SYNTHESIZING RESULTS FROM THE TRIAL STATE ASSESSMENT

Stephen W. Raudenbush, Randall P. Fotiu, Yuk Fai Cheong, and Zora M. Ziazi
Department of Counseling, Educational Psychology, and Special Education, 246 Erickson Hall
Michigan State University, East Lansing, MI 48824-1034

Key Words: Assessment, NAEP, Hierarchical linear models, Gibbs sampling

Major purposes of the Trial State Assessments (TSA) in mathematics are to compare states in terms of their mean mathematics proficiency, to assess how well sub-groups of students such as females, ethnic minorities, and socially disadvantaged students are faring in the several states (Mullis, Dossey, Owens, and Phillips, 1993), and to assess relationships between policy-relevant predictors and math proficiency. It is then possible to compare states over time using repeated cross sections of the TSA.

Of course, strong causal inferences will not be justified based on cross-sectional data, or even on the basis of repeated cross-sections, especially given the absence of a measure of prior attainment in the NAEP data. Rather, policy analyses of TSA data are designed to produce suggestive results; not to prescribe policy, but to stimulate new thinking about the sources of variation in mathematics proficiency at each level of the system, with special emphasis on the state level, about the predictors of variation at each level, about the opportunities and targets for intervention at each level, and about the plausible mechanisms for school improvement.

In pursuing these purposes a number of methodological challenges emanating from the design of the TSA immediately confront the data analyst. These include the special structure of the outcome data which, by design, are incomplete for every student; the within-state design that includes both clustering and stratification; and the problem of incorporating heterogeneity between states. In this paper, we propose and illustrate a comprehensive and broadly applicable strategy for coping with these challenges. Our strategy has two stages: a within-state analysis and a between-state analysis. The within-state analysis uses a hierarchical linear model to handle the clustered character of the sample. This analysis is replicated for each plausible value and the results pooled as recommended in Little and Schenker (1994) and Mislevy (1992) using a specialized version of the HLM program (Bryk, Raudenbush, and Congdon, 1994) originally adapted for multiple plausible values by Arnold, Kaufman, and Sedlacek (1992). The output for each state is a vector of parameter estimates and their estimated sampling variance matrix. These then provide input data for the second stage of the analysis, a Bayesian synthesis of findings across states. Taken together, the two stages have the structure of a planned "meta-analysis" (Glass, 1976) in which each state's separate analysis constitutes a "study" and the between-state analysis combines these results.

Data Analytic Challenges and Strategies

Within each state, the sampling and measurement design unique to the TSA pose challenges that are not easily addressed with standard data analysis methods. Even when those are addressed, however, one must decide how to combine information across states. Below we outline a two-stage ("within-state" - "between-state") approach to the analysis of TSA.

Within-State Analysis

Let us consider first the data yielded within each participating state. The sampling plan was designed primarily to make inferences about student math proficiency. The students within each state were selected as the result of a two-stage cluster sample with stratification at the first stage. Specifically, schools were first stratified on the basis of urbanicity, minority concentration, size, and area income and then a) schools were selected at random within strata with a probability proportional to student grade level enrollment; and b) students were systematically selected from a list of students, given a random starting point, within schools. Overall, about 100 schools per participating state were selected with approximately 25 to 30 students selected from each school. An appropriate analytic method for these data, given the goals of the study and the design, is a two-level hierarchical linear model (e.g., Aitkin and Longford, 1980; de Leeuw and Kreft, 1986; Raudenbush and Bryk, 1986; Goldstein, 1987) in which students are "level-1" observations, schools are "level-2" observations, and each student's data are weighted inversely proportional to that student's selection given the stratification. The "HLM" program of Bryk, Raudenbush, and Congdon (1994) allows incorporation of design weights at level 1 reflecting the sampling design for inferences directed primarily at student characteristics. A comprehensive coverage of the sampling design and the construction of weights are presented by Johnson, Mazzeo, and Kline (1993).

