (Raudenbush & Liu, 2001). Indeed, the capability to examine non-linear trends over the course of a study is a key reason to use GMA over IGPP. In general, all polynomial functions that differentiate between groups need to be considered in determining the model-based estimate of end-of-study means for each group, and Eq. 7 is a simplified formula that is applicable when only a linear trend without an intercept is included in the GMA model. (See Atkins, 2005, for an example of an approach that derives separate Level 1 linear and quadratic coefficients as a result of Level 2 coefficients associated with the between-condition differences for these two polynomials, and how these separate equations can be used to obtain an estimate of the end-of-study mean for each group. The difference between these two estimated means can then be divided by the estimate of the population standard deviation to obtain $d_{\text{GMA-RAW}}$.)

Conclusions

Researchers who conduct controlled clinical trials need to calculate effect sizes that are comparable across studies and suitable for inclusion in meta-analyses that combine studies using different statistical designs (Olejnik & Algina, 2003). Such comparability and synthesis are needed to advance the fields of treatment and prevention sciences, particularly when combined with rigorous criteria for primary research (e.g., Flay, Biglan, Boruch, Castro, Gottfredson, Kellam, et al., 2005). Effect-sizes for GMA can be expressed in same metric as those calculated from classical designs by dividing the model-based estimates of mean differences at the end of the study by the within-group standard deviation of raw scores that estimates the variability in the outcome measure at baseline.

Acknowledgments

This project was supported by Award Number R01HD046364 from the Eunice Kennedy Shriver National Institute of Child Health & Human Development. The content is solely the responsibility of the author and does not necessarily represent the official views of the Eunice Kennedy Shriver National Institute of Child Health & Human Development or the National Institutes of Health.

The author thanks Deborah Capaldi, Xiaofeng Liu, and William Shadish for their helpful comments on a draft of this article.

References

American Psychological Association. Publication manual of the American Psychological Association. Vol. 5th ed.. American Psychological Association; Washington, DC: 2001.

Antoni MH, Lechner SC, Kazi A, Wimberly SR, Sifre T, et al. How stress management improves quality of life after treatment for breast cancer. Journal of Consulting and Clinical Psychology 2006;74:1143–1152. [PubMed: 17154743]

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]

Atkins DC, Berns SF, George WH, Doss BD, Gattis K, Christensen A. Prediction of response to treatment in a randomized clinical trial of marital therapy. Journal of Consulting and Clinical Psychology 2005;73:893–903. [PubMed: 16287389]

Atkins DC, Eldridge KA, Baucom DH, Christensen A. Infidelity and behavioral couple therapy: Optimism in the face of betrayal. Journal of Consulting and Clinical Psychology 2005;73:144–150. [PubMed: 15709841]

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

Beevers CG, Miller IW. Unlinking negative cognition and symptoms of depression: Evidence of a specific treatment effect for cognitive therapy. Journal of Consulting and Clinical Psychology 2005;73:68–77. [PubMed: 15709833]

Blumenthal JA, Babyak MA, Carney RM, Keefe FJ, Davis RD, et al. Telephone-based coping skills training for patients awaiting long transplanation. Journal of Consulting and Clinical Psychology 2006;74:535–544. [PubMed: 16822110]

Bock, RD. Within-subject experimentation in psychiatric research.. In: Gibbons, RD.; Dysken, MW., editors. Statistical and methodological advances in psychiatric research. Spectrum; New York: 1983. p. 59-90.

Brown EC, Catalano RF, Fleming CB, Haggerty KP, Abbott RD. Adolescent substance use outcomes in the raising healthy children project: A two-part latent growth curve analysis. Journal of Consulting and Clinical Psychology 2005;73:699–710. [PubMed: 16173857]

Christensen A, Atkins DC, Berns S, Wheeler J, Baucom DH, Simpson LE. Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples. Journal of Consulting and Clinical Psychology 2004;72:176–191. [PubMed: 15065953]

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

Curran PJ, Muthen BO. The application of latent curve analysis for testing developmental theories in intervention research. American Journal of Community Psychology 1999;27:567–595. [PubMed: 10573835]

DeGarmo DM, Patterson GR, Forgatch MS. How do outcomes in a specified parent training intervention maintain or wane over time? Prevention Science 2004;5:73–89. [PubMed: 15134313]

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

Dimidjian S, Hollon SD, Dobson KS, Schmaling KB, Kohlenberg RJ, et al. Randomized trial of behavioral activation, cognitive therapy, and antidepressant medication in the acute treatment of adults with major depression. Journal of Consulting and Clinical Psychology 2006;74:658–670. [PubMed: 16881773]

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.

Feingold A. Assessment of journals in social science psychology. American Psychologist 1989;44:961–964.

