Florida Journal of Educational Research, 2004, Vol. 42, pp. 1 - 21

A Cautionary Note on Shrinkage Estimates of School and Teacher Effects

Richard L. Tate

Florida State University

Abstract

Increasing use of "shrinkage" estimates of school and teacher effects on student achievement in educational accountability programs has been associated with arguments that multilevel models are more appropriate for the hierarchical structure of the school situation. Such estimates are usually presented as statistically optimal in that they minimize the mean square error of the estimates, a desirable property achieved by intentional introduction of a bias into the effect estimate. There is little evidence that those designing accountability programs are aware of the possible problematic nature of the differential bias associated with shrinkage estimates. In particular, intuitive rankings of school or teacher effects that are based on observed achievement means can under some circumstances be dramatically changed when shrinkage estimates are used. It is argued that all stakeholders in educational accountability programs should be aware of and agree to this feature of a system based on shrinkage estimates.

The design of any performance-based school or teacher accountability program requires a decision of how measures of student performance should be incorporated. Although a survey of the many methods that have been proposed for this purpose is well beyond the scope of this brief cautionary note, it is important to note that many have used *model-based estimates* of average student performance. Such estimates are based on the formal specification of a model of student achievement or student gain, a model that allows the rigorous determination of the precision of the estimates and, if desired, the incorporation of student and school demographic variables to provide a "value-added" assessment. Such model-based approaches have long been used in state-wide assessments (see, e.g., Tate, 2001, for a comparative study based on Florida data), and are also being proposed to address problems of reliability and validity associated with the current non-model-based implementation of the *No Child Left Behind* legislation (e.g., Thum, 2003).

In recent years, model-based "shrinkage estimates" of school and teacher effects have been increasingly proposed for use in educational accountability programs (e.g., Phillips & Adcock, 1997, Sanders, Saxton, & Horn, 1997, and Webster & Mendro, 1997). This approach to estimation is permitted by a decision to consider schools or teachers to be random (rather than fixed) effect factors in the statistical model underlying effect estimation. A random effects factor assumes that the schools or teachers are a random sample from a population of interest, with the associated analyses providing estimates of the mean and variance of the effects in the population and shrinkage estimates of the individual effects. A general family of statistical procedures that allows the inclusion of random effect factors is known by various names, including hierarchical linear models (e.g., Bryk & Raudenbush, 1992), mixed effects models (e.g., McLean, Sanders, & Stroup, 1991), multilevel models (e.g., Goldstein, 1987, 1995), and random coefficient models (e.g., Longford, 1993). Discussions and illustrations of the estimation of school and/or teacher effects with such models (to be referred to here as multilevel models) are given, for example, in Aitkin and Longford (1986), Bryk and Raudenbush (1989), Goldstein (1983, 1984, 1997), Longford (1985), Pituch (1999), Raudenbush and Bryk (1986, 1989), Raudenbush and Wilms (1995), Sanders, and Horn (1994), and Wilms and Raudenbush (1989).

A shrinkage estimate of the effect of, say, an individual teacher can be viewed as an optimal combination of two sources of information, the information available for that specific teacher and information about all teachers being evaluated. Assume, for a simple example, that one goal of an accountability program is to rank all teachers in a district with respect to the average achievement of the students in their classes at the end of the school year. (The modeling approach discussed here can also be applied to other student outcomes, such as attitude and attendance.). A ranking based on shrinkage estimates would, for each teacher, combine the observed student achievement mean for that teacher with the overall average of the student means for all of the teachers in the district. The weight placed on the observed mean in this combination would depend in part on the amount of information available for the individual teacher. If an individual teacher's mean were based on a large number of students, the weight on the observed mean would be large and the resulting shrinkage estimate would be not be much different from the observed mean. On the other hand, if the class size for a teacher were very small, the shrinkage estimate would shrink the observed mean toward the grand mean of all of the teachers.