However, provision must also be made for the special character of the outcome variable used in NAEP as a result of the matrix sampling scheme in which each student was observed on only a subset of relevant items. Rather than yielding a single measured variable, NAEP produces five "plausible values" -- random draws from the posterior distribution of each student's "true" outcome

given the subset of items observed on that student (Johnson, Mazzeo, and Kline, 1993). To cope with this problem, Arnold et al. (1992) modified the HLM program to compute a separate analysis for each of the five plausible values and then to synthesize the results via an adaptation of Rubin's (1987) recommended approach to the analysis of multiply imputed data. This approach is described in detail in Little and Schenker (1994) and Mislevy (1992). This approach takes into account the extra uncertainty that arises because multiple plausible values rather than a single observed outcome were available.

Between-State Analysis

We expect that mean proficiency and "status gaps" in proficiency will vary from state to state. This heterogeneity across states is both of interest substantively and of concern methodologically. Substantively, state-to-state differences pose important questions for state and national policy-makers. Such questions may be addressed in one of two ways. First, it may be that once student background, school context, and policy-relevant predictors are controlled within states, little between-state variation will remain to be explained. Such a result would be informative and would motivate a study of how the key policy-relevant variables are distributed across states. Second, especially to the extent state-level heterogeneity persists even after controlling relevant covariates within states, it may be helpful to use state differences in income, funding, and policy as predictors of state differences in outcomes. This paper adopts the first strategy.

Methodologically, such state-level heterogeneity plays a role similar to that of between-cluster variance in that analyses that ignore such heterogeneity will often produce biased results. In particular, standard errors for effects of predictors defined on states will be negatively biased. To address these issues, we employ a Bayesian framework in which, given the predictors in the model, the state effects are viewed as exchangeable and therefore random (DeFinetti, 1964; Lindley and Smith, 1971). The variance assigned to state-level heterogeneity represents our uncertainty about the source of that heterogeneity.

Why Choose a Bayesian Synthesis Across States?

The analyses we have proposed so far involve estimation of variation between students within schools, between schools within states and between states; and the formulation of prediction models to account for such variation at each level. Formally, this structure represents a three-level hierarchy. Hence, one might contemplate the use of a now-standard three-level hierarchical linear model with estimation via maximum likelihood as a basis for the analysis. Bryk and Raudenbush (1992, Chapter 8) review applications of this model. However, there are several compelling reasons to reject this choice, and new methods of analyses appear necessary for these data.

First, although the data are very dense at level 1 (the student level with about 2500 students per state) and level 2 (the school level with about 100 schools per state), the data are comparatively sparse at level 3 (the state level with 41 states and territories). Standard applications of hierarchical linear models condition estimates of all regression coefficients and their standard errors on point estimates of the variance-covariance parameters (Raudenbush, 1988). However, level-3 variance and covariance components are likely to be estimated with moderate or poor precision (depending on the research question). In this setting, conditioning on point estimates can lead to inferential errors, especially, in our case, regarding inferences about relationships between state-level predictors and outcomes.

A more suitable analytic choice is a Bayesian analysis in which a non-informative prior distribution is specified for all state-level parameters. All inferences about regression coefficients at any level will then fully take into account the uncertainty about the unknown state-level variances and covariances. Moreover, inferences about these variances themselves will be important, since they signify the degree of state-to-state heterogeneity in the key outcomes of the study. Such inferences will be based on their posterior distributions. Such posterior distributions can be presented in a way that is readily accessible to non-technical audiences. Seltzer (1993) clearly explicates the advantages of this approach. In contrast, the maximum likelihood approach bases inferences about such variances on the large sample normal approximation to their sampling distribution. This approach will often be seriously inaccurate, especially when the number of level-3 units is small and the degree of heterogeneity at that level is modest. In these cases, the sampling distribution of the level-3 variance will tend to be highly skewed, contrary to normal theory (see Rubin, 1981, for an analogous case and a lucid discussion of this problem). The Bayes model in our case will be estimated via Gibbs sampling (e.g., Gelfand and Smith, 1990; Tanner and Wong, 1987; Seltzer, 1993).

