

Little Practical Difference and Pie in the Sky: A Response to Thum and Bryk and a Rejoinder to Sykes

William J. Webster
Robert Mendro
Timothy Orsak
Dash Weerasinghe
Karen Bembry

Dallas Public Schools

This response addresses the primary methodological issues discussed in Thum and Bryk with a point-by-point approach, provides a general response to Sykes, and gives an update of the Dallas Teacher Evaluation System since the original paper was prepared. It would be useful, however, to address globally the issues brought up by the respondents before launching into specific remarks. In regard to Thum and Bryk, based on the hundreds of analyses and dozens of systematic research studies conducted in the Dallas Public Schools over the past decade, involving 200 schools, at least 6000 teachers, and at least 300,000 students, there is no one best way to estimate teacher and school effects. (Olson and Webster, 1986; Webster and Olson, 1988; Webster, Mendro, and Almaguer, 1993; Webster, et. al., 1995.) Rather, carefully thought out prediction models, whether they are conducted in one or two stages, are ordinary least-squares regression models using appropriate interactions or are hierarchical linear models, or whether they treat predictor variables as fixed or random produce practically identical results. (Since gain models measure a different outcome, we cannot comment on them until we have more thoroughly analyzed the outcomes and their relation to the raw data. If anything attracts our attention to the gain model, it is Bereiter's (1963) suggestion that gain scores may not

have the seemingly straightforward intrinsic meaning imputed to them when subjected to careful inspection.)

There are some exceptions to this rule. For example, consider models where only school level data are available because student-level data have been aggregated to the school level and models where the only student assessment data available is the outcome measure which is then regressed against classifying variables to produce results. Both are seriously biased with respect to the initial achievement levels of students or the classificatory variables which are related to their outcomes.

Once the information from typically used classificatory variables and their interactions, from pretest measures (for a single year or many years in a longitudinal model), and from school or classroom level variables have been used in a prediction model, there is very little systematic information left. From a practical standpoint, there is no significant variation left which can be accounted for and resulting estimates of overall effects are nearly identical to those from another model using the same information. The point is that our experience with the data has shown that there is only so much systematic variation in these data and that a carefully thought out and explicated model will produce results virtually identical to those from another such model.

Of course, this is a simplification. We would not be using a hierarchical linear model for our own school and teacher effectiveness analyses if we did not believe, on the basis of extensive investigations summarized in our published results, that the multi-level approach offered a way of obtaining estimates of effects that were minutely improved over estimates from other models. However, the basic characteristics of all of the models are similar. The correlations of resulting estimates of effectiveness at the school level

and of effectiveness components at the student level with the variables controlled in the model are all acceptably low or practically zero, and the intercorrelation of the effectiveness estimates is quite high.

Considering Sykes, twenty-five years experience in operating testing systems, fifteen years experience in developing curriculum-referenced testing programs, and ten years of devising and implementing fair and accurate accountability systems has shown the impracticality of “authentic” or performance tests in large scale accountability programs. This is not to say that performance assessments have no place. We firmly believe that they should be used as an integral part of the curriculum. It is simply in large scale accountability testing that these measures have yet to show practical applicability, especially when compared to commercially available or carefully constructed local assessment instruments. (Webster and Schuhmacher, 1973; Webster, 1991.)

There are certainly drawbacks to any real accountability system. Implementing systems of school and teacher evaluation, assessing school effectiveness, or making any attempt to assign accountability to individuals or organizations will tempt a small percentage of individuals to indulge in unethical behavior to beat the system. However, this is not limited to testing systems. Systems of teacher evaluation rating teacher performance in a small number of visits have produced cheating and distortion on a scale equal to or exceeding any done with testing systems. This lamentable conclusion in no way justifies giving up accountability systems.

A more serious concern is the damage done to the curriculum by those who limit it to what they perceive will be covered on a test instrument. We feel that there is a very direct answer to this concern. Assessments must cover a sufficient portion of the

curriculum and assess higher-order skills with enough thoroughness that this broad array of skills and abilities is taught in order to meet the assessments. If assessments are sufficiently broad and deep, there is then nothing wrong with teaching to the assessment.

