ED 471 666 TM 034 695

AUTHOR	Thum, Yeow Meng
TITLE	Measuring Student and School Progress with the California API. CSE Technical Report.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.; National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO	CSE-TR-578
PUB DATE	2002-10-00
NOTE	36p.
CONTRACT	R305B960002-01
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	Academic Achievement; *Achievement Gains; Bayesian Statistics; Elementary Secondary Education; Meta Analysis; *School Statistics; State Programs

ABSTRACT

This paper focuses on interpreting the major conceptual features of California's Academic Performance Index (API) as a coherent set of statistical procedures. To facilitate a characterization of its statistical properties, the paper casts the index as a simple weighted average of the subjective worth of students' normative performance and presents its estimation in the form of a linear model. The paper also illustrates several problems with this index for the study of a school's year-to-year progress. In current use, the API lacks realistic estimates of precision and, on closer examination, is seen to misrepresent student and school performance conceptually. The paper presents an alternative analysis of the API index, based on a Bayesian meta-analysis of results from school-specific multilevel models of longitudinal student test scores. A display is introduced for the precision of estimated relative gains of each school in the form of a profile that represents the probability that a gain estimate exceeds set fractions of the distance the pretest is from the statewide target of 800. Along with estimates of their reliabilities, researchers also produced rank estimates of school API gains rather than simply ranking schools. The approach is illustrated with an elementary school cohort who too the Stanford 9 at the Long Beach Unified School District in spring 2000. An appendix contains a method for inference on the Public Schools Accountability Act ratio. (Contains 6 tables, 6 figures, and 58 references.) (Author/SLD)

Measuring Student and School Progress With the California API

CSE Technical Report 578

Yeow Meng Thum
Graduate School of Education and Information Studies
University of California, Los Angeles/CRESST

October 2002

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 3.1: Methodologies for Assessing Student Progress
Michael Seltzer and Yeow Meng Thum, Project Directors, CRESST/UCLA, Graduate School of Education and Information Studies

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002-01, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

Measuring Student and School Progress with the California API

Yeow Meng Thum*

Graduate School of Education and Information Studies
University of California, Los Angeles / CRESST
October 2002

Abstract. This paper focuses on interpreting the major conceptual features of California's Academic Performance Index (*API*) as a coherent set of statistical procedures. To facilitate a characterization of its statistical properties, we first cast the index as a simple weighted average of the subjective worth of students' normative performance and present its estimation in the form of a linear model. In the process, we illustrate with an example several problems with this index for the study of a school's year-to-year progress. In its current usage the API lacks realistic estimates of precision and, on closer examination, further misrepresents conceptually student and school performance. We present an alternative analysis of the API index, based on a Bayesian meta-analysis of results from school-specific multilevel models of longitudinal student test scores. We introduce a display for the precision of estimated relative gains of each school in the form of a profile that represents the probability that a gain estimate exceeds set fractions of the distance the pretest is from the statewide target of 800. Along with estimates of their reliabilities, we also produce rank estimates of school API gains rather than simply ranking schools. We illustrate our approach with an elementary school student cohort who took the Stanford 9 at the Long Beach Unified School District in the Spring of 2000.

Keywords. Academic Performance Index, Acceptability curves, Bayesian meta-analysis, Index score, Latent variable regression, Measuring progress, Multilevel modelling, Ratios, Ranks, Reliability, School performance.

1 Introduction

Like many systemic efforts across the country, California has taken serious steps to build a statewide accountability system to help public schools better gauge and improve the academic performance for all of their students. The Public Schools Accountability Act (PSAA), which became law in 1999 (SB 1X, Chapter 3 of 1999), requires that the State Board of Education (SBE) design a statewide numerical index, the *Academic Performance Index* (API), for measuring the

> " ... performance of schools, especially the academic performance of pupils, and demonstrate comparable improvement in academic achievement by all numerically significant ethnic and socioeconomically disadvantaged subgroups within schools." *(PSAA, Article 2, Section 52052 (a))*

The importance of the API, officially dubbed the "cornerstone" of California's accountability effort, cannot be overstated. In this paper, we will examine the major conceptual features of this instrument

* This research was supported in part by the Center for Research on Evaluation, Standards, and Student Testing (CRESST). The author thanks Professor María Castro of the Facultad de Educación, Universidad de Murcia, Campus de Espinardo, 30100 Murcia (Spain), for close collaboration during 2000 that laid much of the groundwork for the ideas dealt with in this paper. We are very grateful to Dr. Lynn Winters of the Long Beach Unified School District for the data. Katharine Fry of CRESST provided invaluable copy-editing assistance. Early versions of this work first appeared in February 2002. Opinions expressed in this paper are however the sole responsibility of this author, as are all remaining errors. Please direct all correspondence to Y. M. Thum, at thum@ucla.edu.

for the measurement of student and school change, leaving aside its use for the accompanying reward system. After we briefly explain how the index defines student and school progress with the help of simple examples, we suggest that in its present usage the API measures neither. We then illustrate an approach to measuring school performance which retains the broader features of the index and that is feasible within the boundaries of the present API information base. In concluding, we hope to have articulated a general procedure for making useful data-driven accountability decisions beyond the immediate concerns of any single index.

2 Unpacking the API

When compared with indicators that simply aggregate students, raw test scores in various way to the school-level, the API formula has two distinctive features. First, it employs a set of weights (reasonable perhaps, albeit subjective) to express the "worth" attached by the SBE to the student's *normative* performance. In particular, the student national percentile rank (NPR) for each test (Language, Mathematics, Reading, Science, History and Social Science, or Spelling) is assigned a value of either 200, 500, 700, 875 or 1000 depending on whether the student NPR falls in the first to fifth NPR quintiles (1-19, 20-39, 40-59, 60-79, 80-99 – the so-called *performance bands*).[1] These five values form the basis of all further API calculations. For each relevant subject area, the API test component is simply the school average of student assigned values. Note further that, by construction, the API ranges from a minimum of 200 (when all scores fall within the lowest quintile) to a maximum of 1000 (when every score equals or exceeds the 80th %tile rank).

2.1 The API Component for a Test, API_{jk}

To fix ideas, suppose we index each of the $B = 5$ performance bands by $b = 1, 2, \ldots, B$. Let the number of students in performance band b be n_b and p_b represents the proportion of students in band b. We further denote the series of values that are set at (200, 500, 700, 875, 1000), the worth assigned to each performance band b, as v_b. If $j = 1, 2, \ldots, N$ denotes a school and $k = 1, 2, \ldots, K$ denotes a specific test, the API_{jk} component estimate is then simply the weighted mean of the student scores v_b's, or

$$\widehat{API}_{jk} = \frac{1}{n} \sum_{b=1}^{B} n_b \times v_b = \sum_{b=1}^{B} p_b \times v_b \ , \tag{1}$$

where n is $\sum_b n_b$. Given (1), the weighted variance of v_b is

$$\mathrm{Var}(v_b) = \frac{1}{(n-1)} \sum_{b=1}^{B} n_b \times \left(v_b - \widehat{API}_{jk}\right)^2$$

and the standard error of the weighted mean, our school $\widehat{API}_{jk}$ component, is the square-root of its sampling variance

$$\mathrm{Var}(\widehat{API}_{jk}) = \frac{\mathrm{Var}(v_b)}{n} \ . \tag{2}$$

[1] The interested reader may visit the PSAA home-page at http://www.cde.ca.gov/PSAA/API/ for additional clarification.

Example 1 (API Component Mean and Standard Error). If the 100 students in your school are placed in the lowest to the highest performance band numbers in the proportions 20%, 20%, 40%, 10%, and 10%, then your school API is 607.5, with a standard error of 24.92, by Equations (1) and (2). If the total number of students is 1000 and the proportions remain as they were, the standard error drops to 7.85. □

2.2 The School API, API_j

The second distinctive aspect of the API is that the results for different subject matter may be weighted differently. Just like the student normative attainment weighting already described, a similar mechanism now expresses the varying worth we attach to the different subjects. The SBE chooses a different scheme for elementary and middle schools and for high schools. In grades 2-8, the Language test counts for 15%, Mathematics for 40%, Reading for 30%, and Spelling for the remaining 15%. For grades 9 and above, the percentile weights are Language, 20%; Mathematics, 20%; Reading, 20%; History and Social Science, 20%; Science, 20%. The API_j for a school is then the weighted sum of subject specific components, API_{jk}.

Even with potentially different subject matter weights, calculating the school API remains straightforward. Let $k = 1, 2, \ldots, K$ indexes the subject matter tests as before and, for each test, w_k denotes the weight for the corresponding test and $\sum_k w_k = 1$. We note that K equals 4 for elementary and middle schools and 5 for high schools. Using Equation (1) above for each test k in school j, the school API, API_j, can be expressed as a weighted composite of the API_{jk} for the different tests

$$\widehat{API}_j = \sum_{k=1}^{K} w_k \times \widehat{API}_{jk} \ . \tag{3}$$

From Equation (3), we see immediately that the standard error is the square-root of the variance of a weighted composite of (independent) components, or

$$\mathrm{Var}(\widehat{API}_j) = \sum_{k=1}^{K} w_k^2 \times \mathrm{Var}(\widehat{API}_{jk}) \ . \tag{4}$$

Equation (4) provides a conservative estimate of the uncertainty we attach to our school API estimate. For now, we will leave aside the issue of whether or not to "pool" variances to "improve" its precision.

Example 2 (School API and its Standard Error). An elementary school received the following subject-specific API mean and standard error estimates: Language (521.65, 13.09), Mathematics (598.60, 12.99), Reading (490.87, 12.46), and Spelling (557.67, 13.38). Weighting Language and Spelling at 15% each, Mathematics at 40% and Reading at 30%, the school API is 548.60 with a standard error of 6.99, by Equations (3) and (4). □

3 The API as a Linear Model

Up to this point, we have described procedures based on well-known results for making statistical inference on independent means and variances. A more comprehensive approach will need to make explicit the assumptions made about the sampling design producing the data, about error distributions, etc., as all these factors are necessary for us to make sense of the additional complications when we wish to monitor performance over time. In this section, we show how the familiar general linear model may be used to represent the school API and to describe how the school APIs change over time.

3.1 The $API_{jk}^{(t)}$ for school j at time t on test k

Let $t = 1, 2, \ldots, T$ index the time period under study. We observe the longitudinal array of performance band frequencies for an elementary school j at time t as

$$n_{jk}^{(t)} = \left[n_{1jk}^{(t)}\; n_{2jk}^{(t)}\; n_{3jk}^{(t)}\; n_{4jk}^{(t)}\; n_{5jk}^{(t)} \right] , \tag{5}$$

with $n_{bjk}^{(t)}$ students scoring in performance band b on test k. If $\boldsymbol{v}$ is the $B \times 1$ outcome vector, the $API_{jk}^{(t)}$, now denoted by $\mu_{jk}^{(t)}$, may thus be represented as a conventional linear model

$$\mathbf{H}_{jk}^{(t)}\boldsymbol{v} = \mathbf{H}_{jk}^{(t)}\mathbf{1}_B\mu_{jk}^{(t)} + \mathbf{H}_{jk}^{(t)}\boldsymbol{e}_{jk}^{(t)} , \tag{6}$$

where $\mathbf{1}_B$ is the $B \times 1$ unit vector. $\mathbf{D}_{jk}^{(t)}$ is a diagonal matrix of the performance band frequencies for time t, *i.e.*, $\boldsymbol{n}_{jk}^{(t)}$, and $\mathbf{D}_{jk}^{(t)} = \mathbf{H}_{jk}^{(t)}\mathbf{H}_{jk}^{(t)\prime}$. We further assume that the residual errors are independently and identically distributed normal, or $\boldsymbol{e}_{jk}^{(t)} \sim \mathcal{N}(\mathbf{0}_B, \sigma^2\mathbf{I})$. Keep in mind that a more elaborate error structure, if better supported by the data, would be preferred should it improve our inference.

Standard results for the general linear model suggest that the solution for $\mu_j^{(t)}$ is simply

$$\hat{\mu}_{jk}^{(t)} = \left[\mathbf{1}_B'\mathbf{D}_{jk}^{(t)}\mathbf{1}_B\right]^{-1}\mathbf{1}_B'\mathbf{D}_{jk}^{(t)}\boldsymbol{v} , \tag{7}$$

$$\hat{\sigma}^2 = \left(\left[\boldsymbol{v} - \mathbf{1}_B\hat{\mu}_{jk}^{(t)}\right]'\mathbf{D}_{jk}^{(t)}\left[\boldsymbol{v} - \mathbf{1}_B\hat{\mu}_{jk}^{(t)}\right]\right) / \left(\mathbf{1}_B'\mathbf{D}_{jk}^{(t)}\mathbf{1}_B - 1\right) , \tag{8}$$

$$\text{s.e.}\left[\hat{\mu}_{jk}^{(t)}\right] = \hat{\sigma}\left[\mathbf{1}_B'\mathbf{D}_{jk}^{(t)}\mathbf{1}_B\right]^{-\frac{1}{2}} . \tag{9}$$

$\mathbf{1}_B'\mathbf{D}_{jk}^{(t)}\mathbf{1}_B$ of course equals to $n_{\cdot jk}^{(t)} = \sum_b n_{bjk}^{(t)}$, the total number of scores. Most statistical software will easily perform the above calculations. We provide a simple example next, employing SAS© PROC GLM.