The possibility of shrinkage estimates of individual effects is almost always presented as an attractive feature of treating schools or teachers as random effects in multilevel models. The shrinkage estimator is optimal from one statistical perspective because it minimizes the expected mean squared error of estimation (MSE). (The MSE is comprised of two components, one due to any systematic estimation bias and one reflecting random variation about the expected value of the estimator.) This minimum MSE property results from the introduction of statistical bias in the estimation, resulting from the shrinkage, to suppress the random component of the MSE. As a result, shrinkage estimation is often viewed as a possible solution to the problem of very unstable estimates of the effects for, say, teachers with very small classes. This is stated in the literature in various ways. For example Goldstein (1997), in considering the estimation of school effects, notes that "The shrinkage estimates therefore are 'conservative,' in the sense that where there is little information in any one school (i.e., few students) the estimate is close to the average over all schools" (p. 380), and Phillips and Adcock (1996) state that a multilevel solution provides “more stable estimates in smaller schools” (p. 4).

This cautionary note is motivated by the concern that little attention is currently given in the applied literature (or in the writings of those responsible for accountability systems) to the possible negative consequences of the estimation bias that is associated with shrinkage estimates. (Comments by Bryk and Raudenbush [1992, p. 129] represent one of the few exceptions to this lack of attention.) Discussions of shrinkage estimates in the statistical literature usually make it clear that the amount of shrinkage associated with the estimated effect of any school or teacher depends in part on the number of students in the school or class, as explained above. However, there is little evidence that those involved in operational accountability programs fully appreciate the implications of such differential bias in the presence of variation of school or class size. The purpose of this cautionary note is to review the nature of the differential bias associated with shrinkage estimates of school or teacher effects and to illustrate how the resulting rankings can reverse common sense rankings based on observed achievement means.

In order to establish the meaning of some basic terms and concepts, it will be necessary to provide below a brief mathematical description of shrinkage estimates. First, the simplest goal of the ranking of class or school means will be addressed with “unconditional” shrinkage estimates in which no attempt is made to control for background variables. Possible reversals of intuitive teacher or school rankings based on observed means due to the use of shrinkage estimates are then illustrated. Finally, the same concerns are briefly considered for the more realistic situation in which there is a desire to control for background variables using conditional" shrinkage estimates.

A Brief Review of Unconditional Shrinkage Estimates

To explain shrinkage estimates, consider a simple multilevel model for teachers (schools will be considered later) in which there is no attempt to mathematically control for background variables. Following the presentation in Bryk and Raudenbush (1992, pp. 17, 39-40), the model for the achievement for the i^{th} student in the class for the j^{th} teacher (Y_{ij}) is given by

$$Y_{ij} = \beta_{0j} + r_{ij}$$

where β_{0j} is the mean student achievement for the j^{th} teacher. The residual, denoted r_{ij} and representing the random student effect defined as the deviation of the achievement for the i^{th} student from the teacher mean, is assumed to be distributed normally with a mean of zero and variance of σ^2. The model for the teacher mean, β_{0j}, is

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

where γ_{00} is the average of the teacher means in the population of teachers (or "grand mean" for short) and the residual u_{0j}, representing the teacher effect, is assumed to be distributed normally with a mean of zero and variance of τ_{00}. The multilevel model can also be represented in a single equation formulation, more common in the mixed effects model statistical literature, by substituting the second equation into the first to obtain

$$Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}$$

This is simply a one-way random effects ANOVA model.

Consider three possible ways to estimate the individual teacher mean, β_{0j}. First, one may simply compute the average of Y_{ij}, denoted $\bar{Y}_{\bullet j}$, for all students in the class of the j^{th} teacher. This observed mean, often called the ordinary least squares (OLS) estimate, is an unbiased estimate of the true teacher mean with an error variance of $V_j = \sigma^2 / n_j$ where n_j is the class size. The error variance of this OLS estimate can be very large for small classes. A second possibility is to use a common estimator for each teacher mean that is equal to $\hat{\gamma}_{00}$, the estimated average of the population of teacher means. This estimate is positively biased for all teachers having true means below the population average and negatively biased for all with true means above the population mean. The random portion of the associated estimation error, though, is usually very small because it is based on the total number of students for all of the teachers.

