

RUNNING HEAD: HLM effect size

Effect size measures for the two-level linear multilevel model

J. Kyle Roberts  
Baylor College of Medicine

James P. Monaco  
VuCOMP, Inc.

---

Paper presented at the annual meeting of the American Educational Research Association, April 9, 2006. Correspondence should be sent to J. Kyle Roberts, One Baylor Plaza, MS: BCM 411, Houston, TX 77030-3411, or [jkrobert@bcm.edu](mailto:jkrobert@bcm.edu). The authors would like to thank Ron Harrist for comments that lead to the development of this manuscript.

## Abstract

With the rise of the use and utility of hierarchical linear modeling (HLM), one question has consistently been posed to authors and on listserves: “How much variance does my model explain?” Answering this question within the HLM framework is not an easy task where it is actually possible to explain “negative variance” when the addition of explanatory variables increases the corresponding variance components (Snijders & Bosker, 1999). Because effect size measures previously proposed consider variance at each level, a single measure is needed which helps researchers interpret the strength of the model as a whole. The purpose of this paper is to provide a history of past HLM effect sizes and present three new measures which consider “whole model” effects.

## Effect size measures for the two-level linear multilevel model

The utility of effect sizes in research interpretation has generated considerable discussion, much of which centers on the role and function of effect sizes, especially concerning the relationship to statistical significance tests (cf. Harlow, Mulaik, & Steiger, 1997). Many authors agree that effect sizes can serve a valuable function to help evaluate the magnitude of a difference or relationship (cf. Cohen, 1994; Kirk, 1996; Schmidt, 1996; Shaver, 1985; Thompson, 1996; Wilkinson & APA Task Force on Statistical Inference, 1999). Their articles, along with current publications (c.f., Knapp & Sawilowsky, 2001a; Roberts & Henson, 2002) continue to debate both the use and utility of measures of effect size when considered both in conjunction with and peripheral from statistical significance testing.

One positive thing that has occurred while researchers began to debate the issue of effect size reporting (e.g., Knapp & Sawilowsky, 2001b; Thompson, 2001) is the encouragement of researchers to consider more than just the magical “p-value” before making interpretations as to the noteworthiness (or lack thereof) of a given study. And although it may seem that the field of research follows changes and adopts a new course with the speed and acuteness of a glacier, the fact that this glacier is moving is predicated by the adoption of such language in the *APA Publication Manual* (2001):

The general principle to be followed, however, is to provide the reader not only with information about statistical significance but also with enough information to assess the magnitude of the observed effect or relationship. (p. 26)

A history of effect sizes has been dealt with exhaustively in Huberty (2002), and does not bear repeating here. One thing absent from Huberty's catalog was the use of effect size indices in multilevel analysis. It was probably wisely absent from Huberty's manuscript, since there is much misconception, and even disagreement, as to the interpretation of these effects. We will quickly list some of the proposed effect size indices for use in hierarchical linear modeling (HLM) and multilevel modeling and give brief explanations as to their utility.

### *Intraclass Correlation*

Intraclass correlation (ICC) is generally thought of as the degree of dependence of individuals upon a higher structure to which they belong; or, the proportion of total variance that is between the groups of the regression equation. Put more succinctly, it "is the degree to which individuals share common experiences due to closeness in space and/or time" (Kreft & Leeuw, 1998, p. 9). Hox (1995) explains the ICC as a "population estimate of the variance explained by the grouping structure" (p. 14). The ICC for a 2-level model can be represented as:

$$\rho_1 = \frac{\tau_0^2}{\tau_0^2 + \sigma^2} = \frac{\sigma_{u0}^2}{\sigma_{u0}^2 + \sigma_{e0}^2}, \quad (1)$$

where the numerator is represented by the variance at the second level of the hierarchy ( $\tau_0^2$ ), and the denominator represents the total variation in the model at both level-2 and level-1 ( $\sigma^2$ ).

Although the ICC actually is *not* a measure of the effect size of an HLM model, it bears mentioning here because it sometimes is wrongly thought of as a measure of the "power" or strength of HLM over ordinary least squares (OLS) regression. Roberts (2002) has rightly pointed out that it would be incorrect to interpret this statistic as a measure of the magnitude of difference between OLS and HLM estimates.