Example 3 (Calculating $API_{jk}^{(t)}$ - Example 1 continued). The following SAS PROC GLM code produces the correct estimates $\hat{\mu} = 607.5$, $\hat{\sigma} = 249.25$, and s.e.$(\hat{\mu}) = 24.92$, for the observed band proportions of (.2, .2, .4, .1, .1) for a school with a total of N= 100 students. In the data set, p, f, and v are the variables for the proportions, frequencies, and values respectively. □

Data and SAS PROC GLM code for Example 3

```
data;
 N=100;
 input p v @@;
 f=N*p;
cards;
0.2  200  0.2  500  0.4  700  0.1  875   0.1 1000
;
proc glm;
  freq f;
  model v= /solution;
```

3.2 The $API_{jk}^{(t)}$ for school j over tests and over time

Extending the above analysis to a time-series ($t = 1, 2, \ldots, T$) with multiple tests ($k = 1, 2, \ldots, K$) is straightforward. Suppose we let $\mathbf{1}_p$ be the $p \times 1$ unit vector and $\mathbf{I}_p$ be the $p \times p$ identity matrix for some p. We further define $\mathbf{D}_j$ to be the $TKb \times TKb$ (block-)diagonal matrix with elements

$$\left[\mathbf{n}_j^{(1)}, \mathbf{n}_j^{(2)}, \ldots, \mathbf{n}_j^{(t)}, \ldots, \mathbf{n}_j^{(T)}\right] ,$$

each block of which represents the number of scores at time point t for the set of tests observed in school j.

The general linear model for the occasion means is thus

$$\mathbf{H}_j(\mathbf{1}_T \otimes \mathbf{1}_K \otimes \boldsymbol{v}) = \mathbf{H}_j(\boldsymbol{\mu}_j \otimes \mathbf{1}_B) + \mathbf{H}_j(\mathbf{1}_T \otimes \mathbf{1}_K \otimes \boldsymbol{e}_j) , \tag{10}$$

where, again, $\mathbf{D}_j = \mathbf{H}_j\mathbf{H}_j'$. Here, $\otimes$ is the left Kronecker product, such that, for any $p \times q$ matrix $\mathbf{A}$ and $r \times s$ matrix $\mathbf{B}$, $\mathbf{A} \otimes \mathbf{B}$ is a $p \cdot r \times q \cdot s$ matrix

$$\mathbf{A} \otimes \mathbf{B} = \{a_{\imath\jmath} \cdot \mathbf{B}\} ,$$

$\imath = 1, 2, \ldots, p$ and $\jmath = 1, 2, \ldots, q$. $\boldsymbol{\mu}_j$ is the $TK \times 1$ vector of means with elements corresponding to each time-point and test. Under similar distributional assumptions, the estimator for Equation (10) is

$$\hat{\boldsymbol{\mu}}_j = \left[(\mathbf{I}_T \otimes \mathbf{I}_K \otimes \mathbf{1}_B)' \, \mathbf{D}_j \, (\mathbf{I}_T \otimes \mathbf{I}_K \otimes \mathbf{1}_B)\right]^{-1} (\mathbf{I}_T \otimes \mathbf{I}_K \otimes \mathbf{1}_B)' \, \mathbf{D}_j \, (\mathbf{1}_T \otimes \mathbf{1}_K \otimes \mathbf{I}_B) \, \boldsymbol{v} . \tag{11}$$

The result may be more obvious by noting that $(\mathbf{1}_T \otimes \mathbf{1}_K \otimes \boldsymbol{v}) = (\mathbf{1}_T \otimes \mathbf{1}_K \otimes \mathbf{I}_B)\boldsymbol{v}$, $(\boldsymbol{\mu}_j \otimes \mathbf{1}_B) = (\mathbf{I}_T \otimes \mathbf{I}_K \otimes \mathbf{1}_B)\boldsymbol{\mu}_j$, and $(\mathbf{1}_T \otimes \boldsymbol{e}_j) = (\mathbf{1}_T \otimes \mathbf{1}_K \otimes \mathbf{I}_B)\boldsymbol{e}_j$. Its standard error s.e.$(\hat{\boldsymbol{\mu}}_j)$ and the residual variance $\hat{\sigma}^2$ can be obtained from formulae analogous to Equations (6). For example, standard errors can be shown, *e.g.*, Bock (1975), to be

$$\text{s.e.} \left[\hat{\boldsymbol{\mu}}_j\right] = \hat{\sigma} \, \times \, \text{diag} \left[(\mathbf{I}_T \otimes \mathbf{I}_K \otimes \mathbf{1}_B)' \, \mathbf{D}_j \, (\mathbf{I}_T \otimes \mathbf{I}_K \otimes \mathbf{1}_B)\right]^{-\frac{1}{2}} . \tag{12}$$

Significant subgroups. The PSAA also requires that the system monitor the progress of significant subgroups (special education status, English language learners, socioeconomic status, gender, and ethnic groupings). Note that we need only to modify slightly Equation (10) as

$$\mathbf{H}_{gj}(\mathbf{1}_T \otimes \mathbf{1}_K \otimes \boldsymbol{v}) = \mathbf{H}_{gj}(\boldsymbol{\mu}_{gj} \otimes \mathbf{1}_B) + \mathbf{H}_{gj}(\mathbf{1}_T \otimes \mathbf{1}_K \otimes \boldsymbol{e}_{gj}) . \tag{13}$$

so that the API, $\boldsymbol{\mu}_{gjk}^{(t)}$, for any significant subgroup $g = 1, 2, \ldots, G$ may also be obtained. To keep the notation simple however, we will not further consider subgroup estimates in the sequel.

Subject matter weighting. Weighting test APIs, $\boldsymbol{\mu}_j$, *e.g.*, with $\boldsymbol{w} = (.15, .4, .3, .15)$ for elementary schools, is equivalent to estimating the linear composite $\boldsymbol{\mu}_j^*$, where

$$\hat{\boldsymbol{\mu}}_j^* = (\mathbf{I}_T \otimes \boldsymbol{w}') \, \hat{\boldsymbol{\mu}}_j , \tag{14}$$

with standard error

$$\text{s.e.} \left[\hat{\boldsymbol{\mu}}_j^*\right] = \hat{\sigma} \, \times \, \text{diag} \left[(\mathbf{I}_T \otimes \boldsymbol{w}' \otimes \mathbf{1}_B)' \, \mathbf{D}_j \, (\mathbf{I}_T \otimes \boldsymbol{w}' \otimes \mathbf{1}_B)\right]^{-\frac{1}{2}} , \tag{15}$$

if we rely on results (11) and (12).

A simple growth model. A similar argument lends itself to estimating growth parameters if we have a longer time-series, $t = 1, 2, \ldots, T$. The linear growth model may now be represented by projecting $\boldsymbol{\mu}_j^*$ on time. For example, we may pose

$$\mathbf{M} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & (T-1) \end{pmatrix} \tag{16}$$

to estimate the API at time point 1 and a linear growth rate. Note that "centering" the linear growth predictor at the last time point can give an alternative and useful interpretation of special interest for accountability purposes. Now β_{0j} is an estimate for the current status of school j, *i.e.*, at time T.

Setting $\boldsymbol{\mu}_j^* = \mathbf{M}\boldsymbol{\beta}_j$, we obtain

$$\hat{\boldsymbol{\beta}}_j = \left(\mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{M}'\hat{\boldsymbol{\mu}}_j^* \tag{17}$$

and, again, standard errors for the growth parameters, s.e. $\left[\hat{\boldsymbol{\beta}}_j\right]$, are

$$\hat{\sigma} \times \operatorname{diag}\left\{\left(\mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{M}'\left[\left(\mathbf{I}_T \otimes \boldsymbol{w}' \otimes \mathbf{1}_B\right)'\mathbf{D}_j\left(\mathbf{I}_T \otimes \boldsymbol{w}' \otimes \mathbf{1}_B\right)\right]^{-1}\mathbf{M}\left(\mathbf{M}'\mathbf{M}\right)^{-1}\right\}^{\frac{1}{2}} . \tag{18}$$

Note that these standard errors are based on prior estimates of the API components specific to each time point and test, *i.e.*, from Equation (15). They are too large because the model degrees of freedom is TK as opposed to T if we are estimating the T annual APIs. If we treat $\boldsymbol{w}$ as weights in the linear model, a more direct estimate of the API for each year is

$$\mathbf{H}_j^*(\mathbf{1}_T \otimes \mathbf{1}_K \otimes \boldsymbol{v}) = \mathbf{H}_j^*(\boldsymbol{\mu}_j^* \otimes \mathbf{1}_B) + \mathbf{H}_j^*(\mathbf{1}_T \otimes \mathbf{1}_K \otimes \boldsymbol{e}_j^*) . \tag{19}$$

The frequencies are now weighted as $\mathbf{H}_j^* = \mathbf{H}_j \times (\mathbf{I}_T \otimes \boldsymbol{w}^* \otimes \mathbf{I}_B)$, and $\boldsymbol{w}^*$ is a diagonal matrix containing the square-root of elements in $\boldsymbol{w}$. The resulting solution takes a form similar to Equation (11). Finally, for the growth model (16), we have

$$\hat{\boldsymbol{\beta}}_j = \left[(\mathbf{M} \otimes \mathbf{1}_K \otimes \mathbf{1}_B)'\mathbf{D}_j^*(\mathbf{M} \otimes \mathbf{1}_K \otimes \mathbf{1}_B)\right]^{-1}(\mathbf{M} \otimes \mathbf{1}_K \otimes \mathbf{1}_B)'\mathbf{D}_j^*(\mathbf{1}_T \otimes \mathbf{1}_K \otimes \mathbf{I}_B)\boldsymbol{v} , \tag{20}$$

where $\mathbf{D}_j^* = \mathbf{H}_j^*\mathbf{H}_j^{*\prime}$. The standard errors are smaller relative to (18), as they are now based on the 2 degrees of freedom for the growth model. We next provide estimates in an example of a school with 4 waves of data.

Example 4 (Calculating and Modelling API_{jk} over Time). In the following data from an elementary school, `test`, `time`, `p`, `f`, and `v` are the variables for the tests, time, proportions, frequencies, and values respectively. `N` defines the total number of students at each time point. `t` codes the variable for linear growth, centered at the first time point. The first block of SAS PROC GLM code produces the following $\hat{\mu}_{jk}^{(t)}$ (s.e.) estimates:

```
Test \ Time         1               2               3               4
-----------   --------------  --------------  --------------  --------------
Language      607.50 (22.39)  607.50 (21.35)  645.00 (20.44)  725.00 (22.39)
Mathematics   637.50 (18.28)  675.00 (19.64)  742.50 (18.28)  692.50 (20.44)
Reading       655.00 (22.39)  575.00 (20.44)  772.50 (20.44)  742.50 (22.39)
Spelling      755.00 (21.35)  692.50 (19.64)  625.00 (21.35)  725.00 (21.35)
```

Weighting tests produces the API estimates (s.e.), $\hat{\mu}_j^{(t)}$: 655.87 (10.96), 637.50 (10.87), 719.25 (10.52), and 717.25 (11.55) for time points 1 to 4, respectively (obtained by the lines beginning with the command "estimate"). In terms of the school growth on the API, $\hat{\beta}_j$, we estimate that the school API is 642.59 (9.18) at time point 1 and over the time span the API grows linearly at about 26.59 (5.01) a year. Given the pooled error variance of $\hat{\sigma} =$ 223.88, the school API is growing at an annual rate of .12 standard deviation units. The alternative solution, given by Equations (19) and (20), is implemented in the second block of SAS PROC GLM code. t recodes the variable `time`, centering it at time point 1. For each test is the corresponding PSAA weight, w. As expected, the standard deviation estimate is appreciably reduced, to 115.45. The estimates and their s.e.'s are 642.65 (8.77) and 27.23 (4.78). While the location parameter estimates are roughly equivalent, they are now determined with more precision. When compared with background variation, the school has been growing at a relatively larger estimate (about twice) of .23 s.d.'s a year. □

Data and SAS PROC GLM code for Example 4

```
data;
 input time test $ N p1-p5 @@;
 t=time-1;
      if test="L" then w=.15;
 else if test="M" then w=.40;
 else if test="R" then w=.30;
 else if test="S" then w=.15;
 array p p1-p5;
 array v v1-v5 (200 500 700 875 1000);
 do over p;
  value=v;
   f=int(p*N);
   output;
 end;
 drop v1-v5 p1-p5;
cards;
1 L 100 .2 .2 .4 .1 .1 1 M 150 .1 .3 .4 .1 .1 1 R 100 .1 .3 .3 .2 .1 1 S 110 .0 .2 .4 .2 .2
2 L 110 .2 .2 .4 .1 .1 2 M 130 .1 .2 .4 .2 .1 2 R 120 .2 .3 .3 .2 .0 2 S 130 .1 .2 .3 .3 .1
3 L 120 .2 .1 .4 .2 .1 3 M 150 .1 .1 .3 .3 .2 3 R 120 .0 .2 .3 .3 .2 3 S 110 .1 .3 .4 .2 .0
4 L 100 .1 .1 .4 .2 .2 4 M 120 .1 .2 .3 .3 .1 4 R 100 .1 .1 .3 .3 .2 4 S 110 .1 .1 .4 .2 .2
;
proc glm;                /*** Estimating Linear Composites ***/
 class time test;
 freq f;
 model value=time*test / noint solution;
  estimate "API1" time*test .15 .4 .3 .15 0 0 0 0 0 0 0 0 0 0 0 0;
  estimate "API2" time*test 0 0 0 0 .15 .4 .3 .15 0 0 0 0 0 0 0 0;
  estimate "API3" time*test 0 0 0 0 0 0 0 0 .15 .4 .3 .15 0 0 0 0;
  estimate "API4" time*test 0 0 0 0 0 0 0 0 0 0 0 0 .15 .4 .3 .15;
  estimate "Int" time*test 0.105 0.28 0.21 0.105 0.06 0.16 0.12 0.06
                           0.015 0.04 0.03 0.015 -0.03 -0.08 -0.06 -0.03;
  estimate "Slope" time*test -0.045 -0.12 -0.09 -0.045 -0.015 -0.04 -0.03 -0.015
                              0.015 0.04 0.03 0.015 0.045 0.12 0.09 0.045;

proc glm;               /*** Linear Fixed Effects Growth ***/
 freq f;
 weight w;
 model value=t / solution;
```