A third option for estimating the teacher mean is the shrinkage estimator (also known as an Empirical Bayes estimator), β_{0j}^{*}, defined as the following weighted composite of the above two estimates

$$\beta_{0j}^{*} = \lambda_j \bar{Y}_{\bullet j} + \left(1 - \lambda_j\right)\hat{\gamma}_{00}$$

where $\lambda_j$is the reliability of $\bar{Y}_{\bullet j}$ as a measure of β_{0j}. The reliability parameter is defined as the proportion of the variance of the observed class mean that is due to the true variance of the class means, i.e., the reliability λ_j is defined as $\tau_{00} / \left(\tau_{00} + V_j\right)$. Rearrangement of this equation indicates that the proportion shrinkage of the deviation $\bar{Y}_{\bullet j} - \hat{\gamma}_{00}$ $\left(\text{i.e.,} \left(\bar{Y}_{\bullet j} - \beta_{0j}^{*}\right) / \left(\bar{Y}_{\bullet j} - \hat{\gamma}_{00}\right)\right)$ is simply 1 - λ_j. Thus, when the reliability of the estimate based on the data from the teacher's class is very high, there is little shrinkage and the shrinkage estimate will be approximately equal to $\bar{Y}_{\bullet j}$. In contrast, a low reliability with associated large shrinkage results in the estimate being "shrunk" towards the grand mean based on all the teachers, $\hat{\gamma}_{00}$. Because of this shrinkage, the Empirical Bayes estimator is biased towards the grand mean, γ_{00}, with a negative bias for teachers with observed class means above the grand mean and a positive bias for those with means below the grand mean.

In practice, it may be recommended that teachers be evaluated by ranking them with respect to the shrinkage estimates of the class means.

The precision of each shrinkage estimate in the ranking could be represented with an error band based on the associated standard error (see, e.g., Equation 3.38 in Bryk and Raudenbush, 1992). Comparisons of teachers with nonoverlapping error bands would be viewed as trustworthy.

Differential Shrinkage: A Potentially Problematic Feature

Consider more closely the reliability λ_j that determines the weights in the Empirical Bayes estimate and the resulting shrinkage of $\bar{Y}_{\bullet j}$ towards $\hat{\gamma}_{00}$. Defining the variance ratio ω as τ_{00}/σ^2 (the ratio of the true variance of the teacher means to the within-class variance of individual achievement) and recalling that the error variance, V_j, of $\bar{Y}_{\bullet j}$ is equal to σ^2/n_j, the reliability can be expressed as

$$\lambda_j = \frac{\omega}{\omega + \frac{1}{n_j}}$$

A graphical representation of this relationship is shown in Figure 1. It is seen that the reliability decreases (and the resulting shrinkage increases) as the class size, n_j, decreases, holding constant the variance ratio.