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

Fidler F, Cumming G, Thomason N, Pannuzzo D, Smith J, et al. Toward improving statistical reporting in the Journal of Consulting and Clinical Psychology. Journal of Consulting and Clinical Psychology 2005;73:136–143. [PubMed: 15709840]

Flay BR, Biglan A, Boruch RF, Castro FG, Gottfredson D, Kellam S, et al. Standards of evidence: Criteria for efficacy, effectiveness, and dissemination. Prevention Science 2005;6:151–175. [PubMed: 16365954]

Fraser MW, Galinsky MJ, Smokowski PR, Day SH, Terzian MR, et al. Social information-processing skills training to promote social competence and prevent aggressive behavior in the third grades. Journal of Consulting and Clinical Psychology 2005;73:1045–1055. [PubMed: 16392978]

Gibbons RD, Hedeker DR, Davis JM. Estimation of effect sizes from a series of experiments involving paired comparisons. Journal of Educational Statistics 1993;18:271–279.

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. Sage; Beverly Hills, CA: 1981.

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]

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

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

Hox, J. Multilevel analysis: Techniques and applications. Erlbaum; Mahwah, NJ: 2002.

Hunter, JE.; Schmidt, FL. Methods of meta-analysis: Correcting error and bias in research findings. Vol. 2nd ed.. Sage; Thousand Oaks, CA: 2004.

Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13–22.

Lipsey MW, Wilson DB. The efficacy of psychological, educational and behavioral treatment: Confirmation from meta-analysis. American Psychologist 1993;48:1181–1209. [PubMed: 8297057]

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

McGrath RE, Meyer GJ. When effect sizes disagree: The case of *r* and *d*. Psychological Methods 2006;11:386–401. [PubMed: 17154753]

Monson CM, Schnurr PP, Resick PA, Friedman MJ, Young-Xu Y, Stevens SP. Cognitive processing therapy for veterans with military-related posttraumatic stress disorder. Journal of Consulting and Clinical Psychology 2006;74:898–907. [PubMed: 17032094]

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]

Mullen, B.; Rosenthal, R. Basic meta-analysis: Procedures and programs. Erlbaum; Hillsdale, NJ: 1985.

Muthen BO, Curran PJ. General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. Psychological Methods 1997;2:371–402.

Muthen, LK.; Muthen, BO. Mplus user's guide. Vol. 4th ed. Muthen and Muthen; Los Angeles, CA: 2007.

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]

Petitti, DB. Meta-analysis, decision analysis, and cost-effectiveness analysis: Methods for quantitative synthesis in medicine. Vol. 2nd ed.. Oxford; New York: 2000.

Raudenbush, SW. Hierarchical linear models to study the effects of social context on development.. In: Gottman, JM., editor. The analysis of change. Erlbaum; Mahwah, NJ: 1995. p. 167-201.

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

Raudenbush, S.; Bryk, A.; Cheong, Y.; Congdon, R.; du Toit, M. HLM 6: Hierarchical linear and nonlinear modeling. Scientific Software International; Lincolnwood, IL: 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. Meta-analytic procedures for social research. Vol. 2nd ed.. Sage; Newbury Park, CA: 1991.

Rye MS, Pargament KI, Pan W, Yingling DW, Shogren KA, Ito M. Can group interventions facilitate forgiveness of an ex-spouse? A randomized clinical trial. Journal of Consulting and Clinical Psychology 2005;73:880–892. [PubMed: 16287388]

Sandler IN, Ayers TS, Wolchik SA, Tein J, Kwok O, et al. The family bereavement program: Efficacy evaluation of a theory-based prevention program for parentally bereaved children and adolescents. Journal of Consulting and Clinical Psychology 2003;71:587–600. [PubMed: 12795581]

Schulz MS, Cowan CP, Cowan PA. Promoting healthy beginnings: A randomized controlled trial of a preventive intervention to preserve marital quality during the transition to parenthood. Journal of Consulting and Clinical Psychology 2006;74:20–31. [PubMed: 16551140]

Shadish, WR.; Cook, TD.; Campbell, DT. Experimental and quasiexperimental designs for generalized causal inference. Houghton-Mifflin; Boston: 2002.

Shadish, WR.; Robinson, L.; Lu, C. ES: A computer program for effect size estimation. Assessment Systems Corporation; St. Paul, MN: 1999.

Simpson JN, McGrath PJ, Yamada JT. Clinical trials in the *Journal of Pediatric Psychology*: Applying the CONSORT statement. Journal of Pediatric Psychology 2003;28:159–167. [PubMed: 12654939]

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

Snedecor, GW.; Cochran, WG. Statistical methods. Vol. 6th ed. Iowa State University Press; Ames: 1967.