3.3 Measuring Progress the PSAA Way.

The above characterization of growth may not be immediately useful when the number of time points are small and an "average gain," *i.e.*, a growth rate, is not the object of inference. Adequate annual progress, as defined by the PSAA, amounts to a 5% increase on the school API estimate from one year to the next, when compared with how far below the school was in the previous year from the statewide target of 800. In terms of the estimated school API for the time points t and $t-1$, from Equation (14), annual growth is estimated by the ratio

$$\lambda_j^{(t)} = \frac{\mu_j^{*(t)} - \mu_j^{*(t-1)}}{800 - \mu_j^{*(t-1)}} \ . \tag{21}$$

Suppose that $T = 4$ and we are interested in the API growth from $t = 1$ to $t = 2$, or $\lambda_j^{(2)}$. Because the numerator is the linear contrast $\boldsymbol{c}_1'\hat{\boldsymbol{\mu}}_j^*$, where $\boldsymbol{c}_1' = [\ -1,\ \ 1,\ \ 0,\ \ 0\ ]$, and the denominator is $800 - \boldsymbol{c}_2'\hat{\boldsymbol{\mu}}_j^*$, where $\boldsymbol{c}_2' = [\ 1,\ \ 0,\ \ 0,\ \ 0\ ]$, we see that

$$\hat{\lambda}_j^{(2)} = \frac{\boldsymbol{c}_1'\hat{\boldsymbol{\mu}}_j^*}{800 - \boldsymbol{c}_2'\hat{\boldsymbol{\mu}}_j^*} \ .$$

Writing $\mathbf{C} = \begin{bmatrix} \boldsymbol{c}_1' \\ \boldsymbol{c}_2' \end{bmatrix}$, the variance-covariance matrix of the numerator and denominator contrasts are

$$\text{Var}\left(\mathbf{C}\hat{\boldsymbol{\mu}}_j^*\right) = \hat{\sigma}^2 \ \times \ \mathbf{C}\left[\left(\mathbf{I}_T \otimes \boldsymbol{w}' \otimes \mathbf{1}_B\right)' \mathbf{D}_j \left(\mathbf{I}_T \otimes \boldsymbol{w}' \otimes \mathbf{1}_B\right)\right]^{-1} \mathbf{C}' \ . \tag{22}$$

Standard errors for the nonlinear parameter $\hat{\lambda}_j^{(t)}$ cannot be obtained as directly, however, although a commonly proposed asymptotic estimator may be derived using the *δ-method* (see Appendix. Letting

$$\text{Var}\left(\mathbf{C}\hat{\boldsymbol{\mu}}_j^*\right) = \begin{bmatrix} \hat{\sigma}_{11}^2 & \hat{\sigma}_{12} \\ \hat{\sigma}_{21} & \hat{\sigma}_{22}^2 \end{bmatrix}$$

for any $t > 2$, then its standard error, s.e. $\left[\hat{\lambda}_j^{(t)}\right]$, is given by

$$\sqrt{\frac{\left(800 - \hat{\mu}_j^{*(t)}\right)^2}{\left(800 - \hat{\mu}_j^{*(t-1)}\right)^4}\,\hat{\sigma}_{11}^2 + 2\frac{\left(800 - \hat{\mu}_j^{*(t)}\right)}{\left(800 - \hat{\mu}_j^{*(t-1)}\right)^3}\,\hat{\sigma}_{12} + \frac{1}{\left(800 - \hat{\mu}_j^{*(t-1)}\right)^2}\,\hat{\sigma}_{22}^2} \ . \tag{23}$$

For large samples, and if the coefficient of variation of the denominator, $\hat{\sigma}_{22}/\hat{\mu}_j^{*(t-1)}$, is less than 10%, this approximation seems reasonably accurate (Cochran, 1977, p. 166).

Example 5 (Ratio of Gain to Distance from Target). We return to the data in Example 4 above, supposing that $n_{jk}^{(t)} = 100$ for simplicity. The APIs for years 1 through 4, according to Equation (14), are 655.875, 637.5, 719.25, and 717.25, respectively. Because the numbers of observations are equal for each test and year, the standard errors for the APIs are uniformly 11.336. (See columns `API` and `Var` in table to follow.) Annual API estimates are also uncorrelated. We estimated the coefficients of variation for years 1 through 3, listed under column `CV`. These are generally in the 1% to 2% range, suggesting that the approximated standard errors for the ratios are acceptably precise. Applying Equations (21) and (23), the ratios of gain for years $t > 2$ over distance of the school's API at time $t-1$ through 3 from the target API of 800, $\hat{\lambda}_j^{(t)}$, and their standard errors, all expressed as percentages, are given in columns `Lambda` and `SE(Lambda)` below.

```
Year    API      Var     CV    Lambda  SE(Lambda)
----  -------  ------  -----  ------  ----------
 1    655.875  128.50  0.017
 2    637.5    128.50  0.018  -12.75    18.938
 3    719.25   128.50  0.016   50.31    11.003
 4    717.25   128.50           -2.48    31.857
```

An important point to note from these results is that, even when the standard errors of the APIs are equal from year to year, the standard errors of their ratios are far less predictable. As can be seen from Equation (23), the sampling variance of ratios of APIs depends on weighted linear composites of the ratios of powers of the API estimates themselves. □

Even as we are able to approximate the standard errors for quantities such as the API growth ratio, its properties need close scrutiny in practice. We will merely list two major concerns here and return to them in Section 7, where we consider better procedures for making inferences regarding ratios of estimates.

First, as the denominator becomes small, ratio estimators such as Equation (21) becomes increasingly inchoate. In our case, such situations arise when schools have initial APIs that are close to the 800 target. When its absolute value is less than 1.0, the ratio is wildly unstable. At zero, it is undefined. When it is negative, the meaning of the resulting ratio bears little resemblance to what is originally intended. Currently, all these difficulties are "avoided" by appealing to various administrative rules that essentially treat schools that are "close" to the target as also meeting the growth target. Second, even when the denominator is reasonable, the approximated standard errors may be large, and they may vary widely from one year to the next, as can be seen in Example 5. Taken together, these considerations suggest that decisions based on these ratio estimates are to be made only with extreme caution. Drawing on Thum (2002), we suggest a remedy in Section 7.

3.4 Missing Test Data

Although the API is not defined for individual students under the PSAA, an API-like index may be calculated for the individual student if the student has her full complement of test scores. It is easy to show that with the full complement of subject matter test scores, the student-level API is exactly its school-level counterpart in that when we apply the API definition at the student-level, the school-level API is exactly the mean of student APIs. However, if a student does not have a test score for, say, Language, his API is not defined (at least not one that is consistent in the sense given above) while the definition of the school-level API is unaffected.

For many analysts, that a student API for students with missing scores is undefined appears as an inconsistency to be remedied. It is sometimes suggested that, when one or several component scores are missing, a student API can still be calculated by first (conveniently) assigning missing scores the weight of zero and then re-distributing the weight for each component at the school-level among the available student scores.

To better understand this proposal, suppose an API-like index weighs its two components, Test A at w_A and Test B at w_B, with $w_A + w_B = 1$. Individual student scores on Test A will be weighted w_A/N_A while scores on Test B will be weighted w_B/N_B, if N_A and N_B are the number of valid scores for Test A and Test B respectively. The goal here is to ensure that $\sum w_A/N_A = w_A$ and $\sum w_B/N_B = w_B$ over students. As

$$\frac{w_A/N_A}{w_B/N_B} \neq \frac{w_A}{w_B}$$

unless $N_A = N_B$, this definition of the student index will be dictated purely by happenstance. For student without valid scores on both Test A and Test B, the ratio of subject matter weights will generally not be $w_A : w_B$ respectively as required but $(w_A/N_A) : (w_B/N_B)$. This approach effectively redefines the mandated subject matter weighting of w_A for Test A and w_B for Test B, rendering the procedure highly unpredictable because it is now sensitive to the specific missing data pattern observed. It alters the weights given to the different subjects matter components even for the students with the full complement of test scores. It also weighs students unequally, depending on their specific missing data pattern when calculating the school aggregate. Because such attempts to provide an estimate for the student index in the presence of missing scores on some components are not in keeping with the original API definitions, they are seriously misguided.

This less-than-satisfactory aspect regarding the (lack of) definition of the student API reveals a significant flaw that arises whenever, as for the API, school progress is defined independently of student growth, a point to be further elaborated when we suggest some alternative approaches to measuring school productivity in the next sections.

4 What the API Measures

At the risk of appearing as if we have put "the cart before the horse," we nevertheless ask: Now that we know how to calculate and also how to make statistical inferences with the school API, what does it all mean? Let us suppose for now that an API component summarizes in some sensible way an aggregate attainment level of its students. Then the key to interpreting the construct being measured by an API component lies with the value assignment approach integral to this index. By assigning a numerical value to the student NPR the API has effectively altered a normative test score to reflect the subjective worth of ordered levels of student normative attainment. Generically, indicators such as the API component produce a subjective scale based on normative performance. Its subsequent employment in various arithmetic comparisons, within and between schools, also reveals the presumption that the new outcome variable is measured on an interval scale.

Many educators will no doubt lament that the new performance scoring scheme has little to say about the level of content or curriculum attainment, a prerequisite of most alternative systems. The API scheme appears to have more in common with the mechanics of commodity pricing, or with earlier developments in the construction of social welfare functions. However, while this approach may represent a critical design flaw for measuring content-anchored learning levels, the goal of formulae like the API is different. It seeks only to depict normative status. Additionally, by focusing on attainment in terms of norms, the API procedure allows for further combining normed achievement results from different tests disregarding content and grade level. The results from any candidate for inclusion need only be expressed in terms of normative attainment. Thus, the API scoring scheme is not an accident of design but a deliberate template to facilitate, by extension, the "rolling in" of other tests and indicators.[2]

Suppose that we accept the proposition that the school API summarizes the normative attainment level of its students; it is then fair to ask: *In what sense does the API measure student and school normative attainment?* The answer to this question turns principally on a set of design issues

[2] Currently, the API rests on student results on the Stanford 9, Form T, which is part of the States Standardized Testing and Reporting (STAR) program. Its mature form will, however, include additional components such as results on the California Standards Tests and the High School Exit Examination, as well as student graduation and attendance rates. By law, test scores are to make up 60 percent of this index. Even more inclusive composite indices have been proposed elsewhere (see Rothstein, 2000).

that has received considerable attention in the methodological research literature on the measurement of growth and change over the past four decades. As defined by the PSAA, the API is a cross-sectional statistic. In an early investigation, which compared alternative designs for tracking progress in schools, Dyer, Linn, and Patton (1969) anticipated the current assessment design concerns regarding cross-sectional analyses of longitudinal data. They concluded that, conceptually, the most appropriate method for measuring growth employed a matched-longitudinal design. In conventional regression terminology, an average of within (student) slopes appeared most promising for yielding useful information about growth when compared with slopes from the total regression slope, or the between regression slope (see also Bryk & Raudenbush, 1992, Table 5.9).

An updated rendition of these concerns may be located in the methodological literature on the measurement of growth and change, and their progressive resolution aided by the application of multilevel modelling. Here, we will not attempt a detailed recounting of the issues as they apply to the measurement of student and school progress. The reader may consult Thum (2002) for a brief survey of the conceptual and statistical issues on the measurement of change as it relates to accountability modelling drawn from early work by, for example, Cronbach and Furby (1970), Rogosa *et al.* (1982), Rogosa and Willett (1985), Willett (1988), Rogosa (1995), and on more recent contributions by Collins (1996), Williams and Zimmerman (1996, Maris (1998), Mellenbergh and van de Brink (1998), Mellenbergh (1999), Raykov (1999)). For further discussions framed around the measurement of school performance, see Willms and Raudenbush (1989), Willms (1992), and Meyer (1996).

Earlier investigations of school and student performance had also stalled until it was recognized that, by design, students are nested within schools and, consequently, statistical analyses should no longer routinely assume that student observations are independent within the school (see Cronbach, 1976; Burstein, 1980a; Burstein, 1980b; Bryk & Raudenbush, 1992, pp. 1–3). The literature on applying multilevel modelling, hierarchical linear modelling, or mixed-effects modelling to measuring student outcomes and school effects has since exploded (Raudenbush & Bryk, 1986; Raudenbush, 1988; Willms and Raudenbush, 1989; Gray *et al.* 1995; Willms, 1992; Goldstein & Spiegelhalter, 1996). Thum (2002) suggested that the recent educational research on monitoring student learning and school productivity for accountability purposes generally (1) endorses a multilevel approach that, within a unified model, allows for simultaneously modelling student and school variability and leads also to a more accurate assessment of uncertainty, and (2) favors defining growth as intra-individual change – as measured by some model of student gain (Sanders & Horn, 1994; Thum & Bryk, 1997; Bryk, Thum, Easton, & Luppescu, 1998; Thum, 2002) rather than relying on an indirect and questionable measure of gain in the predicted residual gain score (Gray *et al.* 1995; Meyer, 1996; Harker & Nash, 1996; Webster & Mendro, 1997). Analytically, we think of a value-added model for school assessment data in terms of its two essential components. First, we measure improvement beginning with a model for student gains. Then, we contextualize student or school improvement in a model with relevant student, family, teacher, cohort, and school correlates.

Collectively, therefore, the literature would caution that, by mistaking cross-sectional school aggregates for representative student behavior, indices such as the school API ignore the important conceptual incongruence between student progress and school productivity, with serious interpretive consequences. Changes in indicators such as the API for a school from one year to another do not substitute adequately for indicators of student growth, even when we restrict the analysis, as is currently the case for the API, to only students who have been in the school or district for two consecutive years. A simple illustration serves to clarify the basic problems. Table 1 presents the API assessment data for a three-year longitudinal cohort of students who attended an elementary

Table 1. API performance data for a LBUSD elementary school.