There are no magic answers to these concerns to be found in performance assessment systems or a yet-to-be-devised system of “real” outcomes and “authentic” attributes. A small percentage of misguided individuals will still be inclined to cheat on these measures if they perceive real rewards or penalties attached to them. (Not to mention a small percentage of parents and students then inclined to cheat in producing products for extended assessments.) The curriculum will still suffer from those who perceive the necessity of limiting it to the outcome tasks. (Not to mention the limitation to the curriculum that comes from the necessarily small number of tasks that can be assessed in practical applications of these systems.) Finally, the reliability and, consequently, the validity of these systems are still not established for use in large scale accountability.

Specific Concerns with Thum and Bryk

Outcome Variables. A number of outcome variables (promotion rate, dropout rate, etc.) have been classified as school covariates. The school covariates are listed in the text in the description of the HLM analysis and are further explicated in our following text. The variables listed in Table 1 of our chapter are all outcome variables, weighted and used in determining school effectiveness. The broad array of outcome variables utilized in the Dallas system is result of many hours of discussion about the purposes of education by the Accountability Task Force which oversees the system. The Task Force is discussed in more detail below.

Two-Phase Analysis. A complaint is registered against the homogenizing of residuals in the predictor space on the grounds that it is “likely to induce unpredictable biases in teacher or school effectiveness estimates” since teachers and schools might primarily serve students drawn from one or a few of the subgroups. The homogenization of the individual predictor space responses is done exactly to address this concern. If a school had a concentration of students from one of these groups, the homogenization is performed to assure that it would have no particular advantage or disadvantage. Thorough examination of our data sets for the past 5 years has shown that there are small perturbations in the means of the residuals and that the variances of the residuals at the lower end of the distribution tend to be larger than those from the upper end. It has also shown that our concerns are largely groundless. Students from any one school do not cluster in any of the particular intervals in the regression space. But, to assure that these perturbations do not add a bias, they are standardized.

We have found from experience with the data sets that the use of these corrections is largely for our own peace of mind. The practical effect is virtually nil in terms of the school effectiveness outcomes. Analyses with the two-stage model and with the one-stage HLM model produce virtually identical effect outcomes with correlations of outcomes above .978. Careful analysis of the data show that students are distributed across the outcome space so completely that there is a wide variety of student ability across schools and in each school and classroom. (District policy prohibits homogeneous grouping except for temporary purposes such as reading groups within a reading class. The data bear out the relatively thorough implementation of this policy.) However, were this not the case, the homogenization would guarantee that no school or teacher gained an

advantage from the mean differences within each adjustment grouping. The comments regarding the reliability of residuals would be of concern except for the fact that correlations between the results produced by one-phase HLM models and the two-phase model are very high.

This brings up an issue regarding the use of the effectiveness data. Were we attempting to research the relationships in the data to show the degree or nature of the interactions in the data set, we would not use the smoothing technique because of its grievous damage to accurate probability estimation. Since we are limiting our analysis to a descriptive situation, there is leeway in the analysis to adjust our data without concern for inferential uses.

Regarding the two-phase analysis, we agree that as a regression analysis it can be done in one step. However, by using two phases, the output of the first phase includes an analysis of mean residuals which shows that the fairness subgroups have been equalized and that the playing field has been leveled. At the direction of the Accountability Task Force (which is described in more detail in the response to Sykes), the two-phase analysis was specifically retained since in their opinion, it offered proof to critics that for every analysis in the system, student inequities had been adjusted within all fairness groups. The continual need to use this output particularly when the system has come under more formal challenges has demonstrated the political insight of the Task Force in desiring to retain the two-phase analysis. If the analysis cannot be demonstrated as fair, the degree of community buy-in to the process definitely falters.