A second compelling reason to avoid standard application of three-level hierarchical linear model is that those models typically require homogeneity of dispersion within level-3 units (or at least within subsets of level-3 units) (Bryk and Raudenbush, 1992, Chapter 8). In many applications this assumption is sensible in light of inefficiencies that arise when separate variance structures are estimated for each level-2 or level-3 unit. However, in our case, it will be more realistic to allow every state to have a unique between-school variance-covariance matrix and a unique within-school variance. Loss of efficiency will not be an important consideration because of the substantial data existing within each state. Our separate, within-state analyses allow such heterogeneity of variance.

A third reason for a novel analytic approach is that classical applications of three-level models assume random sampling at each level. However, 41 states and territories volunteering to participate in the 1992 Trial State Assessment cannot be viewed as a simple random sample of states or territories. They are better viewed as strata. However, the classical approach to stratification -- to represent stratum effects as fixed effects -- contradicts our goal of modeling variation in state outcomes. Again, the Bayesian approach with a non-informative prior seems a sensible choice. In the Bayes view, unexplained state-level heterogeneity is modeled by variances and covariances that represent the investigator's uncertainty about the processes the produce it -- rather than representing sampling errors arising from a formal sampling process.

Computations for Bayes estimation via Gibbs sampling are generally intensive. However, the main computational burden in our analysis will be the state-by-state analyses using a two-level hierarchical model with estimation via maximum likelihood. Once these computations are completed, the second-stage analysis -- that is, the between-state analysis using Bayesian estimation, will be computationally undemanding since it will be based on a sample size of 41. Thus, efficient computational methods -- via maximum likelihood -- will be used where the data are dense, that is, at levels 1 and 2. Computationally intensive methods will be used where they are most needed -- at level 3 where the data are comparatively sparse and the cost of using such methods is modest.

Illustrative Examples

Two analyses illustrate the logic of the approach we propose. The first simply estimates each state's mean math proficiency and standard error and then synthesizes these results in order to estimate the extent of state heterogeneity. The second analysis attempts to account for this heterogeneity by formulating and estimating two-level hierarchical models within each state.

Example 1: Assessing Heterogeneity in Unadjusted State Means

We first formulate within each state a two-level hierarchical model with no covariates. The output for each state is an estimated mean and its standard error, which provide input into a between-state synthesis.

Level-1 model. The level-1 units are students; within each school, the student-level outcome depends

$$Y_{ijk} = \pi_{jk} + e_{ijk}, \quad e_{ijk} \sim N(0,\sigma_k^2), \qquad (1)$$

only on the school mean according to the model where Y_{ijk} is the proficiency score for student $_i$ in school j and state k; π_{jk} is the school mean, and e_{ijk} is a random error assumed independently and normally distributed with variance σ_k^2.

Level-2 model. The level-2 units are the schools, and each school's mean is postulated to vary randomly around the state mean according to the model

$$\pi_{jk} = \beta_k + u_{jk}, \quad u_{jk} \sim N(0,\omega_k^2). \qquad (2)$$

Thus, β_k is the mean outcome for state k and u_{jk} is the random school effect assumed independently and normally distributed with variance ω_k^2.

$$Y_{ijk} = \beta_k + u_{jk} + e_{ijk}, \qquad (3)$$

Combined model. Substituting Equation 2 in Equation 1 yields the single "combined equation" which is recognizable as a one-way analysis of variance model with random school and person effects. Estimates of the two variance components, σ_k^2 and ω_k^2, incorporates variation associated with the cluster sample so that the maximum likelihood (ML) estimate of β_k and its standard error will incorporate the extra variation arising from the clustered nature of the sample.

Two modifications are needed to adapt estimation of each state's parameters given the structure of the TSA data. First, incorporation of student-level design weights assures that unequal probability of selection of sub-groups of students will not bias estimation of the state's mean proficiency. Second, the outcomes Y_{ijk} are, in fact, plausible values based on incomplete data rather than observed scores based on a full complement of item-level data. Thus, the ML estimation is replicated for each of five plausible values and the results pooled as specified in Little and Schenker (1994).

Bayesian synthesis. The output from the within-state analysis are , the ML estimate of the state mean and its estimated variance. Note that the variance estimate incorporates uncertainty associated with multiple plausible values, stratification, and clustering.