		With Missing Tests (N=4192)				No Missing Tests (N=4060)			
Year	Worth	Lang	Math	Read	Spell	Lang	Math	Read	Spell
1998	200	113	93	124	118	97	77	117	100
	500	59	45	48	46	55	42	44	41
	700	23	42	29	24	22	39	29	23
	875	24	32	10	19	23	31	10	19
	1000	7	15	4	21	7	15	4	21
1999	200	137	117	173	166	134	112	170	161
	500	90	77	85	74	88	76	84	73
	700	61	68	58	49	59	65	58	47
	875	44	57	23	37	44	57	23	37
	1000	23	38	13	30	23	38	13	30
2000	200	169	116	177	150	166	113	176	147
	500	113	117	132	104	112	116	132	104
	700	82	91	77	84	81	89	76	84
	875	68	87	49	82	67	86	49	80
	1000	38	59	30	51	37	59	30	48

school in the Long Beach Unified School District (LBUSD) in 2000. Columns 3-6 display all available scores over the years from 1998 to 2000 for SAT 9 Language, Mathematics, Reading and Spelling for students in the cohort. Columns 7-10 are tallies for those students with a full complement of test scores. Estimates and standard errors for the annual school APIs and their differences using all available scores are given in columns 2-5 in Table 2. These fixed-effects analyses assumed that for the same period all students gained the same amount, however unlikely this seemed in practice. In standard regression terminology, they represent estimates from the total regression, which treats all observations as independent and any individual differences in gains as random error. Note that with missing test data, a student API is not defined and only the school API conforms to the PSAA definition. Although a student analog to the school API may be defined when there is no missing data (columns 6-9), its school mean, it is argued, still does not represent how much students in the school have progressed on average because individual student gains are not distinguishable in the model. Differences in mean attainment from one year to another can yield a very different picture than the average gain made by an identified (matched) cohort of individuals. Again, in regression terminology, a conceptually more congruent indicator of overall student progress is the school average of student within regression slopes.

Foreshadowing the multilevel modelling which we will recommend for analyzing API growth in the following sections, columns 10-11 in Table 2 display the means and their standard errors that estimate how much students have individually gained on average on the student API defined and tracked over time for each student. As is evident, the average of student gains over time, 55.91 and 37.48 for 1999 and 2000 respectively, can be quite different from gains computed from the averages at each time point, 39.41 and 50.29. These results, stemming form differences in design and analysis, *i.e.*, between a fixed-effects (Panel (a)) and a random-effects (Panel (b)) model, are reproduced in Figure 1.

Table 2. API Estimates and SE for elementary school data given in Table 1.

	With Missing Tests				No Missing Tests					
	Status		Gains		Status		Gains		Gains†	
Year	Est	se	Est	se	Est	se	Est	se	Est	se
1998	445.56	9.34	445.56	9.34	457.90	9.81	457.90	9.81	454.28	14.23
1999	494.91	7.41	49.35	11.92	497.31	7.51	39.41	12.35	55.91	12.69
2000	548.33	6.45	53.42	9.82	547.60	6.51	50.29	9.94	37.48	7.82

† *Random effects model results.*

5 Measuring Student and School Progress in Long Beach USD Using the API

In the following sections, we explore a multivariate multilevel model for measuring student and school productivity in terms of the API similar to that employed by (Thum, 2002) for modelling growth in scale scores[3] for a set of elementary schools in Arizona. In this analysis, we are interested in estimating the amount of growth in the school API for a cohort of students who are in the Long Beach Unified School District in 2000. We will employ only student scores with the full complement of subject matter test results.[4] In all, we used 389,184 test scores from 41,920 students nested within 69 LBUSD schools. Considering the size of this problem, and we have in mind an even larger analysis involving every student the entire state system over a period of time, we adopted a two-stage procedure similar to Raudenbush, Fotiu, and Cheong (1999).

The initial step is to estimate a multivariate two-level model for each school. In our within-student repeated measures model, we estimate a student's API in 1998, and his API gain in 1999 and in 2000, treating test scores for each student as correlated over time. At the student-level, we pool student API estimates to arrive at an estimate for the school. We provide a computational strategy for estimating the school API, μ_j, with some suggestions for a SAS PROC MIXED specification.

The second step performs a random-effects multivariate meta-analysis employing the school-level 1998 API estimates, its 1999 and 2000 API gains as summary inputs, along with their precision estimates.[5] Although our analysis shares similar motivation as described by Raudenbush, Fotiu, and Cheong (1999), our present study does not attribute the variability in school performance to various school-level factors. Furthermore, we take this Bayesian turn in our analysis principally as a vehicle for effecting simple and direct estimation of functions of parameter estimates, such as the PSAA productivity ratio, $\lambda_j^{(t)}$, given by Equation (21). We will show how our analysis enables straightforward inferences on school API status and gain estimates, and their ranks, in addition to the all-important API productivity ratio, $\lambda_j^{(t)}$.

[3] Though a more generic term, scale scores here refer to scores based on item-response theory models.

[4] The issue of missing tests will require a separate and lengthier treatment. When a test is unavailable, it gets at the very definition of the API itself. As such, it cannot be mistaken for a missing data problem.

[5] See Smith *et al.* (1995) and Normand (1999) for a general discussion of the Bayesian approach to random-effects meta-analysis.

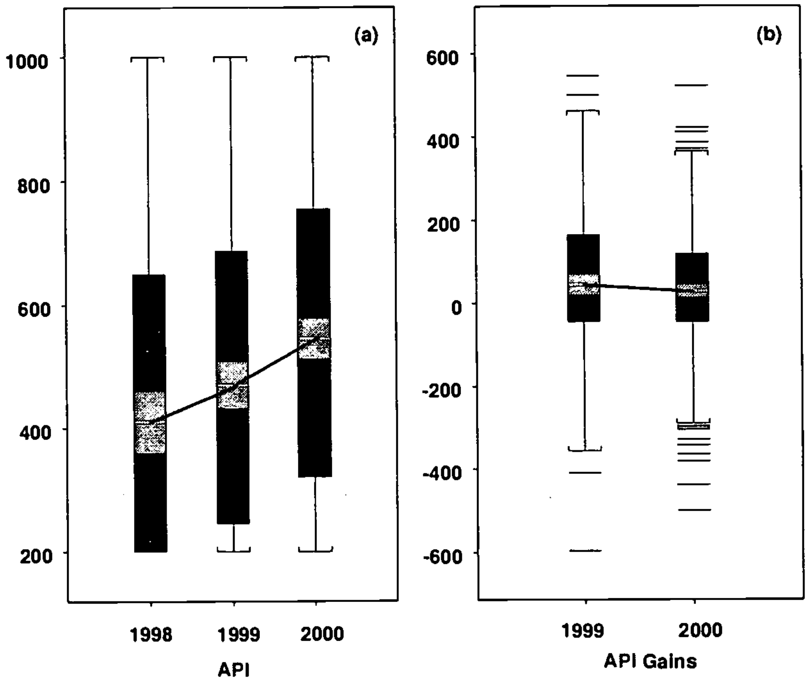

Fig. 1. Comparing school API estimates can be misleading for understanding growth.

5.1 Step 1: A Multivariate Multilevel Model for the School API, μ_j

To facilitate our rather abbreviated discussion of multivariate multilevel modelling, we employed as much as possible the notation of Thum (1997).[6] If student i in school j scores $v_{ijk}^{(t)}$ on test k at time t, then we may estimate the student API, $\pi_{ij}^{*(t)}$, by weighting each Equation (24)

$$\textbf{Student Growth Model}: \qquad v_{ijk}^{(t)} = \pi_{ij}^{(t)} + e_{ijk}^{(t)} \tag{24}$$

by its corresponding subject matter weight, $\sqrt{w_{ijk}^{(t)}}$. The residual, $e_{ijk}^{*(t)}$, is often assumed to be identically and independently distributed normal, or $e_{ijk}^{*(t)} \sim \mathcal{N}(0, \theta_j^2)$. In an application with multiple outcomes, we may also consider alternative error structures. Assuming that the error variances are heterogeneous and that the errors are correlated over time, for example, we may set

$$e_{ij}^{*} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K \otimes \mathbf{\Theta}_j) \ ,$$

where $\mathbf{\Theta}_j$ is the unstructured $T \times T$ matrix. When some of the time points are unobserved, rows and columns in $\mathbf{\Theta}_j$ will be deleted accordingly, leading to patterns of $\mathbf{\Theta}_{ij}$ that are unique to the

[6] Thum (1997) gave a detailed development of a multivariate generalization of the standard two-level multilevel linear model that essentially replaced the univariate level one mixed-effects model with a multivariate mixed-effects analysis of covariance model. Other sources, from the Bayesian hierarchical modelling perspective, include Congdon (2001, ch. 8) and Raudenbush and Bryk (2001, ch. 13). Recently, Yang *et al.* (2001) employed a multilevel multivariate model for examination results to assess the effects of self-selection on exam subjects by individuals.

student (see the discussion in Thum (1997, pp. 82–83) regarding the treatment of missing outcomes when the errors take a more general structure than *i.i.d.*).

Students within each school j contribute to an estimate of the school API, which we now denote as $\mu_j^{(t)}$, for each time point t as given by

$$\textbf{Students within School}: \qquad \pi_{ij}^{*(t)} = \mu_j^{(t)} + u_{ij}^{(t)} \; . \tag{25}$$

The school-level residuals, $\boldsymbol{u}_{ij}$, are assumed to be identically and independently distributed T-variate normal, $\boldsymbol{u}_{ij} \sim \mathcal{N}_T(\mathbf{0}_T, \boldsymbol{\Upsilon}_j)$.

For longitudinal data, we may also model student scores with, for example, a design for linear growth such as implied by Equation (16). For our illustration with three time points, it will be more useful to employ a design on time that provides direct estimates of the student API at time point 1, the gain between time point 2 compared with time point 1, and the gain from time point 2 to time point 3. Thus, setting $\boldsymbol{\pi}_{ij}^* = \mathbf{M}_{ij}\boldsymbol{\beta}_{ij}$, the unknown parameters, $\boldsymbol{\beta}_{ij}$, now correspond to each columns in

$$\mathbf{M}_{ij} = \begin{pmatrix} 1\,0\,0 \\ 1\,1\,0 \\ 1\,1\,1 \end{pmatrix} \; .$$

The alternative school-level model may be written as

$$\textbf{Students within School}': \qquad \boldsymbol{\beta}_{ij} = \boldsymbol{\gamma}_j + \boldsymbol{r}_{ij} \; , \tag{26}$$

with $\boldsymbol{r}_{ij} \sim \mathcal{N}_3(\mathbf{0}_3, \boldsymbol{\Psi}_j)$. It may then be shown that $\boldsymbol{v}_{ij}$ is marginally distributed multivariate normal as

$$\boldsymbol{v}_{ij} \sim \mathcal{N}(\mathbf{M}_{ij}\boldsymbol{\gamma}_j, \boldsymbol{\Sigma}_j) \tag{27}$$

where $\boldsymbol{\Sigma}_j = \mathbf{M}_{ij}\boldsymbol{\Psi}_j\mathbf{M}'_{ij} + \boldsymbol{\Theta}_{ij}$. Using standard results for the multilevel model, *e.g.* from Thum (1997), the estimated profile of API growth factors for school j is given by

$$\hat{\boldsymbol{\gamma}}_j = \left[\sum_{i=1}^{n_{ij}} \mathbf{M}'_{ij}\widehat{\boldsymbol{\Sigma}}_j^{-1}\mathbf{M}_{ij} \right]^{-1} \sum_{i=1}^{n_{ij}} \mathbf{M}'_{ij}\widehat{\boldsymbol{\Sigma}}_j^{-1}\boldsymbol{v}_{ij} \tag{28}$$

with variance-covariance matrix, $\text{Var}(\hat{\boldsymbol{\gamma}}_j)$, or

$$\hat{\mathbf{V}}_j = \left[\sum_{i=1}^{n_{ij}} \mathbf{M}'_{ij}\widehat{\boldsymbol{\Sigma}}_j^{-1}\mathbf{M}_{ij} \right]^{-1} \; . \tag{29}$$

The reader should note that our model specifies unique components of variance for each school j because there is no compelling reason to assume that these structural factors should be identical across the diverse student bodies in the school system.[7]

Example 6 (A Multivariate Multilevel Model for School j). To estimate the 1998 API status and gains in 1999 and 2000 for school 407, we employ the school data file (`Sch407`) containing the following variables: `stuid` identifies the student; `subject` identifies the subject matter; `year` identifies the year; `yr1`, `yr2`, and `yr3` encode the year as given by the matrix $\mathbf{M}_{ij}$;

[7] It is also clear from the previous sections that the analysis here may be easily adapted to provide grade or significant subgroup specific estimates.

score2 is the outcome $v_{ijk}^{(t)}$ weighted by its corresponding weight wgt2, or $\sqrt{w_{ijk}^{(t)}}$. For the data described, the SAS PROC MIXED statements below will produce school estimates $\hat{\gamma}_j$, as well as REML estimates of the variance components, $\hat{\mathbf{\Psi}}_j$ and $\hat{\mathbf{\Theta}}_j$. The option covbi in the model statement prints the estimate for $\hat{\mathbf{V}}_j$. For an introduction to multilevel modelling using PROC MIXED, consult Singer (1998). □

Data and SAS PROC MIXED code for Example 6

```
proc mixed data=Sch407 covtest noclprint=10 noitprint empirical noprofile method=reml;
      class stuid year subject;
       model score2=wgt2*yr1 wgt2*yr2 wgt2*yr3
                  / noint solution covbi ddfm=residual;
       random wgt2*yr1 wgt2*yr2 wgt2*yr3 / sub=stuid type=un;
       repeated year/ sub=subject(stuid) type=un rcorr;
 title1 "1998 API Status and Gains in 1999 and 2000 in School 407";
```

5.2 Step 2: A Bayesian Meta-Analysis

In the previous section, we show how we may obtain the restricted maximum likelihood estimate of $\hat{\gamma}_j$ and its variance-covariance matrix $\hat{\mathbf{V}}_j$ for each individual school j. While it is now useful to regard the estimates $(\hat{\gamma}_j, \hat{\mathbf{V}}_j)$ as providing reasonable information about the "true" performance at each school site, γ_j, the information on individual school performance can be made more precise by viewing individual schools as members in a group or sub-groups. In general, pooling school estimates in a multilevel model produces an improved version of the same school estimates $(\tilde{\gamma}_j)$, termed *empirical Bayes* or *shrinkage* estimates, that, although biased, have minimum mean squared error (deGroot, 1975). The model will also provide an estimate of the average performance of the system of schools. We also gain some assessment of the graded heterogeneity of school performance that is sensitive to the relative precision of the information from individual schools. However, because our group of schools is not a random sample in any sense, like Raudenbush, Fotiu, and Cheong (1999), we would adopt a model aggregation view of this model (also see Spiegelhalter & Marshall, 1998, and Congdon, 2001, ch. 5).