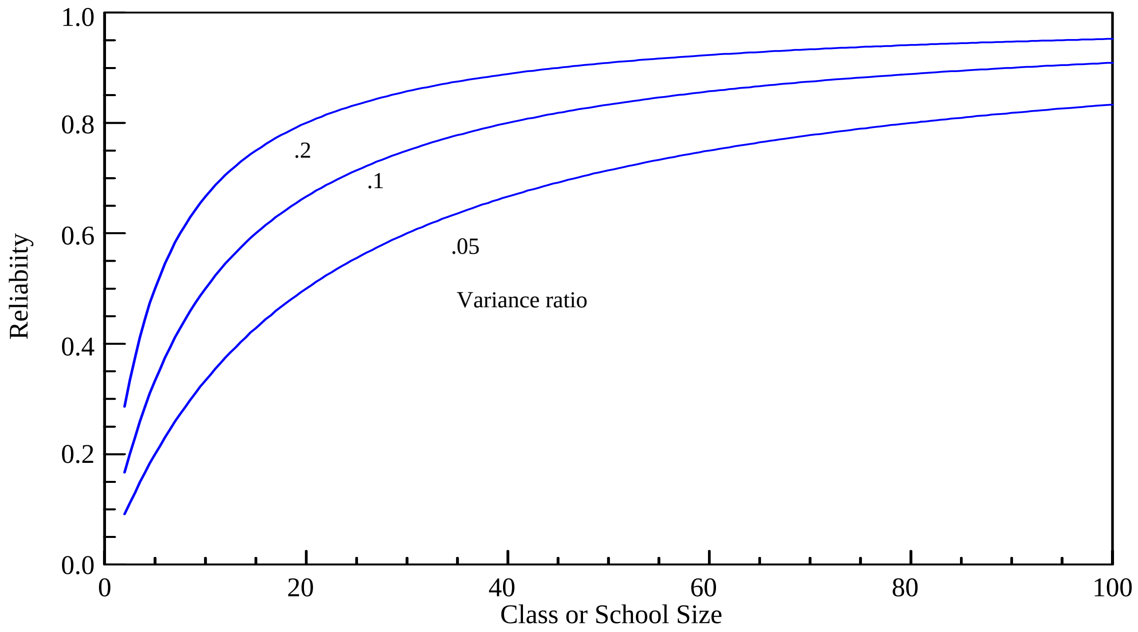
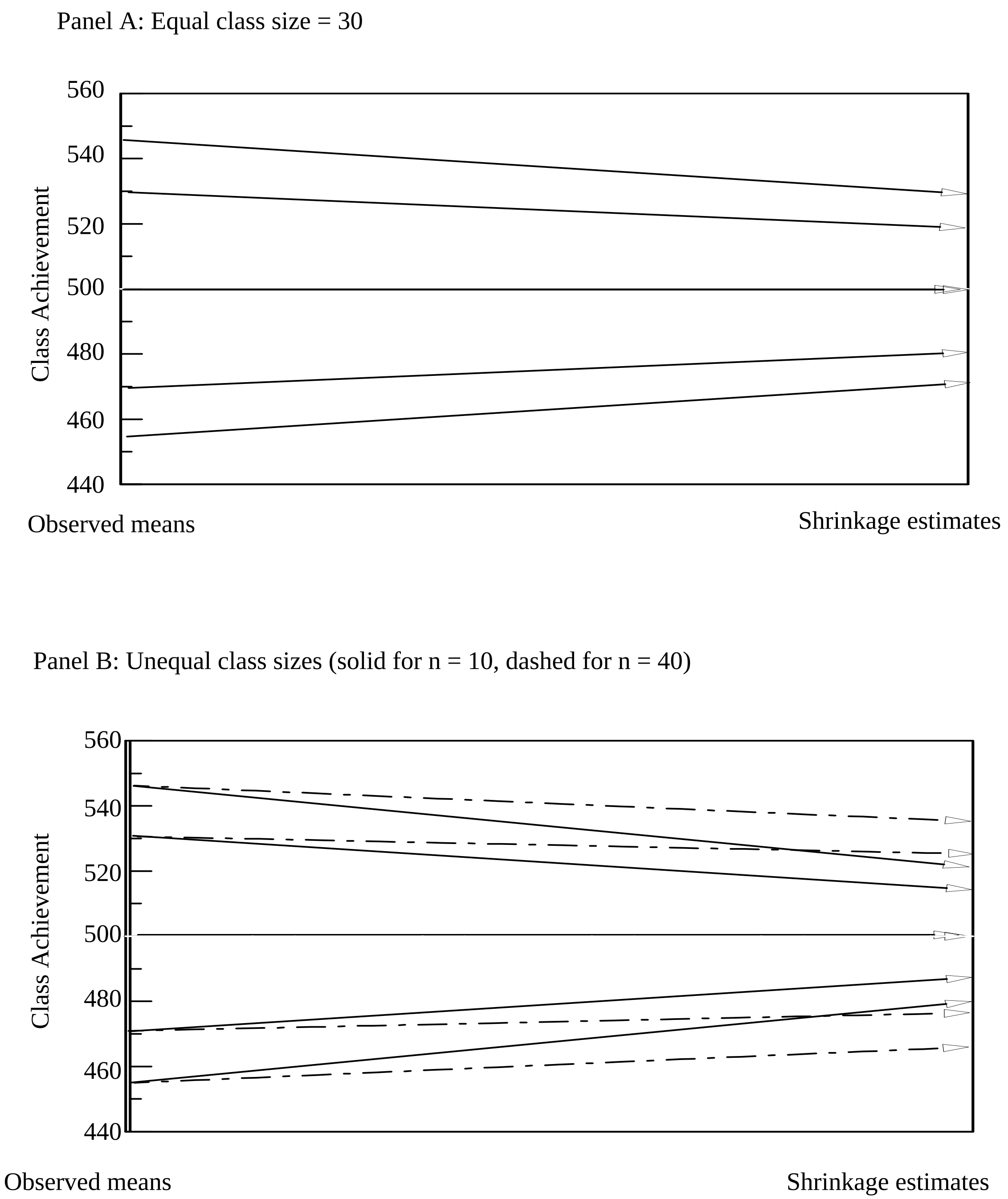