Exchangeable prior for β. The state means are assumed a priori exchangeable, implying that we have no prior knowledge about the magnitude of the means of given states. We therefore assume

$$\beta_k|\gamma \sim N(\gamma,\tau). \qquad (4)$$

Here γ represents the prior location of the state means, a kind of "national mean," though it cannot be viewed as representative of the entire US population of eighth graders; and τ represents the heterogeneity of the state means. The estimated variances V_k are assumed equal to their true values. Though this assumption cannot be true, its falsehood will have essentially no consequences on

inference given the large sample sizes within states.

Non-informative priors for γ, τ. We have essentially no knowledge about the location of the state means or their heterogeneity. We therefore assume a priori that this pair of parameters are independent and non-informative on their parameter spaces. Technically,

$$p(\gamma) = C_\gamma, \quad -\infty < \gamma < \infty \qquad (5)$$

Here the C_γ is an arbitrarily small constant. This flat prior assures that the posterior density of the parameter will be proportional to the likelihood, which is determined by the data.

In general, the variance-covariance matrix, τ, is assumed to have an inverse Wishart prior distribution given by

$$\tau \sim W^{-1}(\psi,\nu) \qquad (6)$$

where Ψ, is the precision matrix of the inverse Wishart distribution and ν is the degrees-of-freedom parameter. This prior distribution is assumed to be non-informative in its contribution to the posterior distribution as the degrees of freedom parameter, ν, approaches 0, and Ψ approaches 0.

Results. The analysis produces approximate marginal posterior distributions for β_k, k = 1,...,41 and for γ and τ.

Figure 1 gives 98% posterior credibility intervals for approximateley half of the state means, β_k. Many of these do not overlap, implying significant between-state heterogeneity. This impression is confirmed by Figure 2 (not shown) which gives the approximate marginal posterior for τ. The figure shows unmistakable evidence of heterogeneity between states (note that zero is not a plausible value for τ). However, there is considerable uncertainty about the magnitude of this heterogeneity. The outcome variable was on a scale with a mean near zero and a variance of approximately unity. The posterior mean of τ is .079, implying that 7.9 % of the variance in the outcome lies between states. However, τ values as small as .04 and as large as .12 are not improbable. Thus, it appears that from 4% to 12% of the variance in the outcome lies between states.

Example 2: Accounting for Heterogeneity in State Means

We now formulate within each state a two-level hierarchical model with covariates. The output for each state is a vector of regression coefficient estimates and their covariance matrix. Among these is the intercept, which will be the focus of the discussion here.

Level-1 model. The level-1 model relates student-level predictors to student outcomes according to the model

$$Y_{ijk} = \pi_{jk} + \sum_{p=1}^{P} \alpha_{pk} X_{pijk} + e_{ijk}, \quad e_{ijk} \sim N(0,\sigma_k^2), \qquad (7)$$

where Y_{ijk} is again the math proficiency score for student i in school j and state k; π_{jk} is the school mean adjusted for that school's means on the X variables, α_{pk} is the regression coefficient associated with each X_{pijk}, which is the student's value on the pth student-level covariate. All covariates were centered around the Michigan means, so that if π_{jk} is positive, it means that school j in state k has a higher adjusted mean than does Michigan.

The following Xs were used in the model: gender, ethnicity (indicators for Hispanic, non-Hispanic black, Asian, and Native American), national origin (indicator for born outside the US), family type (indicators for single parent and both parent with other type as the reference group), parental education (indicators for high school degree, high school plus, and bachelors degree), amount of time watching television, mobility, home literacy environment (indicators for receiving a newspaper, having more than 25 books, and subscribing to magazines), academic level (indicators for taking algebra and taking pre-algebra), and teacher preparation (teachers' years of experience teaching math, and indicators for having majored in math as an undergraduate, having majored in math education, and having an advanced degree). Thus, in all there were P = 23 student-level covariates in the model within each state, not an inordinate number given state sample sizes averaging about 2,500 students.

Level-2 model. The level-2 model relates school-level covariates to school intercepts according to the model

$$\pi_{jk} = \beta_k + \sum_{q=1}^{Q} \delta_{qk} W_{qjk} + u_{jk}, \quad u_{jk} \sim N(0,\omega_k^2). \qquad (8)$$

All W's were deviated around the Michigan mean so that β_k is the adjusted mean for state k with positive values meaning that state's adjusted mean is higher than Michigan's. Note that ω_k^2 is the residual variance between schools within state k.