We relate the performance estimate $(\hat{\gamma}_j, \hat{\mathbf{V}}_j)$ for each school j to its "true" but unknown population value, γ_j, using the meta-analysis model

$$\hat{\gamma}_j \mid \gamma_j \sim \mathcal{N}(\gamma_j, \hat{\mathbf{V}}_j) \tag{30}$$

and

$$\gamma_j \mid \zeta \sim \mathcal{N}(\zeta, \mathbf{\Phi})\,. \tag{31}$$

Because we supply the precision estimates via $\hat{\mathbf{V}}_j$ in Equation (30), this otherwise familiar multivariate two-level model is often known as a "v-known" problem.

We employ noninformative priors throughout to reflect our lack of specific prior information that would influence the results of our analysis.[8] Our prior for ζ, which may be interpreted as the performance profile for this collection of LBUSD schools, is accordingly assumed to be uninformative, each of which is distributed as

$$\zeta_s \sim \mathcal{N}(0.0, \mathrm{c}_s)$$

[8] Another compelling perspective for us is that this choice of prior indirectly reveals our interest in obtaining a likelihood-based solution via a Bayesian setup of the problem. For more elaboration, see Wasserman (2000).

for some suitably large constant c_s (so that ζ_s "can be anywhere"). We further assume that the noninformative prior for $\mathbf{\Phi}$, the variance-covariance matrix of the school performance factors $\boldsymbol{\gamma}_j$ in the collection of schools, is an inverse Wishart

$$\mathcal{W}^{-1}(\boldsymbol{\Upsilon}, \nu)$$

with precision $\boldsymbol{\Upsilon}$ on ν degrees of freedom.

We illustrate this computational strategy with a sketch of effective WINBUGS© program specification (Spiegelhalter *et al.* 1999). In this application, the value for ν is 3. Reasonable alternative initial values for $\boldsymbol{\Upsilon}$ are conveniently culled from our set of $\hat{\mathbf{V}}_j$'s. Ten thousand updates give a stable solution. Convergence issues for MCMC and on approaches for assessing convergence are treated by Cowles and Carlin (1995) and Brooks and Roberts (1998). Basic WINBUGS program code for our model is given in Example 7.

Example 7 (A Bayesian Meta-analysis via WINBUGS). Suppose we have the estimates for the 1998 API status and the gains in 1999 and 2000 for each j of M schools stored in a matrix `g[j,1:3]` and the *inverse* of their corresponding variance-covariance matrix stored as `V[j,1:3,1:3]`. We relied on WINBUGS for fitting our model. The Markov chain Monte Carlo (MCMC) approach employed in this program simulates the posterior distribution for each parameter by repeatedly drawing values from the appropriate full conditionals. Within a loop for schools, indexed by `j`, the following statements gives the core of a WINBUGS program:

`WINBUGS code segment for Example 7`

```
# Model section
  model
   {
    for(j in 1:M){
         g[j,1:3] ~ dmnorm(gm[j,],V[j,,]);
        gm[j,1:3] ~ dmnorm(ze[],omega.gm[,]);
         }
# Specify Priors
     for(i in 1:3){
       ze[i] ~ dnorm(0.0,.0001);}
     omega.gm[1:3,1:3] ~ dwish(R[,],3);
       for(k in 1:3){ for(l in 1:3){
        Sig2.gm[k,l]<-inverse(omega.gm[,],k,l); }}
    }
```

The correspondence between terms in the code and our model as specified in Equations (30) and (31) is strikingly clear and so our code requires no further clarification. The important point to note is that, by convention, WINBUGS assumes `omega.gm` to be $\mathbf{\Phi}^{-1}$. Thus our outputs from this specification are simulations from the marginal posterior distribution of $\boldsymbol{\gamma}_j$ in `gm[j,]`, of $\boldsymbol{\zeta}$ in `ze[]`, and of $\mathbf{\Phi}$ in `Sig2.gm[,]`. For an introduction to multilevel modelling using WINBUGS, consult Spiegelhalter *et al.* (1999). □

6 API Growth for the Long Beach USD 2000 Longitudinal Cohort

Table 3 displays the aggregate results from our Bayesian meta-analysis. Note in particular that tabled entries are selected features (mean, standard deviation, and values corresponding to the 2.5%, the 50%, and the 97.5%) of the simulated marginal posterior distribution of each estimate.

Our set of LBUSD schools typically attained an API of 545.80 ($\hat{\zeta}_1$) in 1998. Judging from the range of estimates ±1.0 × SD from the posterior mean, about 68% of LBUSD schools obtained an

Table 3. Results for the Long Beach USD 2000 Cohort.

Parameter		Posterior Estimates				
		Mean	SD	2.5%	50%	97.5%
System Average						
1998 APIs	$\hat{\zeta}_1$	545.80	14.37	517.40	545.90	573.40
1999 API Gains	$\hat{\zeta}_2$	43.39	2.80	37.94	43.34	48.95
2000 API Gains	$\hat{\zeta}_3$	21.93	2.99	15.96	21.95	27.80
Variance Components						
1998 APIs	$\hat{\phi}_{11}$	117.30	10.34	98.83	116.70	139.60
1999 Gains and 1998 Status	$\hat{\rho}_{12}$	-0.31	0.12	-0.52	-0.32	-0.07
2000 Gains and 1998 Status	$\hat{\rho}_{13}$	-0.52	0.09	-0.68	-0.52	-0.32
1999 API Gains	$\hat{\phi}_{22}$	21.02	2.30	16.96	20.87	26.01
1999 and 2000 Gains	$\hat{\rho}_{23}$	0.06	0.14	-0.21	0.06	0.33
2000 API Gains	$\hat{\phi}_{33}$	23.35	2.22	19.49	23.17	28.16

1998 API status between 428.50 ($\hat{\zeta}_1 - \hat{\phi}_{11}$) and 663.10 ($\hat{\zeta}_1 + \hat{\phi}_{11}$). Schools gained 43.39 ($\hat{\zeta}_2$) on the API in 1999. About 70% of the schools gained from 22.27 ($\hat{\zeta}_2 - \hat{\phi}_{22}$) to 64.41 ($\hat{\zeta}_2 + \hat{\phi}_{22}$). The school system then saw a less impressive gain, of about 21.93 ($\hat{\zeta}_3$), in 2000. Gains have about the same spread in 2000 ($\hat{\phi}_{22} = 21.02$) as in 1999 ($\hat{\phi}_{33} = 23.35$) but are typically smaller, ranging from -1.42 ($\hat{\zeta}_3 - \hat{\phi}_{33}$) to 45.28 ($\hat{\zeta}_3 + \hat{\phi}_{33}$). It appears that, in 2000, a fair number of schools, as many as 15%, posted no gain on the API at all.

6.1 Relating Gains to Initial Status: From Associations to Predictions

Is there a relationship between where a school starts out in 1998 and how much it gains in 1999 and in 2000? The scatter plots in Figure 2 suggest that while 1998 API attainment is negatively correlated with 1999 gain, at about $-.31$ ($\hat{\rho}_{12}$ in Table 3), an even stronger relationship, $\hat{\rho}_{13} = -.52$, characterizes school 1998 API status and school API gain in 2000. Moreover, school API gains in 1999 are not predictive of gains in 2000 ($\hat{\rho}_{23} \simeq 0$).

Beyond considering such correlations, we may estimate predictive models of (true) gains from (true) initial status that take into account the uncertainties in our status and gains estimates. Because the estimator for 1998 status and subsequent gains are latent variables, the resulting so-called *latent variable regression* (LVR) model.[9] for predicting 1999 API gains from 1998 API status begins with Equation (30), but replaces (31) with

$$\begin{aligned} \gamma_{j1} &\sim \mathcal{N}(\zeta_1, \phi_{11}) \ , \\ \gamma_{j2} &\sim \mathcal{N}(\xi_{j1}, \phi_{22}) \ , \\ \xi_{j1} &= \zeta_2 + \zeta_4 \times (\gamma_{j1} - \zeta_1) \end{aligned} \tag{32}$$

[9] Recently, Raudenbush and Bryk (2001, pp. 361-364) provided an example of a multilevel model with LVR for clarifying the gender gap in growth rates for mathematics attainment in high schools. Latent variable regression applications in the Bayesian hierarchical modelling framework are given by Congdon (2001, section 8.6.2) and by Seltzer, Choi, and Thum (2001). These possibilities relate directly to earlier suggestions by Raykov (1993) who pointed out that the residual from an LVR with covariates, formulated within the structural equations modeling framework (SEM), may be employed as a residualized *true* gain estimate.

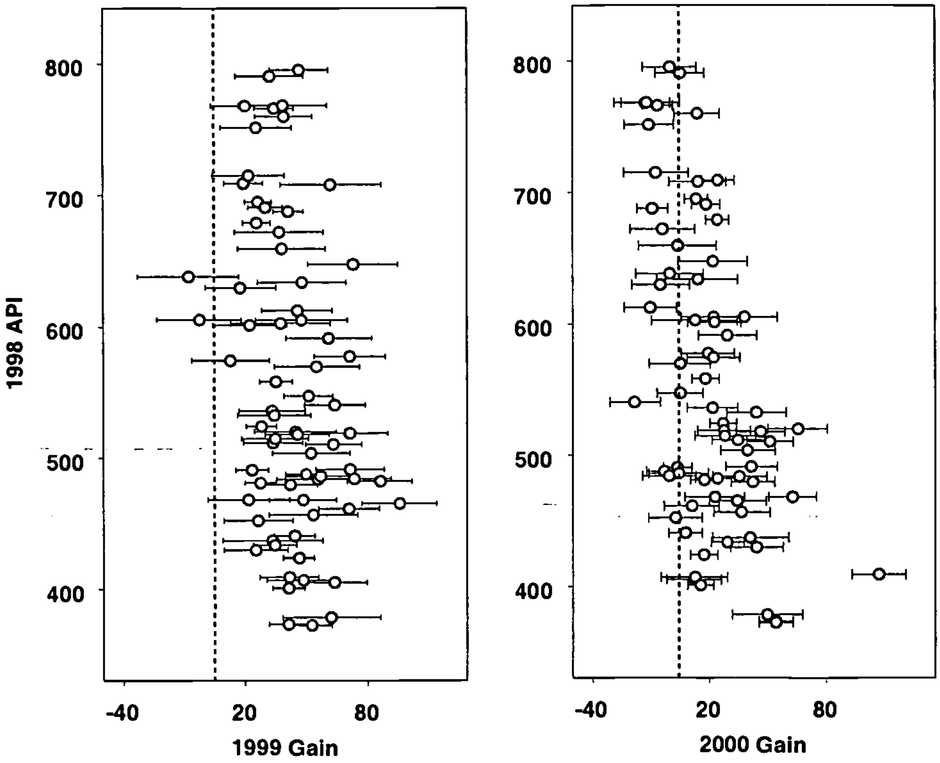

Fig. 2. School API gain estimates and 95% interval estimates in 1999 and 2000 against their 1998 API status.

for regressing 1999 API gains on 1998 API status, and with

$$\begin{aligned}\gamma_{j3} &\sim \mathcal{N}(\xi_{j2}, \phi_{33}) \ , \\ \xi_{j2} &= \zeta_3 + \zeta_6 \times (\gamma_{j1} - \zeta_1) \end{aligned} \tag{33}$$

for predicting 2000 gains using 1998 API status.

In this LVR model, ξ_{j1} and ξ_{j2} represent the new latent outcomes (in place of γ_{j2} and γ_{j3} respectively) conditional on our measure of 1998 API status, γ_{j1}. The model, now comprising Equations (30), (32), and (33), and with uncorrelated residual errors, is also easily implemented in WINBUGS by straightforward elaborations of the code segment provided in Example 7. Our results in Table 4 show that the average API gain in 1999 is 43.47 ($\hat{\zeta}_2$), and for each unit increase in a school's 1998 API status, schools gain at a rate that is about 6% API points less ($\hat{\zeta}_4 = -.06$) on average. Controlling for 1998 status explains only about 4% of the variation in 1999 gains (by stating the variance reduced as a proportion of the parameter variance in the unconditional model). Adjusting for its 1998 API status, a typical school expects to gain at a rate of about a tenth of an API point less for each point increase in its 1998 API ($\hat{\zeta}_5 = -.10$). More importantly, 1998 API status accounts for some 13% of the between school variation in 2000 API gains in 2000.

That initial status has, in this instance, better predicted more distal achievement gains at the school-level when we restrict ourselves to longitudinal student cohorts may very well be an

Table 4. Latent Variable Regression Results for the Long Beach USD 2000 Cohort.