Fixed or Random Slopes. The choice of fixed or random slopes depends in our view on the nature of the sources of variation in the slopes. The slopes are modeled using

a number of school parameters at the second level. These include, as mentioned in the paper, percent of minority students, percent black, percent low income, percent mobility, etc. To the extent that slopes vary as a result of these factors, their use adjusts the differences. We choose a random model to control for the effects of possible interactions of concomitant variables in specific school settings. If we had evidence of an interaction of school effect with these variables we would use the fixed model since to use the random model would mask these effects. Other analyses of the data that we have conducted regarding slopes within subgroups do not lead us to suspect an interaction of the slopes with instructional effect. However, our analyses of fixed and random results shows that our worries about effects due to these factors and the possibility of interactions are both very remote. Correlations of school effects from fixed and random slope models are above .985.

Variable Treatment. In our analyses, we grand mean center all variables at level 1. Also, regarding our misstatement of the HLM solution, our statement should have read “The HLM equations are solved and empirical Bayes residuals used for school effects.” The following gives a more explicit statement of our model.

Y_{lij} = Outcome variable of interest for each student i in school j . l is a measure for grade/subject/year.

X_{1ij} = Black English Proficient Status (1 if black, 0 otherwise).

X_{2ij} = Hispanic English Proficient Status (1 if Hispanic, 0 otherwise).

X_{3ij} = Limited English Proficient Status (1 if LEP, 0 other).

X_{4ij} = Gender (1 if male, 0 if female).

- X_{5ij} = Free or Reduced Lunch Status (1 if subsidized, 0 otherwise).
- X_{6ij} = School Mobility Rate (same for all i in each j).
- X_{7ij} = School Overcrowdedness (same for all i in each j).
- X_{8ij} = Block Average Family Income
- X_{9ij} = Block Average Family Education Level
- X_{10ij} = Block Average Family Poverty Level
- X_{kij} = indicates the variable k for i^{th} student in school j for $i = 1, 2, \dots, I_j$ and $j = 1, 2, \dots, J$.

Student Level Variables:

- r_{ij}^{95} = Posttest Residual Score from fairness stage for measure l for i^{th} student in school j . In this paper it represents *ITBS* Reading 1995 or *ITBS* Mathematics 1995.
- r_{hij}^{94} = h^{th} predictor used to estimate r_{ij}^{95} for i^{th} student in school j . This is a Pretest Residual score from the fairness stage. In this paper it represents *ITBS* Reading 1994 and *ITBS* Mathematics 1994.
- r_{lij} = $Y_{lij} - \hat{Y}_{lij}$ from OLE

School Level Variables:

- W_{1j} = School Mobility
- W_{2j} = School Overcrowdedness
- W_{3j} = School Average Family Income
- W_{4j} = School Average Family Education

- W_{5j} = School Average Family Poverty Index
 W_{6j} = School Percentage on Free or Reduced Lunch
 W_{7j} = School Percentage Minority
 W_{8j} = School Percentage Black
 W_{9j} = School Percentage Hispanic
 W_{10j} = School Percentage Limited English Proficient

The MLR and HLM model used:

Stage 1:

$$\begin{aligned}
 Y_{ij} = & \Lambda_0 + \Lambda_1 X_{1ij} + \Lambda_2 X_{2ij} + \Lambda_3 X_{3ij} + \Lambda_4 X_{4ij} + \Lambda_5 X_{5ij} + \Lambda_6 X_{6ij} + \Lambda_7 X_{7ij} + \Lambda_8 X_{8ij} + \Lambda_9 X_{9ij} \\
 & + \Lambda_{10} X_{10ij} + \Lambda_{11} (X_{1ij} X_{4ij}) + \Lambda_{12} (X_{2ij} X_{4ij}) + \Lambda_{13} (X_{3ij} X_{4ij}) + \Lambda_{14} (X_{1ij} X_{5ij}) + \\
 & \Lambda_{15} (X_{2ij} X_{5ij}) + \Lambda_{16} (X_{3ij} X_{5ij}) + \Lambda_{17} (X_{4ij} X_{5ij}) + \Lambda_{18} (X_{1ij} X_{4ij} X_{5ij}) + \Lambda_{19} (X_{2ij} X_{4ij} X_{5ij}) + \\
 & \Lambda_{20} (X_{3ij} X_{4ij} X_{5ij}) + \varepsilon_{ij}
 \end{aligned}$$

where $\varepsilon_{ij} \sim \text{i.i.d.} \sim N(0, \sigma^2)$.

$$r_1^{95} = Y_1^{95} - \hat{Y}_1^{95}$$

$$r_1^{94} = Y_1^{94} - \hat{Y}_1^{94}$$

Y_1^{94}, Y_1^{95} = Student's scores in 93/94 and 94/95 respectively, for math and reading.