### *Proportion Reduction in Variance*

The process of building a multilevel model often begins with a null model (also called the baseline model by Hox, 2002). In this baseline model, just the grand mean is fit in the model such that:

$$y_{ij} = \gamma_{00} + u_{0j} + e_{ij}, \quad (2)$$

where,  $\gamma_{00}$  is the model grand mean (or intercept),  $u_{0j}$  is the random group level effect with variance  $\sigma_{u0}^2$ , and  $e_{ij}$  is the person level effect with variance  $\sigma_e^2$ . This baseline model's variance estimates serve as a benchmark for determining the  $R^2$  at each level of the hierarchy. By using variance estimates from the null model (e.g.,  $\sigma_{u0b}^2$ ) and variance estimates from the model where all predictors are entered (e.g.,  $\sigma_{u0m}^2$ ), the percent reduction in variance between the null model and the complete model can be estimated by:

$$R_2^2 = \frac{\sigma_{u0b}^2 - \sigma_{u0m}^2}{\sigma_{u0b}^2}, \quad (3)$$

for the percent reduction in level-2 variance and by:

$$R_1^2 = \frac{\sigma_{eb}^2 - \sigma_{em}^2}{\sigma_{eb}^2}, \quad (4)$$

for the percent reduction in level-1 variance, where  $|b$  and  $|m$  represent the baseline and full models, respectively. This formula is reflected in many forms by Hox (2002, p. 64), or conversely as:

$$R_2^2 = \frac{\tau_{00}(null) - \tau_{00}(full)}{\tau_{00}(null)} \quad (5)$$

by Raudenbush and Bryk (2002, p. 74) and Kreft and de Leeuw (1998, p. 118).

Although perceived as a tool for noting the reduction in variance at each level of the model, Hox (2002) and Snijders and Bosker (1999) are quick to caution researchers against

directly interpreting this statistic, since it is possible to obtain negative values for  $R^2$  with these formulas when either  $\sigma_e^2$  is a biased estimator or when level-2 predictors are included in the model. This is a difficult concept to grasp, because in normal OLS models, the addition of variables to the model can only help prediction of the dependent variable (raise  $R^2$ ), not hurt prediction. A negative  $R^2$  value in HLM might be wrongly interpreted to mean that the predictor variables are performing at worse levels than just the grand mean as a predictor.

Negative variance can occur in an example where we use a variable that has almost no variation at one of the levels. Consider the case of a model where we have one single level-1 predictor. It would be safe to assume that the addition of this variable would reduce both the between and within groups variance. If we add a group-level predictor, then we could expect that it would reduce only the between-groups variance, not the within-groups variance, ultimately increasing the estimate for the population variance  $\hat{\sigma}_{u0}^2$ . For example, consider the output in Table 1 from a two-level model.

---

Insert Table 1 about here

---

The data in this example were adapted from a hypothetical dataset written to illustrate multilevel models (Roberts, 2004). In model M1, just a single level-1 predictor is included in the model with variance estimates of  $\hat{\sigma}_e^2 = 0.651$  and  $\hat{\sigma}_{u0}^2 = 80.923$ . If we are to consider this in terms of equation 4, we could see that we actually would be decreasing the variance explained (or increasing  $\hat{\sigma}_{u0}^2$ ) at the second level with the introduction of this predictor (hence negative variance explained between the two models).

Although the amount of variance explained is noteworthy at level-1 ( $R_1^2 = (1.979 - .0651)/1.979 = 0.671$ ), the amount of variance explained at the second level is actually -2.381 ( $R_2^2 = (23.923 - 80.890)/23.923 = -2.381$ ). Not only is this number troubling, but it is counter-intuitive to the way most researchers think about the effectiveness of a model. If we were to interpret this model without previous knowledge of multilevel models, we might be inclined to say that the addition of the predictor “ses” is a worse predictor of “math achievement” at the second level than if we had no predictor at all, being that it explains -238% variance!