		Posterior Estimates				
Parameter		Mean	SD	2.5%	50%	97.5%
System Average						
1998 APIs	$\hat{\zeta}_1$	545.40	14.32	516.50	545.50	573.10
1999 API Gains	$\hat{\zeta}_2$	43.47	2.84	38.03	43.46	49.14
1998 API Rate	$\hat{\zeta}_4$	-0.06	0.02	-0.10	-0.06	-0.01
2000 API Gains	$\hat{\zeta}_3$	21.99	3.03	16.04	21.98	27.91
1998 API Rate	$\hat{\zeta}_5$	-0.10	0.02	-0.12	-0.10	-0.06
Variance Components						
1998 APIs	$\hat{\phi}_{11}$	118.40	10.59	100.10	117.60	141.60
1999 API Residual Gains	$\hat{\phi}_{22}$	20.40	2.25	16.44	20.27	25.14
2000 API Residual Gains	$\hat{\phi}_{33}$	20.24	1.98	16.72	20.13	24.50

indication that failing schools have worked harder at meeting the system challenge while schools that are performing well initially have less room to grow. Rather than speculate about alternative explanations here, we will leave these intriguing questions for a separate study. In the next sections, we focus instead on some troublesome issues when using our estimates, $\tilde{\gamma}_j$, to characterize the productivity of individual schools.

6.2 Working with School API Estimates: Reliabilities and Comparisons

We will employ our school-level API estimates, $\tilde{\gamma}_j$, in various ways to characterize individual school productivity. Columns 3–8 of Table 5 displays the means and standard deviations of the simulated marginal posterior distributions[10] of the API estimates for 1998, 1999, and 2000 for each school. Means and standard deviations of the simulated marginal posterior distributions for individual school 1999 and 2000 API gains are given in columns 10–11 and columns 13–14 respectively (both means are plotted against the mean for the school 1998 API status in Figure 2 above). Note further that, because we are employing available data for the 2000 cohort, there will be more data at the school-level as the year increases from 1998 to 2000. This aspect of the data design explains why the reported standard deviations get smaller with year. Similarly, our pooled school estimate of within-student gains employs matched student data for every t and $t-1$; thus standard deviations of school gain estimates are smaller in 2000 than in 1999.

In terms of estimates of their 1998 status or their 1999 and 2000 gains, no schools stand out in Table 5 more than School 460.[11] Student who are in this school in 2000 generally started at ($\tilde{\gamma}_{46,1} = 409.3$), gained about 41.66 ($\tilde{\gamma}_{46,2}$) on the API in 1999, and added a very high gain in 2000 of about 107 ($\tilde{\gamma}_{46,3}$). Although the identity of the school is unknown to us, and we do "cut" the data differently, we are able to confirm one such spectacularly performing school among the LBUSD elementary schools in 2000.

[10] The marginal posterior distributions approximate the "sampling distributions" of each school estimate. Thus its standard deviation may be interpreted as the (conventional) standard error.

[11] School 460 is unit 46 on our subscript j.

Reliability. When we work with school-level estimates it is important to question their reliability. Conventional approaches have essentially sketched prototypical estimates for outcomes, drawn from psychometric test reliability studies that are perhaps keyed on general factors such as school size, school income and ethnic composition, etc. While it is important to understand the inherent reliability of our instruments, these estimates are generally irrelevant because our interest goes beyond the question about how predictable a test is for some typical examinee. Assuming we only field tests with reasonable reliability, we also need an estimate of the precision of the individual student score to weight our analysis accordingly (see Bryk *et al.*, 1998; AERA, APA, & NCME, 1999, p. 29). In the case of the API, we need to start the analysis with some assessment of the misclassification errors that are incurred when individuals are recast into the PSAA performance deciles.[12] Unfortunately, this information is not available for our analysis, and we proceed for now on the necessary assumption that the precision of scores is comparable. Our analysis will then be more concerned about the precision of estimates for the school, as conveyed by comparing the variability of individual components with the estimated background variation. In the context of our multilevel model, the reliability of each of the individual school growth factors, $\hat{\gamma}_s$, is measured by

$$\hat{\kappa}_s = 1.0 - \left\{ [\hat{\mathbf{\Phi}}^{-1} + \hat{\mathbf{V}}_j^{-1}]^{-1} \hat{\mathbf{\Phi}}^{-1} \right\}_{ss} , \tag{34}$$

where $\{\mathbf{Z}\}_{ss}$ denotes the s diagonal element of the matrix $\mathbf{Z}$.[13]

Estimates of the reliability of each school gain estimates in 1999 ($\hat{\kappa}_2$) and 2000 ($\hat{\kappa}_3$) are given in column 12 and column 15 of Table 5. Approximate standard errors for the reliability estimates are also available from our analysis. As is evident from Table 5, the reliabilities of these gain estimates are relatively high and uniform, as indicated by their estimated posterior means (SDs) of .81 (.03) and .88 (.02) for 1999 and 2000 respectively.[14] When reliabilities are uniformly high, as is the case with both school gains estimates, shrinkage is relatively uniform. Consequently, we do not expect to see any dramatic re-ordering of schools when we employ the empirical Bayes school estimates, $\tilde{\gamma}_j$, instead of the school estimates, $\hat{\gamma}_j$, from our separate multilevel analysis.

Comparing Schools. Comparing two schools, or subsets of schools, based on a performance component (such as the schools' 2000 API gains) can be done just as easily. Using the MCMC computational approach, we simply monitor and examine the distribution of the difference between $\tilde{\gamma}_{j,3}$ and $\tilde{\gamma}_{j',3}$ for two schools indexed by j and j'. For example, to contrast the performance in 2000 by School 431 with that by School 613, we examine the marginal posterior of $(\tilde{\gamma}_{20,3} - \tilde{\gamma}_{51,3})$. We find that, in 2000, School 431 gained considerably more than School 613 on the API, with a posterior mean at 29.25 and standard deviation at 6.275 (see Panel (a) in Figure 3). Clearly, any contrast between pairs of schools, or between a school and the mean of a previously identified school group, can be similarly obtained. However, when estimating multiple contrasts, the Type 1 error rate for each (pairwise or more) comparison needs to be adjusted in order to hold the average error rate at some pre-specified level. Goldstein and Healy (1995) proposed a procedure that aids visual determination of the significance of a contrast by adjusting the confidence intervals of the schools

[12] Although not always delivered to the user, standard errors of measurement (sem) are available for scale scores. Classification errors require further effort on the part of the user.

[13] See Bryk & Raudenbush (1992, p. 43, equation 3.51) for a brief discussion of the *multivariate reliability matrix*, $\hat{\mathbf{\Phi}}[\hat{\mathbf{\Phi}} + \hat{\mathbf{V}}_j]^{-1}$.

[14] This interpretation of reliability will nonetheless depend, through $\mathbf{\Phi}$, on the particular collection of schools employed in the analysis. It is relatively unproblematic when the set of schools approaches a simple random sample, or the schools make up the "population."

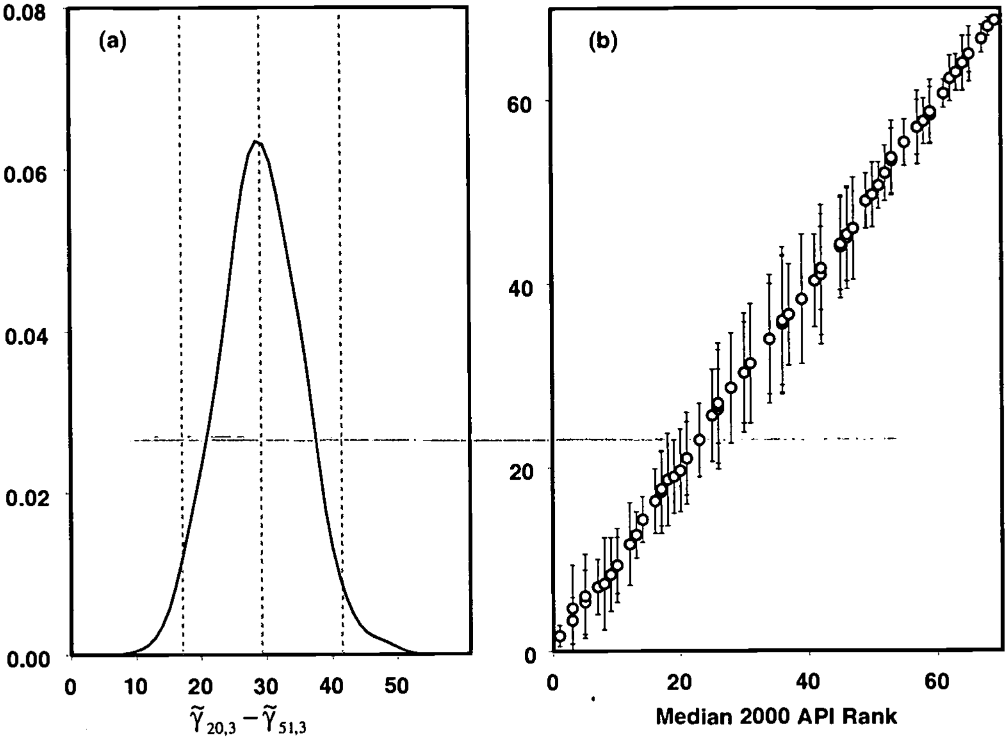

Fig. 3. (a) Comparing School 431 and School 613 on their 2000 API gains via the simulated posterior distribution of $(\tilde{\gamma}_{20,3} - \tilde{\gamma}_{51,3})$. Reference lines mark estimated mean difference at the 2.5%, the mean, and 97.5% points. (b) Ranking (with ties) of LBUSD Schools according to their median APIs in 2000, set within their estimated 95% credibility intervals.

being compared, so that non-overlapping intervals mean a significant difference at the specified level. Again, when employing MCMC estimation, similar confidence statements may be estimated directly by simulating the sampling distribution of the desired contrast.

We also need to caution against the common practice of simply ranking schools according to components of their performance estimates, whether employing $\hat{\gamma}_j$ or $\tilde{\gamma}_j$. Because ranks based on estimates are themselves estimates, they should be treated as such. Arguing as did Laird and Louis (1989), we concur that school rankings should reflect the imprecision of school estimates in the distribution of the ranks. As shown by Goldstein and Spiegelhalter (1996), it is straightforward to estimate the ranking among schools for any of the components in $\tilde{\gamma}$ when using the MCMC approach. As an illustration, we have provided the posterior mean of each school's median rank on the 2000 API status $(\tilde{\gamma}_{j1} + \tilde{\gamma}_{j2} + \tilde{\gamma}_{j3})$ in column 9 of Table 5. Panel (b) in Figure 3 orders each school according to its median rank set within the 95% credibility interval of the school's rank estimates. As with many such orderings, whether by the size of their estimates or their ranks, clear separations between schools on a criterion emerge only when the units are quite distinct. Making explicit the uncertainty in rank estimates will also help to inform rankings when shrinkage among the school estimates is modest.

Quantifying the statistical separation between a randomly selected pair of schools in terms of their ranks will also require adjusting their individual credibility intervals to control for a prespecified Type 1 error rate, for reasons similar to those that Goldstein and Healy (1995) provided for making inferences on multiple comparisons on a continuous outcome. Using MCMC estimation, we may define and monitor the ranks for a subset of schools to achieve similar results. Without such adjusted intervals, schools cannot be judged to be equal in rank when their 95% credibility intervals overlap on a plot such as Panel (b) in Figure 3.

7 Measuring Productivity: The PSAA Ratio vs Productivity Profile

As explained earlier in Section 3.3, the PSAA evaluates school productivity by comparing the productivity ratio estimates, $100 \times \lambda_j^{(t)}$, from Equation (21) to selected percentile targets, say $100 \times \alpha_\ell$, where $\alpha_\ell = .00, .03, .05, .10, .15, .20, .25$, etc. We caution that such a casual use of the productivity ratio may be misleading because any comparisons using estimates should take into account their relative imprecision. We describe two obstacles when we attempt a direct comparison to determine if a school's productivity ratio $\lambda_j^{(t)}$ is at least as large as some α_ℓ.[15] We then suggest the use of a probability statement instead that not only presents how much relative gain a school has made towards to the 800 API target over the year but also an estimate of the level of confidence in our judgement.

We have previously shown in Section 3.3 that when the denominators in the productivity ratio approach zero, taking either small positive or negative values, these ratios are unstable. This has been clearly the case for several of our schools. Columns 2 and 3 in Table 6 provide the posterior means and standard deviations of the PSAA ratio for 1999 and columns 11 and 12 give the estimates for 2000, based on the shrunken estimates $\tilde{\gamma}_j$. In particular, the estimated 1999 productivity ratios for Schools 427, 428, 440, 442, 449, 450, 458, and 628 are unstable. For 2000, in addition to these schools, we also have several instances, such as Schools 440 and 442, where schools lose ground on the API. The precision of ratio estimates is also wildly inflated under these conditions, thus making them unusable for inference.

Leaving aside schools whose base year API status closes on or exceeds the target of 800,[16] we nevertheless find that both the estimated ratios ($\tilde{\lambda}_j^{(t)}$) and their standard deviations (SD($\tilde{\lambda}_j^{(t)}$)) are usable. Under these favorable circumstances, our shrunken ratio estimator, $\tilde{\lambda}_j^{(t)}$, compares reasonably closely with its counterpart, $\hat{\lambda}_j^{(t)}$, which is based on $\hat{\gamma}_j$ from our separate school models. This is clearly depicted in Panel (a) of Figure 4. However, Panel (b) in Figure 4 also suggests that we obtain relatively higher precision if we employ our shrunken estimates, $\tilde{\lambda}_j^{(t)}$. Their standard deviation estimates, SD($\tilde{\lambda}_j^{(t)}$), are, on the whole, systematically smaller than the standard errors SE($\hat{\lambda}_j^{(t)}$) approximated by the $\delta-$method (see Appendix; also Indurkhya *et al.*, 2001).

Even when working only with reasonable estimates ($\tilde{\lambda}_j^{(t)}$, SD($\tilde{\lambda}_j^{(t)}$)), how can a claim that $\lambda_j^{(t)}$ at least meet some α_ℓ be crafted that will take into account the inherent uncertainty of our ratio estimator? Thum (2002), drawing on recent advances by O' Hagan *et al.*(2000, 2001) in analyzing cost-effectiveness ratios for selecting among competing clinical alternatives, suggests that a simple restatement of

$$\text{Probability}\left(\lambda_j^{(t)} \geq \alpha_\ell\right)$$

[15] 5% is the primary threshold for a school to have shown adequate improvement. Other conditions, such as attaining a gain of 5 API points or more, may be similarly treated.