Stage 2:

$$r_{ij}^{95} = \beta_0 + \beta_1 r_{1ij}^{94} + \beta_2 r_{2ij}^{94} + \delta_{ij}$$

$$\beta_{kj} = \gamma_{k0} + \gamma_{k1} W_{1j} + \gamma_{k2} W_{2j} + \gamma_{k3} W_{3j} + \gamma_{k4} W_{4j} + \gamma_{k5} W_{5j} + \gamma_{k6} W_{6j} + \gamma_{k7} W_{7j} + \gamma_{k8} W_{8j} + \gamma_{k9} W_{9j} + \gamma_{k10} W_{10j} + u_{kj}$$

for $i = 1, 2, \dots, I_j$

$j = 1, 2, \dots, J$

$k = 0, 1, 2.$

where $E(\delta_{ij}) = 0$, $\text{Var}(\delta_{ij}) = \sigma^2$, $E(u_{kj}) = 0$, $\text{Var}(u_{kj}) = \sigma^2$, and $\delta_{ij} \perp u_{kj}$. All level 1 and level 2 variables are grand mean centered.

School Rankings

The school rankings are obtained from the empirical Bayes residual for β_{00} , which is u_{00}^* , where

$$u_{00}^* = \beta_{00}^* - (\hat{\gamma}_{00} + \sum_{s=1}^{S_q} \hat{\gamma}_{0s} W_{sj})$$

$$\beta_{00}^* = \lambda_0 \bar{r}_{\cdot 0}^{95} + (1 - \lambda_0) \hat{\gamma}_{00}$$

$$\lambda_0 = \frac{\text{Var}(\beta_{00})}{\text{Var}(\bar{r}_{\cdot 0}^{95})}$$

Teacher indices

$$S_{ij} = r_{ij}^{95} - \hat{r}_{ij}^{95adj}$$

$$\mu = \frac{\sum_{j=1}^J \sum_{t=1}^{T_j} \sum_{k=1}^{K_{tj}} \bar{s}_{tj}^k}{\sum_{j=1}^J \sum_{t=1}^{T_j} K_{tj}}$$

$$\sigma^2 = \frac{\sum_{j=1}^J \sum_{t=1}^{T_j} \sum_{k=1}^{K_{tj}} (\bar{s}_{tj}^k - \mu)^2}{\sum_{j=1}^J \sum_{t=1}^{T_j} K_{tj}}$$

To calculate the B.L.U.P. of \bar{s}_{tj}^k for the t^{th} teacher in school j ,

Let

$$\bar{s}_{tj} = \frac{\sum_{k=1}^{K_{tj}} \bar{s}_{tj}^k}{K_{tj}}$$

$$\sigma_{tj}^2 = \frac{\sum_{k=1}^{K_{tj}} (\bar{s}_{tj}^k - \bar{s}_{tj})^2}{K_{tj}}$$

is the error variance for TEI for teacher t in school j ,

then

$$TEI_{tj} = \mu + (\bar{s}_{tj} - \mu) \left(\frac{\sigma^2}{\left(\sigma^2 + \frac{\sigma_{tj}^2}{k_{tj}} \right)} \right).$$

Response to Issues Raised by Sykes

The Sykes paper demonstrates straightforwardly why the public is losing confidence in public education. If one follows his argument to its logical conclusion, no one can be held accountable for anything. To indulge in a diatribe against standardized testing without presenting a meaningful alternative is of little value. “Authentic assessment” is, because of generally low reliability when employed in accountability systems, neither authentic nor assessment. The Dallas system does in fact include some “authentic assessment” in that the *Texas Assessment of Academic Skills* includes a writing sample that is used as one of the outcome measures in the system. It is, of course, the least reliable measure imposed on the district by the state.