*Explained Variance as a Reduction in Mean Square Prediction Error*

Snijders and Bosker (1999) argue for a slightly different approach to computing  $R^2$  values in multilevel models by computing the model’s associated mean square prediction error. The  $R^2$  for level-1 is then computed as one minus the combined variance at both levels for the full model divided by the combined variance for the null model, or:

$$R_1^2 = 1 - \frac{\text{var}(Y_{ij} - \sum_h \gamma_h X_{hij})}{\text{var}(Y_{ij})} = 1 - \frac{\hat{\sigma}^2(\text{full}) + \hat{\tau}_0^2(\text{full})}{\hat{\sigma}^2(\text{null}) + \hat{\tau}_0^2(\text{null})}, \quad (6)$$

where  $Y_{ij}$  is the outcome variable,  $\gamma_h$  represents the coefficient for outcome variable  $X_{hij}$  for all  $h$  variables,  $\hat{\sigma}^2$  is an estimate of the variance at the first level, and  $\hat{\tau}_0^2$  is an estimate of the variance at the second level.

The level-2  $R^2$  is then found by dividing the  $\hat{\sigma}^2$  by the group cluster size ( $B$ ), or by the average cluster size for unbalanced data, such that:

$$R_2^2 = 1 - \frac{\text{var}(\bar{Y}_{.j} - \sum_h \gamma_h \bar{X}_{h.j})}{\text{var}(\bar{Y}_{.j})} = 1 - \frac{\frac{\hat{\sigma}^2(\text{full})}{B} + \hat{\tau}_0^2(\text{full})}{\frac{\hat{\sigma}^2(\text{null})}{B} + \hat{\tau}_0^2(\text{null})}. \quad (7)$$

In this formula, it is easy to see that the  $R^2$  estimate at level-2 is similar the  $R^2$  for level-1, having just reduced the level-1 variance to represent an average variance for each group. Although this estimation differs from the previous definition of  $R^2$  (Equations 3 & 4), it is still possible to obtain “negative” values for  $R^2$ .

Using Table 1,  $R^2$  at level-1 is:

$$R_1^2 = 1 - \frac{0.651 + 80.890}{1.979 + 23.923} = -2.148,$$

and for level-2 is:

$$R_2^2 = 1 - \frac{\frac{0.651}{10} + 80.890}{\frac{1.979}{10} + 23.923} = -2.356.$$

Once again this solution is extremely problematic, as we have again obtained negative values for  $R^2$ .

#### Proposed New Distance Measures for Calculating $R^2$ in Multilevel Models

In turning our thoughts to multiple regression, the multiple  $R^2$  can be thought of as the correlation between a function of the predictor variables and the dependent variable. Another way to think about this value is that it is the correlation between the dependent variable,  $y$ , and  $\hat{y}$ , the values of the dependent variable predicted from the independent variables. This association can be seen in the following figure where  $x_1$ ,  $x_2$ , and  $x_3$  are all predictors of  $y$ .

---

Insert Figure 1 about here

---

Therefore we could write a formula to represent Figure 1 as:

$$R_{y(x_1, x_2, x_3)}^2 = R_{y\hat{y}}^2. \quad (8)$$

If we are to think theoretically about this formula, we can describe  $\hat{y}$  as simply any given person's predicted  $y$  score based on the weights derived for each independent variable. The distance between an individual's predicted score,  $\hat{y}$ , and their observed score,  $y$ , would simply be thought of as error, or variance unaccounted for.

Although this is a relatively simple formula, it can be applied easily to HLM, if we are to think of HLM as belonging to the General Linear Model of statistics. If we consider that the point of any analysis is to try to produce a series of coefficients that closely approximates an individual's original score, then we can see that in HLM, the  $\hat{y}$  is simply any predicted value based on a set of regression coefficients derived from the HLM model, and the error term is simply the distance between  $\hat{y}$  and  $y$ . In the case of hierarchical linear modeling, these weights are derived through maximum likelihood estimates of the fixed effects, with the individual estimates being the product of empirical Bayesian estimates.