The following Ws were specified in the model: school-median income, instructional dollars per pupil, percent minority, location (with indicators for rural and urban), an indicator for whether the school offers algebra, and a scale indicating the disciplinary climate of the school.

Combined model. Substituting Equation 2 in Equation 1 yields the single "combined equation"

$$Y_{ijk} = \beta_k + \sum_{q=1}^{Q} \delta_{qk} W_{qjk} + \sum_{p=1}^{P} \alpha_{pk} X_{pijk} + u_{jk} + e_{ijk}, \qquad (9)$$

which is recognizable as a random intercepts regression with random school and person effects. Estimates of the two variance components σ_k^2 and ω_k^2, incorporate variation associated with the cluster sample so that the maximum likelihood (ML) estimate of β_k and its standard error will incorporate the extra variation arising from the clustered nature of the sample.

Bayesian synthesis. The synthesis followed the same form as in the case of the unconditional model except that the input was the adjusted state mean and its variance rather than the unconditional mean.

Results. The analysis produced approximate marginal posterior distributions for β_k, k = 1,...,41 and for γ and τ. Figure 3 gives 98% posterior credibility intervals for the state adjusted means, β_k. The vast majority of these now overlap, in contrast to the unconditional case where many did not overlap, implying far less heterogeneity between states than in the unconditional case. This impression is confirmed by Figure 4 (not shown) which gives the approximate marginal posterior for the conditional τ. Although the figure shows evidence of heterogeneity between states (note that zero is not a plausible value for τ), there is every reason to believe that the magnitude of this heterogeneity is small. The posterior mean of τ is .012, implying that 1.2 % of the variance in the outcome lies between the adjusted means of the states. Moreover, the unknown value of τ is unlikely to exceed .03 or 3% of the total variance in the outcome. Thus, it appears that from .004% to 3% of the variance in the outcome lies between state adjusted means.

Comment. Considerable controversy has surrounded the question of whether and how to compare states in terms of mean proficiency. Some have argued that unadjusted means simply reward those states with the most advantaged compositions and give impoverished states and those with large number of immigrant and ethnic minority students no chance to "shine." However, the National Governing Board of NAEP has taken a strong stand against adjusted means, arguing that educators should not set up low expectations for more disadvantaged states.

A technical argument against standard approaches to adjustment is that if one controls for student demographic background without specifying explanatory variables representing policy and practice, the adjustments for background will be biased. That is, given the typical positive correlation between student social advantage and exemplary school practice, a model that omits school practice will over-estimate the effect of social advantage, thus leading to biased estimates of adjusted means.

The analysis reported above offers an alternative, namely, to formulate models that include social composition and school policy and practice. Such a model accounts for nearly all of the variation between states, as a comparison of Figures 2 and 4 shows. The key issue is then to compare states, not on their adjusted means, but, rather, on the values taken on by key policy-relevant predictors. For example, our results indicated consistent and positive effects of teacher preparation and a positive disciplinary climate. It would therefore make sense for states to examine their performance on these variables and to use such information to assess policy options for improving schooling. (Complete set of tables and figures available upon request.)

References

Arnold, C.L., Kaufman, P.D., & Sedlacek, D.S. (1992). School effects on educational achievement in mathematics and science: 1985-1986 (Research and Development Report). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

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

Bryk, A.S., Raudenbush, & Congdon, R.T. (1994). An introduction to HLM: Computer Program and Users' Guide. Version 2. Chicago: Department of Education, University of Education.

Bryk, A.S., Raudenbush, S.W. (1992). Hierarchical linear models for social and behavioral research: Applications and data analysis methods. Newbury Park, CA: Sage, 1992.

DeFinetti (1964). Foresight: its logical laws, its subjective sources. In Studies in Subjective Probability, (H.E. Kyburg, Jr., and H.E. Smokler, Eds.) pp. 93-158. New York: Wiley.

de Leeuw, J., & Kreft, I. (1986). Random coefficient models for multilevel analysis. Journal of Educational Statistics, 11(1), 57-85.