Figure 1. Observed mean reliability as a function of class or school size.

When the class sizes for all teachers are approximately equal, the bias introduced by the shrinkage estimator would not be a problem when making a relative comparison of teachers. To illustrate, consider a situation in which the test has been scaled in the population of students to have a mean of 500 and a standard deviation of 100. Assume further that the variance ratio ω is equal to 0.1. Since the variance of Y in the random effect ANOVA model has a variance of $\tau_{00} + \sigma^2$, it can be determined that the true variance of the teacher effects is equal to 909.1 with a standard deviation of 30.2. For a class size of 20, the distribution of observed class means would have a variance of $\tau_{00} + \sigma^2/20 =$ 1363.6 and a standard deviation of 36.9. Consider five different teachers ranked from highest to lowest with respect to their class achievement means, with means of 545, 530, 500, 470, and 455, respectively. (These observed means correspond to +1.2, +0.8, 0.0, -0.8, and -1.2 standard deviations about 500 in the distribution of class means). If all of these teachers have the same class size of 20, the reliability of $\bar{Y}_{\bullet j}$ will be 0.667 and the resulting shrinkage estimates of the five teacher means will be equal to 530, 520, 500, 480, and 470, respectively (assuming that the estimated grand mean is 500). These results are represented in Panel a of Figure 2. Although the apparent variability of the estimated teacher means has been reduced, the shrinkage has not resulted in any change in the original ranking of the teachers based on the observed class means.

When the class sizes of teachers are different, though, the situation may become more problematic. Consider another example in which all conditions are identical to those stated above except that now there are 10 teachers, five with classes of 10 students each and five with classes of 40 students. The reliability λ_j for the teachers with the small classes is 0.5 while that for the teachers with large classes is 0.8. As a result, the $\bar{Y}_{\bullet j}$ values for the small classes are shrunk farther towards the mean of 500 than the $\bar{Y}_{\bullet j}$ values for the large classes. Assuming the same $\bar{Y}_{\bullet j}$ values of 545, 530, 500, 470, and 455, the resulting shrinkage estimates for the teachers with the small classes are 522, 515, 500, 485, and 478, respectively.



Panel A: Equal class size = 30

Panel B: Unequal class sizes (solid for n = 10, dashed for n = 40)

Figure 2. A comparison of shrinkage estimates for equal and unequal class sizes.

In contrast, for the same $\overline{Y}_{\bullet j}$ values, the shrinkage estimates for the teachers with large classes are 536, 524, 500, 476, and 464, respectively. As shown in Panel b of Figure 2, the variable reliability over teachers due to different class sizes results in differential shrinkage, a shrinkage that results in large changes in the original rankings of the teachers. For example, the teacher with a class of 10 students having an observed class mean of 545 is ranked lower on the shrinkage estimates than the teacher with 40 students having a class mean of 530.

Reversals of original teacher rankings (i.e., those based on observed means) operate in the opposite direction for teachers with class means below the teacher effect average. That is, the $\overline{Y}_{\bullet j}$ values are shrunk more in the positive direction for teachers with smaller classes. For example, Panel b of Figure 2 indicates that a teacher with a class size of 10 and a $\overline{Y}_{\bullet j}$ value of 455 will, based on shrinkage estimation, rank higher than a teacher with a class size of 40 and a $\overline{Y}_{\bullet j}$ value of 470. In sum, teachers with $\overline{Y}_{\bullet j}$ values above the teacher average are benefited by having large classes, while those with $\overline{Y}_{\bullet j}$ values below the teacher average are benefited by having small classes.