Although it would seem that a researcher could simply correlate these  $\hat{y}$  and  $y$  values to obtain an estimate of  $R^2$ , we must remember and maintain in HLM that we wish to honor the procedure by which the estimates were obtained. In ordinary least squares regression, we can typically compute the total variance in the model as:

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 . \quad (9)$$

In a multilevel model, we must remember that we typically define the null model as having both fixed effects (the grand mean for the dependent variable) and random effects (the variation of each group's mean around that grand mean). In defining our model in this manner, we cannot simply compute the total variance in the same manner as it is computed in Equation 9.

Instead, the total variance must be the predicted values of  $y$  when only the grand mean is used as a predictor, or:

$$\tilde{y}_{ij} = \gamma_{00}(\text{cons}) + u_{0j} . \quad (10)$$

Notice that the random estimate for individuals has been left out of equation 10. By doing this, the grand mean value of the dependent variable is being estimated for the entire model and the random level-2 deviate  $u_{0j}$  (also known as  $\beta_{0j}$  for each group). We then can compute the total amount of variance in this model as:

$$\sigma_{total}^2 = \sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ij} - \tilde{y}_j)^2 = e_{ij}^2 , \quad (11)$$

where  $\tilde{y}_j$  is the random estimate for group  $j$  and  $y_{ij}$  is the original outcome score for person  $i$  in group  $j$ . This number is the same value as the sum of the square of all of the residual values  $e_{ij}$ , but is distinctly different from the variance estimate from the unbiased estimator of  $\sigma^2$  which contains a correction factor for the  $Q + 1$  regression parameters such that:

$$\hat{\sigma}^2 = \sum e_{ij}^2 / (n - Q - 1) . \quad (12)$$

As noted in the OLS model, the error variance in a model can be viewed as the distance between  $\hat{y}$  and  $y$ . Likewise, in HLM after the  $\hat{y}$  are calculated using the full model, the error variance for the total model can be explained as:

$$\sigma_{error}^2 = \sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ij} - \hat{y}_{ij})^2 , \quad (13)$$

where  $\hat{y}_{ij}$  is the predicted value for person  $i$  in group  $j$  based on the full model:

$$\hat{y}_{ij} = \gamma_{00} + \gamma_{q0}X_{ij} + \gamma_{0q}W_j + u_{0j} + e_{ij} , \quad (14)$$

and  $W_j$  is a level-2 predictor. Once these values are computed, an  $R^2$  value may be computed as:

$$1 - \frac{\sigma_{error}^{\prime 2}}{\sigma_{total}^{\prime 2}} . \quad (15)$$

*R<sup>2</sup> Measure that Incorporates a Gaussian Probability Density Function*

So far, these formulas are not very dissimilar from the previously proposed estimators of variance explained, with the difference being that they do not use the unbiased estimator for variance. However, consider if we were to aggregate this formula to the level-2 grouping structure such that we gain an  $R^2$  value for each level-2 group and then average across all groups. Doing so would further enhance the above formulas such that the estimate of variance explained would be defined by:

$$R_j^2 / k , \quad (16)$$

where  $k$  is the number of level-2 groups and,

$$R_j^2 = 1 - \frac{e_i^2(\text{full})}{e_i^2(\text{null})} , \quad (17)$$

with  $e_i^2$  representing the measure of each residual for the  $i$ th person in each distinct group  $j$  for both the full model and the null model. Although this would represent an “average”  $R^2$  for the entire model by producing a mean  $R^2$  based on each group’s  $R^2$ , it does not take into account that the original estimation method used to produce these values was based on a probability of inclusion in the model from maximum likelihood estimates. Unless each group has exactly the same sample size and the same probability of selection, it would not follow to use Equations 16 and 17 to solve for a model  $R^2$ .