Spybrook, J.; Raudenbush, SW.; Liu, X.; Congdon, R. Optimal design for longitudinal and multilevel research: Documentation for the “Optimal Design” software. University of Michigan; 2006. Unpublished manuscript

Tolan P, Gorman-Smith D, Henry D. Supporting families in a high-risk setting: Proximal effects of the SAFEChildren preventive intervention. Journal of Consulting and Clinical Psychology 2004;72:855–869. [PubMed: 15482043]

Trudeau L, Spoth R, Lillehoj C, Redmond C, Wickrama KAS. Effects of a preventive intervention on adolescent substance use initiation, expectancies, and refusal intentions. Prevention Science 2003;4:109–122. [PubMed: 12751880]

van Lier PAC, Muthen BO, van der Sar RM, Crijnen AAM. Preventing disruptive behavior in elementary schoolchildren: Impact of a universal classroom-based intervention. Journal of Consulting and Clinical Psychology 2004;72:467–478. [PubMed: 15279530]

Weisz JR, Jensen-Doss A, Hawley KM. Evidence-based youth psychotherapies versus usual clinical care: A meta-analysis of direct comparisons. American Psychologist 2006;61:671–689. [PubMed: 17032068]

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

Table 1

Mock Data Used for Analysis of Different Clinical Trial Designs

Subj ID	Tx	T1	T2	T3	T4	T4 - T1
			Control Group			
101.00	−.50	3.00	5.00	5.00	7.00	4.00
102.00	−.50	4.00	4.00	6.00	6.00	2.00
103.00	−.50	4.00	5.00	7.00	8.00	4.00
104.00	−.50	5.00	6.00	6.00	8.00	3.00
105.00	−.50	5.00	6.00	7.00	8.00	3.00
106.00	−.50	5.00	7.00	7.00	7.00	2.00
107.00	−.50	5.00	6.00	8.00	8.00	3.00
108.00	−.50	6.00	6.00	7.00	9.00	3.00
109.00	−.50	6.00	8.00	9.00	10.00	4.00
110.00	−.50	7.00	7.00	8.00	9.00	2.00
Mean		5.00	6.00	7.00	8 .00	3.00
			Intervention Group			
		T1	T2	T3	T4	T4 - T1
111.00	.50	3.00	5.00	7.00	9.00	6.00
112.00	.50	4.00	7.00	9.00	11.00	7.00
113.00	.50	4.00	6.00	8.00	11.00	7.00
114.00	.50	5.00	7.00	9.00	10.00	5.00
115.00	.50	5.00	6.00	9.00	11.00	6.00
116.00	.50	5.00	7.00	10.00	10.00	5.00
117.00	.50	5.00	8.00	8.00	11.00	6.00
118.00	.50	6.00	7.00	9.00	12.00	6.00
119.00	.50	6.00	9.00	11.00	13.00	7.00
120.00	.50	7.00	8.00	10.00	12.00	5.00
Mean		5.00	7.00	9.00	11.00	6.00

Note. The standard deviation associated with each mean is 1.1547. Subj ID = subject number, Tx = treatment code (−.50 = control, .50 = treatment), T4 - T1 = change (gain) score.

Psychol Methods. Author manuscript; available in PMC 2009 July 19.

Table 2

Formulas for Growth-Modeling Effect Sizes in the Clinical Literature Applied to Mock Data in Table 1

Study	Reported Formula	*d*
Antoni et al., 2006	$d = \beta/(\tau)^{1/2}$	29.36
Atkins, Berns et al., 2005	$r = (t^2/(t^2 + df))^{1/2}$	4.15
Atkins, Eldridge et al., 2005	$d = \beta/(\tau)^{1/2}$	29.36
Brown et al. 2005	$d = \beta/(\tau)^{1/2}$	29.36
Christensen et al., 2004	$d = \beta/(\tau)^{1/2}$	29.36
Dimidjian et al., 2006	$d = \beta/(\tau)^{1/2}$	29.36
Fraser et al., 2005	$d = \beta/(\tau + \sigma^2)^{1/2}$	.89
Rye et al., 2005	$d = \beta/\sigma$	.89
Sandler et al., 2003	$d = t\,(n_1 + n_2)/(df)^{1/2}(n_1 n_2)^{1/2}$	4.15
Schulz et al., 2006	$r = (t^2/(t^2 + df))^{1/2}$	4.15
Tolan et al., 2004	$d = \beta/(\tau)^{1/2}$	29.36

Note. *d* is the effect size when the reported formula was applied to the statistics generated from HLM analysis of the mock data in Table 1 (with *r*-to-*d* conversion applied when tabled formula was expressed in the *r* metric). In HLM, σ^2 refers to the Level 1 error variance.