[16] Such cases require a different interpretation of the PSAA productivity ratio and will be treated elsewhere.

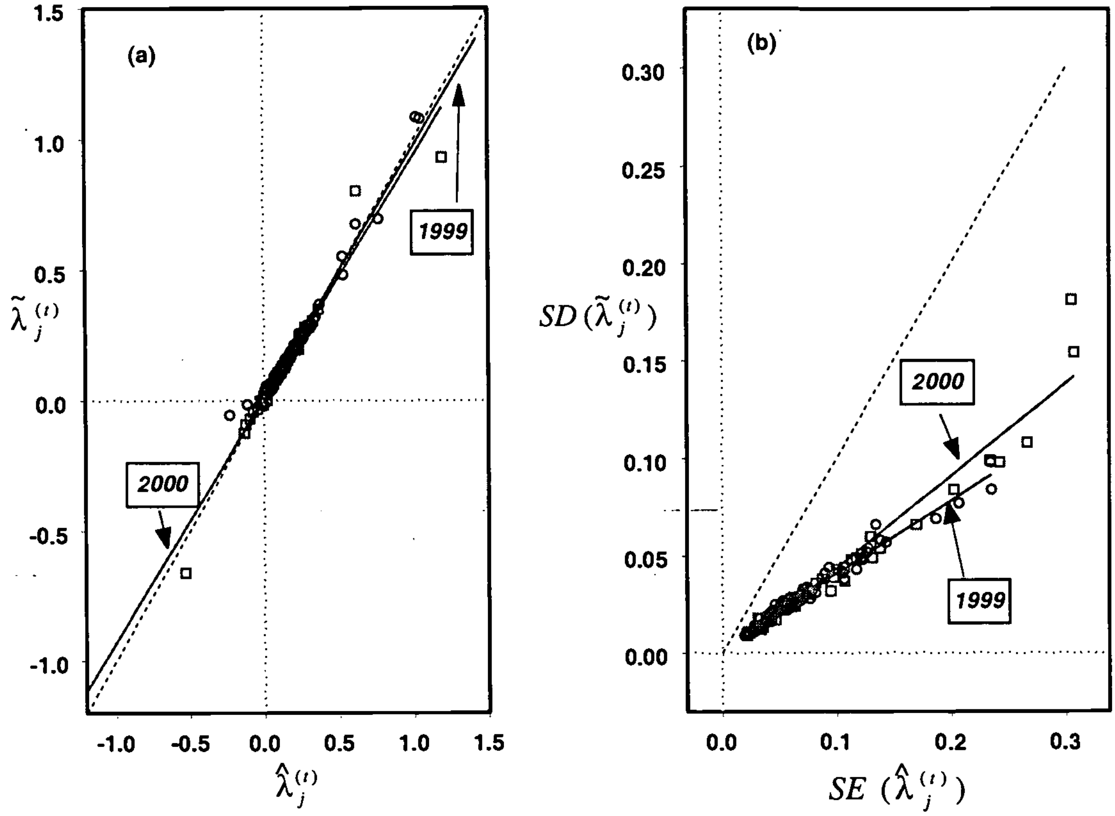

Fig. 4. Estimates of "reasonable" productivity ratios. (**a**) Estimates based on the $\delta-$Method for separate multivariate multilevel HMs for each school, $\hat{\lambda}_j^{(t)}$, are in good agreement with ratios based on the shrunken school estimates, $\tilde{\lambda}_j^{(t)}$, from the meta-analysis. For 1999 (○), these ratios correlate .99 and for 2000 (□), they correlate .97. (**b**) Precision measures do not agree as well, with SE($\hat{\lambda}_j^{(t)}$) being consistently larger than SD($\tilde{\lambda}_j^{(t)}$). (Equivalent estimates will fall about the dotted reference line in each panel.)

is a useful probability statement of the form

$$P_{j\ell}^{(t)} = \text{Probability}\left(\tilde{\gamma}_{jt} \geq \alpha_\ell \times \left[800 - \sum_{s=1}^{t-1} \tilde{\gamma}_{js}\right]\right) , \qquad (35)$$

where $t = 2, 3$ and, under our present parameterization (Section 5.1), $\sum_{s=1}^{t-1} \tilde{\gamma}_{js}$ estimates the API status at time $(t-1)$. Then, Equation (35) simply states the probability that a school's gain estimate is at least as large as some pre-specified fraction of the distance its estimated pre-test is from the 800 API target.[17] Additionally, by varying α_ℓ and graphing the probabilities $\{P_{j,0}^{(t)}, P_{j,0.03}^{(t)}, P_{j,0.05}^{(t)}, \cdots\}$, we generate a *productivity profile*[18] for each school that answers the question: *How much is gained, and at what precision?* Note that the approach is effective whenever we wish to assess whether a difference in the school means exceeds some preset difference, ω, such as for $(\tilde{\gamma}_{j,3} \geq (\tilde{\gamma}_{j',3} + \omega))$ above.

[17] It should be clear that the ratio of any two estimated quantities may be similarly displayed to facilitate their comparison.

[18] It is termed the *acceptability curve* in a cost-benefit analysis.

Table 5. Posterior Means and Standard Deviations of 1998, 1999, and 2000 School API Estimates and Gains in 1999 and 2000 for the Long Beach USD 2000 Longitudinal Cohort, along with School Rank on the 2000 API Status, as well as the Reliability of Gain Estimates, $\hat{\kappa}_{js}$.

		School API Estimates and SD							School API Gains Estimates and SD					
		1998		1999		2000			1999			2000		
School	j	Mean	S.D.	Mean	S.D.	Mean	S.D.	Rank	Mean	S.D.	$\hat{\kappa}_{j2}$	Mean	S.D.	$\hat{\kappa}_{j3}$
404	1	634	16.52	682.1	14.69	696.4	12.81	52	48.03	10.85	0.69	14.38	10.31	0.72
407	2	456.8	13.75	510.4	11.62	547.3	10.97	18	53.55	10.83	0.69	36.87	7.234	0.86
410	3	520.1	12.11	564.9	10.96	630.5	8.744	42	44.8	10.1	0.74	65.55	7.631	0.85
412	4	484.4	11.74	557.8	10.68	557.6	10.47	21	73.42	9.043	0.79	-0.1371	6.724	0.88
413	5	591.5	13.34	652.6	12.24	682.3	11.86	51	61.15	10.47	0.71	29.63	7.523	0.85
414	6	601.5	16.14	623.9	14.45	646.8	14.45	46	22.39	9.146	0.77	22.87	6.933	0.87
415	7	602.9	18.52	640.4	17.02	653.3	15.6	47	37.51	12.29	0.59	12.88	10.91	0.68
416	8	605.4	16.97	653.2	15.24	675.7	14.37	50	47.72	11.27	0.65	22.52	8.352	0.81
417	9	484	12.42	539.9	10.42	575.6	9.521	25	55.91	10.2	0.72	35.73	6.949	0.87
418	10	574.7	14.73	587.5	12.58	609.9	12.57	36	12.76	9.645	0.76	22.42	6.816	0.87
419	11	468.3	12.13	490.1	9.628	552.8	9.363	20	21.71	10.12	0.73	62.71	6.017	0.90
420	12	659.5	18.82	697.7	16.47	701.9	15.93	53	38.2	10.77	0.70	4.185	9.839	0.74
421	13	612.6	15.06	658.4	13.66	648.8	13.79	46	45.78	8.654	0.80	9.626	6.595	0.88
422	14	629.8	14.37	647.5	12.88	643.1	12.16	45	17.75	8.732	0.80	4.469	7.268	0.86
425	15	532.7	14.79	567	14.3	611.5	13.6	36	34.34	9.05	0.78	44.43	7.485	0.85
427	16	768.5	16.94	807.4	14.49	795.9	14.23	65	38.86	10.93	0.69	11.49	8.358	0.81
428	17	795.5	11.93	842.3	9.461	842.5	10.02	69	46.81	7.221	0.86	0.1906	6.763	0.88
429	18	430	10.18	455.3	9.766	499.9	9.42	10	25.31	7.861	0.83	44.61	6.697	0.88
430	19	518.2	12.09	563.9	11.59	610.4	11.35	36	45.65	7.837	0.84	46.54	6.144	0.90
431	20	479.8	9.946	522	8.78	564.7	8.215	23	42.19	6.736	0.88	42.68	5.513	0.92
432	21	482.4	10.25	568.5	9.383	592.9	9.044	30	86.17	8.002	0.83	24.4	5.76	0.91
433	22	518.9	13.16	590.3	11.99	618.2	11.59	39	71.41	9.203	0.78	27.88	6.849	0.87
434	23	570	14.36	625.3	13.2	630.6	13.07	42	55.28	10.39	0.71	5.327	7.659	0.84
435	24	452.4	11.85	478.9	11.37	481.9	11.26	5	26.46	8.504	0.81	3.004	6.756	0.88
436	25	486.2	10.73	543.4	10.08	548.1	9.745	19	57.23	8.726	0.79	4.634	7.539	0.85
437	26	465.3	11.17	561.2	11	595.9	10.54	31	95.85	9.369	0.78	34.72	7.279	0.86
439	27	461.4	10.86	532.4	9.943	543.6	9.613	17	70.97	7.365	0.85	11.18	6.967	0.87
440	28	768.2	11.46	788.4	9.18	776.3	9.09	63	20.17	8.403	0.81	-12.01	6.11	0.90
441	29	647.6	17.02	720.7	14.92	742.9	13.56	59	73.11	11.03	0.69	22.19	8.93	0.78
442	30	790.9	12.23	823.2	10.05	828.5	9.572	68	32.3	8.343	0.81	5.271	6.194	0.90
443	31	511.8	12.64	545.6	11.71	580.7	11.3	26	33.78	7.219	0.86	35.11	6.804	0.88
444	32	715.1	13.74	737	12.5	730.2	12.03	57	21.85	8.91	0.79	-6.821	8.065	0.83
445	33	638.3	18.63	630.3	15.84	630.5	15.03	42	-7.94	12.51	0.62	0.1346	8.511	0.81
446	34	405.3	11.01	469.3	10.74	481.6	10.48	5	63.96	7.824	0.84	12.34	6.972	0.87
447	35	514.9	14.39	549.7	13.15	578	12.38	26	34.77	8.188	0.81	28.34	7.875	0.84
448	36	468.2	11.76	516.7	11.39	539.8	11.06	17	48.52	8.269	0.81	23.04	7.647	0.84
449	37	708.2	19.92	770.4	17.18	784.8	15.96	64	62.19	12.33	0.61	14.45	7.35	0.85
450	38	751.6	13.34	777.3	11.74	766.9	11.47	62	25.63	8.742	0.80	-10.32	6.201	0.89
451	39	536.3	12.55	569.6	11.69	591.5	11.4	30	33.31	8.142	0.82	21.86	6.572	0.89
453	40	577.8	12.17	649.2	10.84	668.7	10.33	49	71.33	8.778	0.80	19.55	6.879	0.87
454	41	503.8	13.1	556.2	10.95	596.3	10.3	31	52.48	9.52	0.75	40.05	6.974	0.87
455	42	540.6	10.52	604.6	10.38	586.9	9.669	28	64.08	7.408	0.85	-17.69	6.506	0.89
457	43	378.8	16.37	441.1	13.56	491	12.93	8	62.28	12	0.62	49.89	8.904	0.79
458	44	760.1	11.26	799.4	10.03	813.5	9.294	67	39.32	7.059	0.87	14.07	5.841	0.91
459	45	437.2	14.99	470.8	12.86	512.1	12.08	12	33.57	12.34	0.60	41.31	9.729	0.75
460	46	409.3	8.27	450.9	8.337	557.9	8.395	21	41.66	7.255	0.86	107	6.768	0.89
461	47	491.4	11.61	562.9	10.71	604.9	9.515	34	71.49	8.294	0.82	41.94	6.526	0.88
466	48	672.1	18.88	709	16.52	705.6	16.15	53	36.91	10.9	0.68	-3.415	8.137	0.82
611	49	373.6	6.69	414.8	6.985	469	7.203	3	41.19	4.702	0.94	54.24	4.207	0.95
612	50	424	6.083	470.5	6.182	487.7	5.977	7	46.42	3.808	0.96	17.21	3.446	0.97
613	51	695.1	7.477	721.7	7.222	735.1	6.916	58	26.57	3.251	0.97	13.42	3.004	0.98
614	52	433.8	7.358	468.4	7.367	498	7.368	10	34.57	5.283	0.92	29.6	4.254	0.95
615	53	687.7	6.885	729.1	6.56	720.6	6.402	55	41.4	3.708	0.96	-8.494	3.874	0.96
616	54	401.1	6.213	442.4	6.496	458	6.423	1	41.35	3.887	0.96	15.61	3.333	0.97
617	55	524.3	7.778	552.4	7.547	579.8	7.262	26	28.09	3.694	0.96	27.44	3.481	0.97
618	56	690.8	9.9	720.9	9.158	739.5	8.903	59	30.18	4.235	0.95	18.57	3.745	0.96
619	57	679.2	8.125	705.2	7.885	729.7	7.44	57	25.91	3.391	0.97	24.51	3.025	0.97
620	58	487.7	8.934	537.6	8.836	535	8.937	16	49.91	5.026	0.93	-2.681	4.432	0.95
622	59	372.8	6.974	425.5	7.294	479.9	7.434	5	52.77	5.001	0.93	54.39	4.247	0.95
624	60	481.2	7.456	508.8	7.209	526.4	7.3	14	27.56	3.767	0.96	17.58	3.6	0.96
625	61	558.4	8.684	593.6	8.309	611.8	8.196	37	35.18	4.013	0.96	18.17	3.562	0.96
626	62	490.7	8.212	514.1	8.31	518.1	8.127	13	23.41	4.204	0.95	3.993	3.578	0.97
627	63	440.9	8.493	485.1	8.432	493.1	8.448	9	44.28	5.029	0.93	7.959	4.274	0.95
628	64	766.3	8.114	800.8	7.185	794.9	7.081	65	34.43	4.818	0.94	-5.9	3.77	0.96
629	65	709.1	8.579	728.5	7.954	753.3	7.537	61	19.4	4.837	0.94	24.79	4.28	0.95
630	66	547.5	11.53	598.7	10.82	604.1	10.53	34	51.25	6.033	0.90	5.306	5.825	0.91
631	67	510.5	10.46	573.9	10.05	625.3	9.509	41	63.4	6.833	0.87	51.32	5.82	0.91
642	68	407.1	17.05	455.6	17.19	468.3	16.63	3	48.44	8.885	0.79	12.7	8.482	0.81
671	69	605.5	13.97	602.9	13.35	641.4	12.18	45	-2.617	10.39	0.72	38.46	8.28	0.82

Table 6. Marginal Posterior Means and Standard Deviations of School PSAA Ratios ($\hat{\lambda}_j^t$) followed by the Marginal Posterior Means of the Probability of a School Achieving Set Proportions (α_ℓ) towards the PSAA Target of 800 on the API for 1999 and 2000. Last row gives the counts of schools gaining at least α_ℓ at 90% confidence.