By inserting the Gaussian probability density function into the above equations, we could gain a measure of  $R^2$  as a function of the probability of inclusion of the given value assuming the model. Doing so would modify equation 11 such that:

$$\sigma_{total}^2 = \frac{\sum_{ij} p(y_{ij})(e'_{ij})^2}{\sum_{ij} p(y_{ij})}, \quad (18)$$

where  $e'_{ij}$  is an estimate of the residual for each  $i$  person in the  $j$ th group in the null model, and:

$$p(y_{ij}) = p(d_{ij} | s_j) p(s_j) \quad (19)$$

where  $p(d_{ij} | s_j)$  is the probability of the person  $i$ , given that they belong to the  $j$ th group, and  $p(s_j)$  is the probability of group  $j$ , given the entire sample of level-2 units. In extrapolating Equation 19 further and applying the probability density function from a Gaussian distribution, it can be shown that:

$$p(y_{ij}) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2 \cdot \sigma_j^2}} \cdot \exp\left[-\frac{(e'_{ij})^2}{2\sigma_{ij}^2}\right] \cdot \exp\left[-\frac{(u'_j)^2}{2\sigma_j^2}\right], \quad (20)$$

with  $\sigma_{ij}^2$  representing the variance of the  $i$  individuals around their  $j$ th group mean for the null model,  $\sigma_j^2$  is the variance of  $\beta_{0j}$  around  $\gamma_{00}$ , and  $e'_{ij}$  and  $u'_j$  are the residual scores for the level-1 and level-2 estimates, respectively.

As would follow from Equation 13, the model error could be thought of as:

$$\sigma_{error}^2 = \frac{\sum_{ij} p(y_{ij})(e''_{ij})^2}{\sum_{ij} p(y_{ij})}, \quad (21)$$

where  $e''_{ij}$  is an estimate of the residual for each  $i$  person in the  $j$ th group in the full model, and:

$$p(y_{ij}) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2 \cdot \sigma_j^2}} \bullet \exp\left[-\frac{(e_{ij}'' )^2}{2\sigma_{ij}^2}\right] \bullet \exp\left[-\frac{(u_j'')^2}{2\sigma_j^2}\right], \quad (22)$$

with  $\sigma_{ij}^2$  representing the variance of the  $i$  individuals around their  $j$ th group mean for the full model,  $\sigma_j^2$  is the variance of  $\beta_{0j}$  around  $\gamma_{00}$ , and  $e_{ij}''$  and  $u_j''$  are the residual scores for the level-1 and level-2 estimates, respectively. The final estimate of variance explained could then be derived from combining Equations 18 and 21:

$$1 - \frac{\sigma_{error}^2}{\sigma_{total}^2}. \quad (23)$$

The strength of a measure such as this is twofold. First, it puts what would normally be a complicated interpretation of a model into a palatable form for the less-informed researcher who might be reading an HLM analysis. Although it often seems that the goal of many statistical concepts is to confuse the graduate student (e.g., the multiplicity of effect sizes currently available for use), in doing so, we only confuse the future researcher, and likewise, future research.

Second, it allows the researcher, outside of the Akaike Information Criterion *AIC* (Akaike, 1987) or the Bayesian Information Criterion *BIC* (Schwarz, 1978), with a single statistic to interpret just how well a model is performing. Since the goal of most research is to find variables that fully describe the variation in the dependent variable, a measure like this could potentially prove very useful in helping the researcher make judgments about the effectiveness of an HLM model.

#### *Group Initiated $R^2$ Based on Weighted Least Squares*

In addition to alternative ways of computing an effect size mentioned above, another type of effect size can be conceived through maximum likelihood methods. In a typical multilevel

ANOVA, the grand estimate for the slope coefficient is simply the weighted least squares estimator (or maximum likelihood estimate)  $\gamma_{00}$  where:

$$\hat{\gamma}_{00} = \sum \Delta_j^{-1} \bar{Y}_{\bullet j} / \sum \Delta_j^{-1} , \quad (24)$$

and  $\Delta_j$  is the sum of the two variance components  $\text{Var}(u_{0j})$  and  $\text{Var}(\bar{e}_{\bullet j})$  (see Raudenbush & Bryk, 2002, p. 40 for further discussion). Put simply, the grand estimate for the mean of all of the groups ( $\gamma_{00}$ ) is the sum of all of the group means ( $\bar{Y}_{\bullet j}$ ) after applying the precision parameter ( $\Delta_j^{-1}$ ) and then dividing by the sum of the precision parameters. The effect of these precision parameters on the grand estimate is to apply more weight to the groups that are measured with more precision (c.f., more level-1 units). This formula for grand estimates also could be applied to the idea of an  $R^2$  effect size measure.