The district’s *Assessments of Course Performance (ACP)*, final examinations in each high school core course, were originally designed by master teachers at the behest of the central “bureaucratic organization” to include performance sections. The plan was to include these parts of the *ACP* in the accountability system as outcomes and to weight them by their reliability. Negative reaction from the field in terms of too much time spent testing for too little information received and from the Board in terms of the costs of this portion of the testing led to the demise of this part of the plan.

This does not imply that performance assessments have no place. Schools are encouraged to develop or adopt expanded measures of student performance and supplied help in the process as well as in developing student portfolios. The teacher evaluation system includes these data, when schools and teachers choose, as part of the needs assessment on the teacher evaluation instrument. But, because of expense in time and money and the unreliability of results, “authentic assessments” are viewed as curriculum measures and instructional tools rather than as large-scale assessment measures and accountability tools.

While the authors are familiar with most of the literature related to “authentic assessment” and, particularly, to its use in large-scale accountability systems, Sykes’ understatement that “technical difficulties in administering and scoring the new assessments continue to make their use problematic as instruments of public policy” says it all. It is not practical in terms of time, money, or information received to include “authentic assessments” as part of a large-scale accountability system. The only reason the Dallas system includes a writing sample is because the State of Texas mandates it and pays for the scoring. Standardized tests, either commercially available or locally developed at the behest of bureaucrats, are not outmoded and still provide accurate, reliable and efficient measures of student achievement. The norm-referenced tests used in Dallas, the survey forms of the *ITBS* and *TAP* which include primarily higher-order skill based items, provide adequate estimates of students’ abilities to read and compute. The curriculum based measures, whether the local measures in the *ACP* or the state measures in the multiple-choice portions of the *TAAS*, provide accurate measures of the elements of the curriculum.

Sykes' contention that the Dallas accountability system reflects a "bureaucratic conception of organization and management that allocates planning and design to the central office and implementation of central directives to subordinate workers in the schools" is wrong. The Dallas accountability system was suggested by a citizen's task force after a comprehensive review of the state of accountability across the country. This task force, consisting of parents, community members, and statisticians from outside education, did not recommend that "authentic assessment" be included in the accountability system for exactly the reasons Sykes mentions, cited earlier. Furthermore, the system was designed and continues to be implemented under the auspices of an Accountability Task Force that includes teachers, parents, administrators, and community people. The teacher evaluation system was designed and continues under the direction of a Teacher Evaluation Task Force that includes 63 members, the majority of whom are teachers. The major teacher organizations have participated in the development of the system and the most prominent of them has publicly endorsed it.

To say more than this would give Sykes' suggestions more value than they deserve. To close, we draw upon Sykes' analogy to the judicial system. A grand jury, considering the Dallas system in light of Sykes' charges and prosecution, would be bound to no-bill the system. To extend the analogy a bit further, a civil suit on the same charges would run a strong chance of being declared frivolous litigation.

Update on the Dallas Teacher Evaluation System

Throughout the 1995-96 school year, the Dallas Public Schools developed and field-tested a new teacher evaluation system. As noted, central to this development was the Teacher Evaluation Task Force. Task Force members include teachers from each of

the ten geographic areas of the city, parents, principals, community members and representatives from teacher organizations as well as a representative from an administrator organization. Based on input received from these various constituents, from the field test which was conducted throughout the school district, and from one of the teacher organizations, the following adjustments to the system have been made.

The emphasis on continuous student improvement remains the focus of the teacher evaluation system. Student achievement data drives the teacher evaluation system just as it drives the principal evaluation process, the campus improvement plan, and the district improvement plan. The alignment of all accountability systems district-wide has been retained. However, how the system would utilize Effectiveness Indices for the teachers has been redefined.