The severity of this problem of rank reversal depends on the extent of the variation of reliability over teachers, which in turn depends on the range of class sizes and the variance ratio (see [5] and Figure 1). Obviously, as the range of class sizes increases, the problem becomes more serious.

The effect of the variance ratio, less obvious from Figure 1, is represented more clearly in Figure 3. The *difference* in reliabilities resulting from two different group sizes is shown as a function of the variance ratio. Considering for now the curve for group (class) sizes ranging from 10 to 40, it is seen that the reliability change is relatively large (say, 0.2 or larger) for variance ratios ranging from approximately 0.01 to 0.3, a range covering likely actual values. In other words, when the range of class sizes is

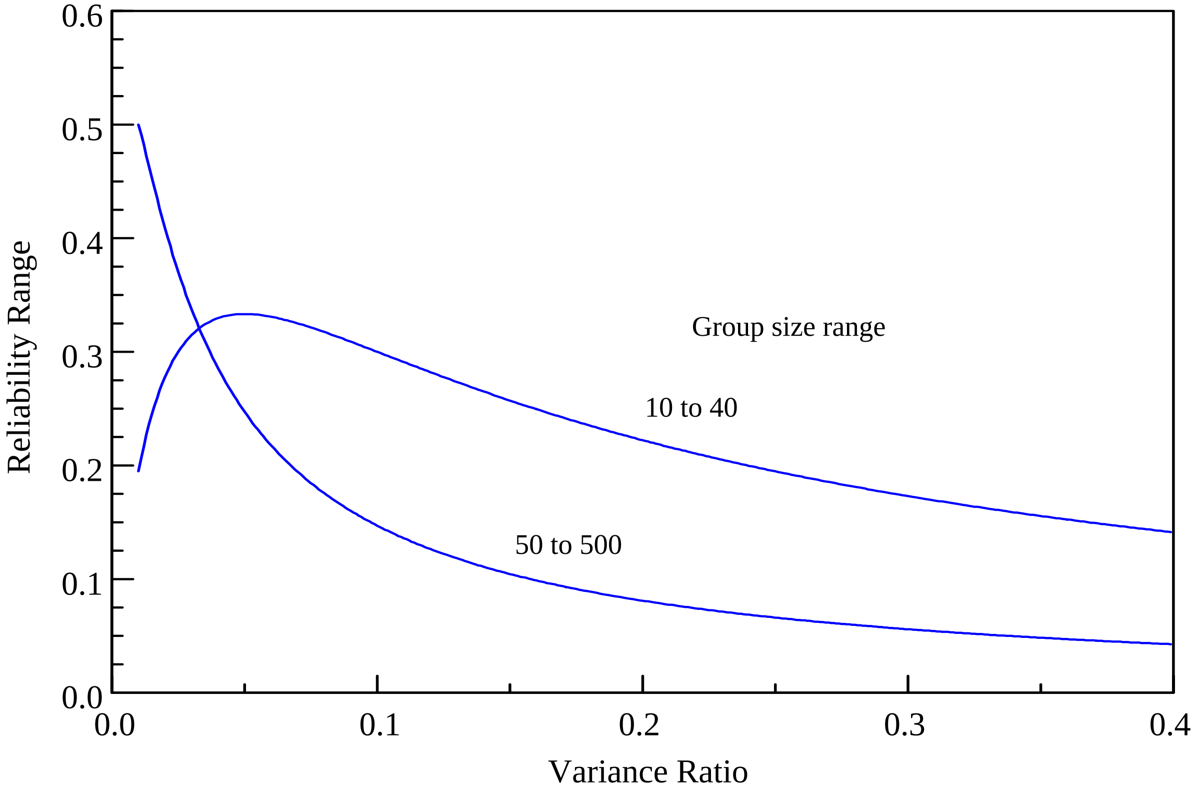


Figure 3. Reliability change associated with two group size ranges as a function of the variance ratio.

relatively large, a significant problem of reversal of teacher rankings would be expected over most reasonable values of the variance ratio.