In typical OLS regression, the multiple  $R^2$  can be expressed as:

$$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} , \quad (25)$$

where  $y_i$  is any given individual's score on the dependent variable,  $\hat{y}_i$  is that individual's predicted score from the linear regression equation, and  $\bar{y}$  is the mean of all individuals on the dependent variable. For any given group within a set of level-2 units, Equation 25 could be considered the mathematical equivalent of:

$$R_j^2 = 1 - \frac{1}{n_j} \sum_{i=1}^{n_j} \frac{(\hat{y}_{ij} - y_{ij})^2}{\sigma_{ij}^2} , \quad (26)$$

where  $n_j$  is the number of people in group  $j$  and  $\sigma_{ij}^2$  is the variance of the individuals in group  $j$ .

For simplification purposes, we will define the latter part of Equation 26 as being an error term

corresponding to a normalized error for a given group. In representing this with the term  $E_i$ , Equation 26 can be thought of as:

$$R_j^2 = 1 - \frac{1}{n_j} \sum_{i=1}^{n_j} E_i . \quad (27)$$

This would mean that the total weighted least squares normalized error for all groups could be thought of as:

$$E_j = \sum_{j=1}^J p(s_j) E_i , \quad (28)$$

where  $p(s_j)$  is the probability of group  $j$  existing given the sample of all  $j$  groups such that:

$$p(s_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \cdot \exp\left[-\frac{(u'_j)^2}{2\sigma_j^2}\right] , \quad (29)$$

with  $u'_j$  being the residual score for group  $j$  around the mean of all groups and  $\sigma_j^2$  the variance of  $j$  groups' means around the grand mean of the dependent variable.

In performing analyses that are not multilevel in nature, we could simply compute the  $R^2$  for each group, and then average these values across all groups to obtain an average group  $R^2$  based on the sample of groups that we drew from the greater population of all level-2 units. As was previously stated, this makes little sense, however, since in multilevel modeling we are producing an equation for each second-level group based on weighted least squares estimators. It seems appropriate, then, to produce an entire model  $R^2$  that is also weighted for the probability of the group from which the estimate was drawn. In expanding Equation 27 to include all groups and also reflect the need to use a weighted estimator,  $R_T^2$  could be thought of as:

$$R_T^2 = 1 - \frac{1}{\sum_{j=1}^J p(s_j) n_j} E_j , \quad (30)$$

where  $E_j$  is the total weighted least squares normalized error for all groups from Equation 28.

We can further deduce that  $R_T^2$

$$\begin{aligned}
&= 1 - \frac{1}{\sum_{j=1}^J p(s_j) n_j} \sum_{j=1}^J p(s_j) E_j, \\
&= 1 - \frac{1}{\sum_{j=1}^J p(s_j) n_j} \sum_{j=1}^J \left( p(s_j) \sum_{i=1}^{n_j} E_i \right), \\
&= \frac{\sum_{j=1}^J p(s_j) n_j - \sum_{j=1}^J \left( p(s_j) \sum_{i=1}^{n_j} E_i \right)}{\sum_{j=1}^J p(s_j) n_j}, \\
&= \frac{\sum_{j=1}^J p(s_j) n_j \left( 1 - \frac{1}{n_j} \sum_{i=1}^{n_j} E_i \right)}{\sum_{j=1}^J p(s_j) n_j}.
\end{aligned}$$

And since we already have defined  $R_j^2$  in Equation 26, we can then interpret:

$$R_T^2 = \frac{\sum_{j=1}^J n_j \cdot p(s_j) \cdot R_j^2}{\sum_{j=1}^J n_j \cdot p(s_j)}, \quad (31)$$

which, theoretically, is simply the weighted least squares average of all of the  $R^2$  values from each group. What is present in Equation 31 is a solution that will produce estimates similar in interpretation to OLS  $R^2$  measures. Of the three possible new measures presented here, Equation 31 is considerably more appropriate than the others, since it honors both the nesting structure of the data and the fact that the model was estimated through weighted least squares estimates.