The initial framework for the teacher evaluation system included a composite Teacher Effectiveness Index for each teacher who administered an achievement test or end-of-course exam. This composite score was used to assign each teacher to one of three tiers: Tier 1 was determined to be the top 40% of teachers across the district, Tier 2 the next 50%, and Tier 3 the bottom 10%. The initial purpose for establishing tiers was to focus the finite resources of the district where they were most needed, presumably in support of the bottom 10% of the teachers. However, the tier system was not well received, and it resulted in negative feelings from teachers that interfered with the ultimate purpose of the system: continuous student improvement. The tier system was voted out by the Task Force, and both the composite score for each teacher and the tier designation was dropped from the Effectiveness Indices format. Also, the recommendation that the name be changed from Teacher Effectiveness Indices to

Classroom Effectiveness Indices was accepted. While leaving the indices unchanged, the name change made the process less intimidating to school staff members. Now, Classroom Effectiveness Indices summarize student residual data computed for each standardized test a teacher administers aggregated by variable or class and utilized strictly for the purpose of diagnosing student growth or lack thereof.

Classroom Effectiveness Indices are one way of diagnosing the needs of students in the revised teacher evaluation system. The process has been expanded to incorporate other forms of needs assessment as well. Each teacher submits an Instructional Improvement Plan that consists of three sections: a *Needs* section, a *Concepts/Content/Strategies* section, and a *Final Evaluation* section. The first section, *Needs*, requires the teacher to list identified needs for his or her students. These needs take two forms: needs of the teacher's students from the previous year as identified in the Classroom Effectiveness Indices, and needs of current students as identified from other available sources. Other data sources may include student profiles, portfolios, teacher-made tests, diagnostic skills analyses, and performance information. If a teacher does not have Classroom Effectiveness Indices, he or she uses the alternative forms of assessments to diagnose current students. Thus, although the system still includes teacher with indices and teachers without indices, this distinction no longer divides teachers into two different groups for evaluation; it simply indicates the potential sources of data available to a teacher for his or her needs analysis.

Each teacher, whether or not he or she has Classroom Effectiveness Indices, lists identified needs in the first section. Instructional practices designed to meet those needs are listed in the *Concepts/Content/Strategies* section, and documentation that will

indicate the instructional practices have been accomplished is listed in the third column. The successful completion of the Instructional Improvement Plan results in a successful evaluation for the year. All teachers will follow the same process for evaluation each year, yet the extent of the evaluation will fluctuate in response to the identified needs of the students. A form for an Instructional Improvement Plan is included.

The Teacher Evaluation Task Force will continue to refine the system during the current school year. The ultimate test of the system will be whether or not student achievement and performance continue to improve.

References

- Bereiter, C. (1963). Some Persisting Dilemmas in the Measurement of Change. In *Problems in Measuring Change* (Ed. by C. W. Harris). Madison, WI: University of Wisconsin Press.
- Olson, G. H. and Webster, W. J. (1986). *Measuring School Effectiveness: A Three-year Study*. A paper presented at the 1986 annual Meeting of the American Educational Research Association.
- Webster, W. J. (1991). An Analysis of Available Student Achievement Data in the Dallas Independent School District, *Executive Summaries of Evaluation Reports*, Dallas, TX: Dallas Independent School District.
- Webster, W. J., Mendro, R. L., and Almaguer, T. O. (1993). *Effectiveness Indices: The Major Component of an Equitable Accountability System*, ERIC TM 019 913.
- Webster, W. J., Mendro, R. L., Bembry, K. L., and Orsak, T. H. (1995). *Alternative Methodologies for Identifying Effective Schools*. A paper presented at the 1995 Annual Meeting of the American Educational Research Association, San Francisco, CA.
- Webster, W. J. and Olson, G. H. (1988). A Quantitative Procedure for the Identification of Effective Schools. *Journal of Experimental Education*, 56, 213-219.
- Webster, W. J. and Schuhmacher, C. C. (1973). A Unified Strategy for Systemwide Research and Evaluation. *Educational Technology*, 13, 5, 68-72.