School Rankings

The example to this point has illustrated the problem of differential shrinkage for teacher rankings. Would the same problem be present in attempts to rank schools based on student achievement? Consideration of Figure 1 may, at first glance, suggest that the problem would be minimal. For the larger numbers of students involved in the determination of school means, the reliability curves are much flatter. On the other hand, the typical variation of school size (say, 50 to 500 students) would be much larger than the usual range of class size. The net effect of the change in both of these two factors is represented by the second curve in Figure 3. For the range of school size of 50 to 500, the associated change in reliability increases with decreasing variance ratio. When one considers that the variance ratio tends to decrease with higher levels of aggregation (the larger the group size, the more individual differences will tend to cancel), it would not be surprising to find variance ratios of 0.1 and lower at the school level. At these lower values, Figure 3 indicates that appreciable reliability changes of 0.15 and higher would be found, again producing problematic reversals in school rankings.

Conditional Shrinkage

There are numerous determinants of student achievement, many of which are not under control of the school or teacher (see, for example, Berk, 1988; Haertel, 1986). It is often argued that a fair evaluation requires that all determinants not under control of the school or teacher be taken into consideration. An oft-proposed alternative to the qualitative consideration of these factors by personnel evaluators is based on the attempt to mathematically model, and thereby control, important determinants of student achievement not under control of the schools or teachers. This model-based approach is often presented as a more rational approach to the difficult task of causal attribution involved in evaluations of schools and teachers. The resulting “conditional” models also provide shrinkage estimates of the teacher or school effects, controlling for the included background variables. Still following the presentation of Bryk and Raudenbush (1992, pp. 21-22, 42-44), consider a simple multilevel model for students nested in schools in which the model at the student level is identical to the student level model considered previously, but the school-level model is now expanded to include average student family SES for the school, denoted *W*, i.e.,

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j}$$

where the residual, u_{0j}, is the effect for school *j* defined as the difference between the observed mean, $\bar{Y}_{\bullet j}$, and the value predicted by the school's value of *W*. This residual is often assumed to be distributed normally with mean zero and variance τ_{00}. Schools would then be ranked with respect to estimates of the residuals, u_{0j}.

The unbiased OLS estimate of the school residual would be

$$\hat{u}_{0j} = \bar{Y}_{\bullet j} - \left(\hat{\gamma}_{00} + \hat{\gamma}_{01} W_j\right)$$

where the term in parentheses is the predicted value of the school mean based on the school's value of *W*. The shrinkage estimate of the same residual is

$$u_{0j}^{*} = \lambda_j \hat{u}_{0j}$$

where the reliability λ_j is defined as before. This equation indicates that the OLS estimate of the school effect is shrunk towards zero. As with the unconditional shrinkage discussed above, varying school sizes will result in differential shrinkage which will produce reversals of a school ranking based on the unbiased OLS estimates of school effects. The severity of the problem would tend to be greater with conditional shrinkage. The variance of the residuals, τ_{00}, will be reduced when some of the variability of the school means is explained by the school variable, *W*. The resulting decrease in the variance ratio, $\omega = \tau_{00} / \sigma^2$, will then produce a larger variation of reliability for the same range of school sizes. For example, it is seen from Figure 3 that a range of school size from 50 to 500 would result in a reliability range of 0.15 when the variance ratio is 0.1. If addition of the control variable *W* explains half of the true variance of the school means, the variance ratio for the conditional model would then be 0.05, producing a range of reliability of 0.25 for the same school size range.

Summary

The shrinkage estimates of teacher and school effects provided by multilevel models offer the important advantage of minimizing the expected mean squared error of the estimates. This desirable property is attained by intentionally introducing estimation bias by shrinking the OLS estimates towards the grand mean (for unconditional estimates) or a predicted value based on group level information (for conditional estimates). It is important to be aware that any variation of group (class or school) size results in a corresponding variation in the amount of shrinkage used in obtaining the estimated school effects. This differential shrinkage can result in final teacher or school rankings that are different from intuitive rankings based on either observed group means or on unbiased estimated effects, differences that may be viewed as unfair by those being evaluated. The designers of any assessment considering the use of shrinkage estimates should determine the magnitude of this problem under the circumstances of interest. All stakeholders (including of course those being evaluated) should be aware of and agree to this feature of shrinkage estimators.