## Discussion

Snijders and Bosker (1999) presented good arguments for instances when Equations 6 and 7 produce results that yield negative values for  $R^2$  in multilevel models. This paper has made an attempt to provide alternative effect size statistics that will not produce negative values for the two-level linear model. While it is sometimes helpful to be able to know the variance accounted for at each level of the HLM model, the language with which researchers must refer to these estimates is, at best, confusing to the non-HLM minded researcher. With the further encouragement from editors to begin reporting effect sizes in all research, it is becoming more necessary for researchers using HLM to be able to explain their results in a way that is common with other statistical methods. Although results from a multilevel model probably will need further explanation, it is hoped that the continued development of these models will help in their proliferation.

There is a caution, however, in making these models more accessible. Just because a researcher has the software and programming skills to utilize complicated techniques does not mean that that technique is warranted. With the growth of a likewise complicated field of statistics, multilevel modeling, Goldstein (1995) voiced similar concerns:

There is a danger, and this paper reminds us of it, that multilevel modeling will become so fashionable that its use will be a requirement of journal editors, or even worse, that the mere fact of having fitted a multilevel model will become a certificate of statistical probity. That would be a great pity. These models are as good as the data they fit; they are powerful tools, not universal panaceas. (p. 202)

It is our sincere hope that developing HLM as a more user-friendly field of statistics will improve its utilization and interpretation, but along with this must come responsibility in evaluating such models.

## References

- Akaike, H. (1987). Factor analysis and the AIC. *Psychometrika*, *52*, 317-332.
- American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.
- Cohen, J. (1994). The earth is round ( $p < .05$ ). *American Psychologist*, *49*, 997-1003.
- Goldstein, H. (1995). Hierarchical data modeling in the social sciences. *Journal of Educational and Behavioral Statistics*, *20*, 201-204.
- Harlow, L. L., Muliak, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.
- Hox, J. (1995). *Applied multilevel analysis*. Amsterdam: TT-Publikaties.
- Hox, J. (2002). *Multilevel Analysis: Techniques and Applications*. Mahwah, NJ: Erlbaum.
- Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement*, *62*(2), 227-240.
- Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746-759.
- Knapp, T. R., & Sawilowsky, S. S. (2001a). Constructive criticisms of methodological and editorial practices. *The Journal of Experimental Education*, *70*, 65-79.
- Knapp, T. R., & Sawilowsky, S. S. (2001b). Strong arguments: Rejoinder to Thompson. *The Journal of Experimental Education*, *70*, 94-95.
- Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Roberts, J. K. (2004). An introductory primer on multilevel and hierarchical linear modeling. *Learning Disabilities: A Contemporary Journal*, *2*(1), 30-38.
- Roberts, J. K. (2002). The importance of intraclass correlation in multilevel and hierarchical linear modeling designs. *Multiple Linear Regression Viewpoints*, *28*(2), 19-31.
- Roberts, J. K., & Henson, R. K. (2002). Correction for bias in estimating effect sizes. *Educational and Psychological Measurement*, *62*(2), 241-253.
- Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods*, *1*, 115-129.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461-464.
- Shaver, J. (1985). Chance and nonsense. *Phi Delta Kappan*, *67*, 57-60.
- Snijders, T., & Bosker, R. (1999). *Multilevel analysis*. Thousand Oaks, CA: Sage.
- Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, *25*(2), 26-30.
- Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education*, *70*, 80-93.
- Wilkinson, L. & American Psychological Association Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanation. *American Psychologist*, *54*, 594-604. [reprint available through the APA Home Page: <http://www.apa.org/journals/amp/amp548594.html>]

Table 1

*Illustration of negative variance with the addition of a level-1 predictor*

Model Formula	Estimate	
	$\sigma_e^2$	$\sigma_{u0}^2$
M0: $science_{ij} = \gamma_{00} + u_{0j} + e_{ij}$	1.979	23.923
M1: $science_{ij} = \gamma_{00} + \gamma_{10}(ses) + u_{0j} + e_{ij}$	0.651	80.890

Figure 1

*Graphical representation of correlation between  $y$  and predictor variables.*