Fotiu, R.P. (1989). A comparison of the EM and data augmentation algorithms on simulated small sample hierarchical data from research on education. Unpublished doctoral dissertation, Michigan State University, East Lansing.

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. Journal of the American Statistical Association, 85, 398-409.

Goldstein, H.I. (1987). Multilevel models in educational and social research. London: Oxford University Press.

Glass, G.V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5, 3-8.

Johnson, E.G., Mazzeo, J, &, Kline, D.L. (1993). Technical report of the NAEP 1992 Trial State Assessment program in mathematics. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Little, R.J.A., & Shenker, N. (1995). Missing Data. In Handbook of Statistical Modeling for the Social and Behavioral Sciences, (Arminger, G., Clogg, C.C., Sobel, M.E., Eds.) pp. 39-75. New York: Plenum Press.

Mislevy, R.J. (1992). Scaling procedures in NAEP. Special Issue: National Assessment of Educational Progress. Journal of Educational Statistics, 17(2), 131-154.

Mullis, I.V.S., Dossey, J.A., Owen, E.H., Phillips, G.W. (1993, April). NAEP 1992 Mathematics Report Card for the Nation and the States. Princeton, NJ: Educational Testing Service.

Rubin, D.B. (1981). Estimation in parallel randomized experiments. Journal of Educational Statistics, 6, 377-400.

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

Seltzer, M.H. (1993). Sensitivity analysis for fixed effects in the hierarchical model: A Gibbs sampling approach. Journal of Educational Statistics, 18(3), 207-235.

Tanner, M. A., & Wong, W. H. (1987) The calculation of posterior distributions by data augmentation (with discussion). Journal of the American Statistical Association, 82, 528-550.

Figure 1

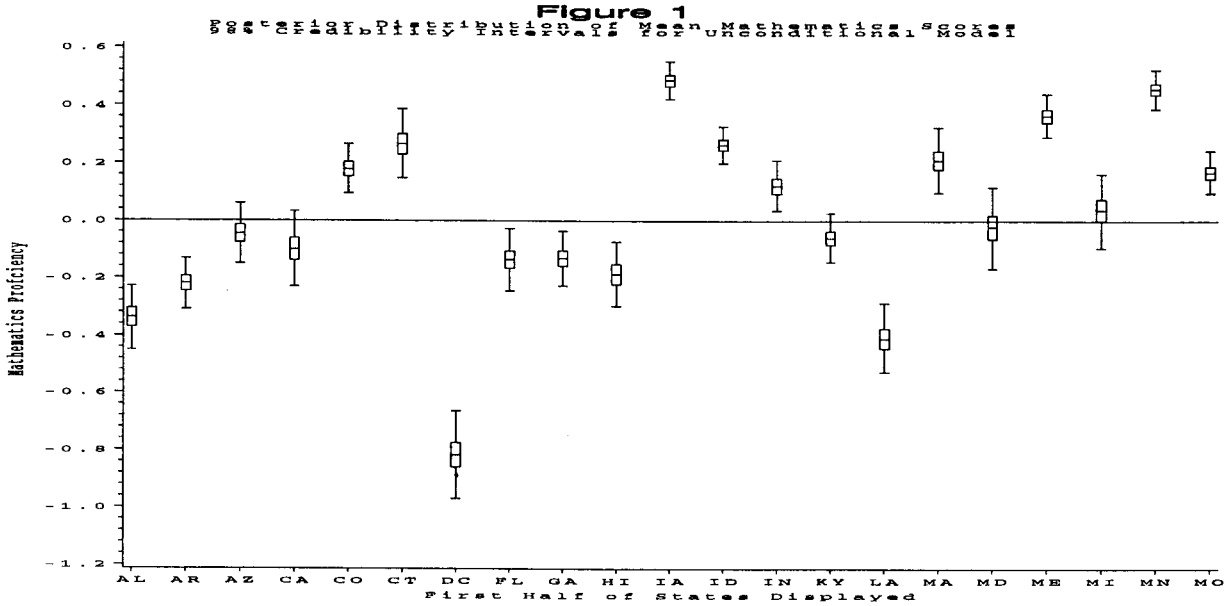

Figure 3