Notes

[1] In the mixed effects model literature, estimation bias is sometimes defined differently (e.g., Robinson, 1991). Shrinkage estimates are often called unbiased estimates in this literature because the expected value of the estimated random effects over the population of groups is equal to the mean of the true random effects. In this literature, these shrinkage estimates are sometimes called BLUP for "Best Linear Unbiased Predictors," with the "unbiased" portion of the label reflecting this meaning of bias. However, based on the current definition of bias reflecting the natural concern of evaluation stakeholders, the shrinkage estimates are biased. That is, given a true random effect for a school or teacher, there is a difference between the expected value of the shrinkage estimates of that effect and the true value.

References

Aitkin, M., & Longford, N. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society, Series A, 149*, 1-43.

Bryk, A.S., & Raudenbush, S.W. (1989). Toward a more appropriate conceptualization of research of school effects: A three-level hierarchical linear model. In R.D. Bock, (Ed.), *Multilevel analysis of educational data*. San Diego: Academic Press (pp. 205-234).

Bryk, A.S., & Raudenbush, S.W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, Ca: Sage.

Glass, G.V. (1990). Using student test scores to evaluate teachers. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation* (pp. 229-240). Newbury Park, CA: Sage.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement, 20*, 369-377.

Goldstein, H. (1984). The methodology of school comparisons. *Oxford Review of Education, 10*, 69-74.

Goldstein, H. (1987). *Multilevel models in educational and social research*. New York: Oxford Press.

Goldstein, H. (1995). *Multilevel statistical models*. London: Edward Arnold.

Goldstein, H. (1997). Methods in school effectiveness research. *School Effectiveness and School Improvement, 8*, 369-395.

Haertel, E. (1986). The valid use of student achievement measures for teacher evaluation. *Educational Evaluation and Policy Analysis, 8*, 45-60.

Longford, N.T. (1985). Mixed linear models and an application to school effectiveness. *Computational Statistics Quarterly, 2*, 109-117.

Longford, N.T. (1993). *Random coefficient models*. Oxford: Clarendon Press.

McLean, R.A., Sanders, W.L., & Stroup, W.W. (1991). A unified approach to mixed linear models. *The American Statistician, 45*, 54-63.

Phillips, G.W., & Adcock, E.P. (1996). *Practical applications of hierarchical linear models to district evaluations*. Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Pituch, K.A. (1999). Describing school effects with residual terms: Modeling the interaction between school practice and student background. *Evaluation Review, 23*, 190-211.

Raudenbush, S.W. (1988). Educational applications of hierarchical linear models. *Journal of Educational Statistics, 13*, 85-116.

Raudenbush, S.W., & Bryk, A.S. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59*, 1-17.

Raudenbush, S.W., & Bryk, A.S. (1989). Quantitative models for estimating teacher and school effectiveness. In R.D. Bock, (Ed.), *Multilevel analysis of educational data*. San Diego: Academic Press (pp. 205-234).

Raudenbush, S.W., & Wilms, J.D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics, 20*, 307-336.

Robinson, G.K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science, 6*, 15-51.

Sanders, W.L., & Horn, S.P. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*, 299-311.

Sanders, W.L., Saxton, A.M., & Horn, S.P. (1997). The Tennessee Value-Added Assessment System (TVAAS): A quantitative, outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press, Inc.

Tate, R. L. (2001). *Annual learning gains/value-added methods for assessing student achievement, teacher effectiveness, and school accountability*. Florida Department of Education, Testing and Evaluation Services: Tallahassee, FL.

Thum, Y. M. (2003). *No Child Left Behind: Methodological challenges & recommendations for measuring Adequate Yearly Progress*. Center for the Study of Evaluation, University of California, Los Angeles.

Webster, W.J., & Mendro, R.L. (1997). The Dallas Value-Added Accountability System. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 81-99). Thousand Oaks, CA: Corwin Press, Inc.

Wilms, J.D., & Raudenbush, S.W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement. 26*, 209-232.