	1999 Gains									2000 Gains								
	$\hat{\lambda}_j^2$		Probability of Achieving α_ℓ							$\hat{\lambda}_j^3$		Probability of Achieving α_ℓ						
School	Mean	S.D.	0	3	5	10	15	20	25	Mean	S.D.	0	3	5	10	15	20	25
404	0.289	0.058	1.00	1.00	1.00	1.00	0.99	0.94	0.76	0.118	0.084	0.91	0.85	0.80	0.61	0.37	0.16	0.04
407	0.156	0.029	1.00	1.00	1.00	0.97	0.58	0.06	0.00	0.127	0.023	1.00	1.00	1.00	0.88	0.16	0.00	0.00
410	0.160	0.033	1.00	1.00	1.00	0.96	0.63	0.11	0.00	0.279	0.027	1.00	1.00	1.00	1.00	1.00	1.00	0.86
412	0.232	0.026	1.00	1.00	1.00	1.00	1.00	0.90	0.25	0.000	0.028	0.49	0.13	0.03	0.00	0.00	0.00	0.00
413	0.293	0.044	1.00	1.00	1.00	1.00	1.00	0.98	0.84	0.201	0.048	1.00	1.00	1.00	0.98	0.86	0.51	0.15
414	0.112	0.043	1.00	1.00	0.92	0.62	0.19	0.02	0.00	0.130	0.039	1.00	0.99	0.98	0.78	0.30	0.03	0.00
415	0.189	0.057	1.00	1.00	0.99	0.94	0.76	0.43	0.14	0.078	0.066	0.88	0.78	0.68	0.38	0.13	0.03	0.00
416	0.245	0.052	1.00	1.00	1.00	1.00	0.97	0.81	0.46	0.153	0.054	1.00	0.99	0.97	0.84	0.53	0.19	0.03
417	0.176	0.029	1.00	1.00	1.00	0.99	0.82	0.21	0.00	0.137	0.025	1.00	1.00	1.00	0.93	0.30	0.00	0.00
418	0.055	0.041	1.00	1.00	0.57	0.13	0.01	0.00	0.00	0.105	0.031	1.00	0.99	0.96	0.57	0.08	0.00	0.00
419	0.065	0.029	1.00	1.00	0.70	0.11	0.00	0.00	0.00	0.202	0.018	1.00	1.00	1.00	1.00	1.00	0.55	0.00
420	0.272	0.069	1.00	1.00	1.00	0.99	0.96	0.85	0.64	0.036	0.099	0.66	0.55	0.46	0.26	0.12	0.04	0.01
421	0.244	0.042	1.00	1.00	1.00	1.00	0.99	0.85	0.45	-0.070	0.049	0.08	0.02	0.00	0.00	0.00	0.00	0.00
422	0.103	0.048	1.00	1.00	0.87	0.53	0.17	0.02	0.00	-0.032	0.049	0.27	0.11	0.04	0.00	0.00	0.00	0.00
425	0.128	0.032	1.00	1.00	0.99	0.81	0.25	0.01	0.00	0.191	0.030	1.00	1.00	1.00	1.00	0.92	0.38	0.02
427	1.386	35.22	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.932	67.25	0.08	0.08	0.09	0.09	0.09	0.10	0.11
428	-4.134	838.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-0.017	0.181	0.51	0.59	0.64	0.75	0.84	0.91	0.95
429	0.068	0.021	1.00	1.00	0.81	0.06	0.00	0.00	0.00	0.129	0.018	1.00	1.00	1.00	0.94	0.13	0.00	0.00
430	0.162	0.026	1.00	1.00	1.00	0.99	0.68	0.07	0.00	0.197	0.025	1.00	1.00	1.00	1.00	0.97	0.45	0.02
431	0.132	0.019	1.00	1.00	1.00	0.95	0.17	0.00	0.00	0.153	0.018	1.00	1.00	1.00	1.00	0.57	0.00	0.00
432	0.271	0.022	1.00	1.00	1.00	1.00	1.00	1.00	0.83	0.105	0.024	1.00	1.00	0.99	0.59	0.03	0.00	0.00
433	0.254	0.029	1.00	1.00	1.00	1.00	1.00	0.97	0.56	0.133	0.031	1.00	1.00	0.99	0.86	0.29	0.01	0.00
434	0.240	0.041	1.00	1.00	1.00	1.00	0.98	0.84	0.41	0.030	0.044	0.76	0.51	0.32	0.05	0.00	0.00	0.00
435	0.076	0.024	1.00	1.00	0.86	0.15	0.00	0.00	0.00	0.009	0.021	0.67	0.16	0.02	0.00	0.00	0.00	0.00
436	0.182	0.025	1.00	1.00	1.00	1.00	0.89	0.24	0.00	0.018	0.029	0.73	0.34	0.13	0.00	0.00	0.00	0.00
437	0.286	0.025	1.00	1.00	1.00	1.00	1.00	1.00	0.93	0.145	0.029	1.00	1.00	1.00	0.94	0.44	0.03	0.00
439	0.210	0.020	1.00	1.00	1.00	1.00	1.00	0.69	0.02	0.041	0.026	0.94	0.68	0.37	0.01	0.00	0.00	0.00
440	0.677	1.355	1.00	1.00	0.99	0.99	0.98	0.97	0.96	-26.71	1728.0	0.02	0.02	0.02	0.01	0.01	0.01	0.00
441	0.482	0.066	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.280	0.108	0.99	0.99	0.98	0.96	0.89	0.78	0.62
442	3.538	159.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-0.058	30.730	0.80	0.84	0.86	0.90	0.94	0.96	0.97
443	0.117	0.023	1.00	1.00	1.00	0.77	0.08	0.00	0.00	0.138	0.025	1.00	1.00	1.00	0.93	0.31	0.01	0.00
444	0.256	0.098	1.00	1.00	0.98	0.95	0.87	0.73	0.53	-0.124	0.154	0.20	0.14	0.10	0.04	0.01	0.00	0.00
445	-0.055	0.084	1.00	1.00	0.09	0.02	0.00	0.00	0.00	0.000	0.051	0.51	0.28	0.16	0.02	0.00	0.00	0.00
446	0.162	0.019	1.00	1.00	1.00	1.00	0.74	0.02	0.00	0.037	0.021	0.96	0.64	0.27	0.00	0.00	0.00	0.00
447	0.122	0.027	1.00	1.00	0.99	0.80	0.14	0.00	0.00	0.113	0.030	1.00	1.00	0.98	0.67	0.10	0.00	0.00
448	0.146	0.023	1.00	1.00	1.00	0.97	0.44	0.01	0.00	0.081	0.026	1.00	0.97	0.88	0.24	0.00	0.00	0.00
449	0.697	0.158	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.805	13.73	0.98	0.97	0.97	0.95	0.93	0.89	0.84
450	0.553	0.279	1.00	1.00	1.00	0.99	0.99	0.98	0.96	-0.660	35.66	0.05	0.04	0.03	0.02	0.01	0.01	0.00
451	0.126	0.029	1.00	1.00	0.99	0.82	0.20	0.00	0.00	0.095	0.027	1.00	0.99	0.94	0.43	0.02	0.00	0.00
453	0.321	0.034	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.129	0.043	1.00	0.99	0.96	0.75	0.32	0.04	0.00
454	0.177	0.029	1.00	1.00	1.00	0.99	0.83	0.21	0.00	0.164	0.027	1.00	1.00	1.00	0.99	0.71	0.09	0.00
455	0.247	0.027	1.00	1.00	1.00	1.00	1.00	0.96	0.45	-0.092	0.036	0.00	0.00	0.00	0.00	0.00	0.00	0.00
457	0.147	0.026	1.00	1.00	1.00	0.97	0.46	0.02	0.00	0.139	0.023	1.00	1.00	1.00	0.95	0.32	0.00	0.00
458	1.086	2.422	1.00	1.00	1.00	1.00	1.00	1.00	1.00	3.555	595.0	0.99	0.99	0.99	1.00	1.00	1.00	1.00
459	0.092	0.032	1.00	1.00	0.90	0.41	0.03	0.00	0.00	0.125	0.028	1.00	1.00	1.00	0.82	0.18	0.00	0.00
460	0.107	0.018	1.00	1.00	1.00	0.65	0.01	0.00	0.00	0.306	0.018	1.00	1.00	1.00	1.00	1.00	1.00	1.00
461	0.232	0.024	1.00	1.00	1.00	1.00	1.00	0.91	0.22	0.177	0.025	1.00	1.00	1.00	1.00	0.86	0.17	0.00
466	0.289	0.077	1.00	1.00	1.00	0.99	0.96	0.88	0.70	-0.044	0.098	0.34	0.22	0.16	0.05	0.02	0.00	0.00
611	0.097	0.011	1.00	1.00	1.00	0.38	0.00	0.00	0.00	0.141	0.011	1.00	1.00	1.00	1.00	0.19	0.00	0.00
612	0.123	0.010	1.00	1.00	1.00	0.99	0.00	0.00	0.00	0.052	0.010	1.00	0.98	0.59	0.00	0.00	0.00	0.00
613	0.254	0.031	1.00	1.00	1.00	1.00	1.00	0.96	0.54	0.172	0.037	1.00	1.00	1.00	0.97	0.72	0.22	0.02
614	0.094	0.014	1.00	1.00	1.00	0.34	0.00	0.00	0.00	0.089	0.012	1.00	1.00	1.00	0.19	0.00	0.00	0.00
615	0.369	0.033	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-0.123	0.060	0.01	0.00	0.00	0.00	0.00	0.00	0.00
616	0.104	0.009	1.00	1.00	1.00	0.65	0.00	0.00	0.00	0.044	0.009	1.00	0.93	0.25	0.00	0.00	0.00	0.00
617	0.102	0.013	1.00	1.00	1.00	0.56	0.00	0.00	0.00	0.111	0.013	1.00	1.00	1.00	0.79	0.00	0.00	0.00
618	0.277	0.038	1.00	1.00	1.00	1.00	1.00	0.98	0.76	0.237	0.049	1.00	1.00	1.00	1.00	0.97	0.77	0.38
619	0.215	0.028	1.00	1.00	1.00	1.00	0.99	0.70	0.11	0.259	0.032	1.00	1.00	1.00	1.00	1.00	0.97	0.60
620	0.160	0.015	1.00	1.00	1.00	1.00	0.74	0.00	0.00	-0.010	0.017	0.27	0.01	0.00	0.00	0.00	0.00	0.00
622	0.124	0.011	1.00	1.00	1.00	0.98	0.01	0.00	0.00	0.145	0.011	1.00	1.00	1.00	1.00	0.33	0.00	0.00
624	0.086	0.011	1.00	1.00	1.00	0.11	0.00	0.00	0.00	0.060	0.012	1.00	0.99	0.80	0.00	0.00	0.00	0.00
625	0.146	0.016	1.00	1.00	1.00	1.00	0.39	0.00	0.00	0.088	0.017	1.00	1.00	0.99	0.24	0.00	0.00	0.00
626	0.076	0.013	1.00	1.00	0.97	0.03	0.00	0.00	0.00	0.014	0.012	0.87	0.10	0.00	0.00	0.00	0.00	0.00
627	0.123	0.013	1.00	1.00	1.00	0.96	0.02	0.00	0.00	0.025	0.013	0.97	0.36	0.03	0.00	0.00	0.00	0.00
628	1.080	0.506	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.687	74.97	0.06	0.06	0.06	0.06	0.06	0.06	0.06
629	0.213	0.049	1.00	1.00	1.00	0.99	0.90	0.61	0.23	0.349	0.060	1.00	1.00	1.00	1.00	1.00	0.99	0.96
630	0.203	0.022	1.00	1.00	1.00	1.00	0.99	0.55	0.02	0.026	0.029	0.82	0.45	0.20	0.00	0.00	0.00	0.00
631	0.219	0.022	1.00	1.00	1.00	1.00	1.00	0.81	0.07	0.227	0.024	1.00	1.00	1.00	1.00	1.00	0.87	0.17
642	0.123	0.022	1.00	1.00	1.00	0.86	0.11	0.00	0.00	0.037	0.024	0.94	0.61	0.29	0.00	0.00	0.00	0.00
671	-0.015	0.054	1.00	1.00	0.11	0.01	0.00	0.00	0.00	0.195	0.038	1.00	1.00	1.00	0.99	0.88	0.45	0.07
			63	63	55	42	24	14	6			45	40	36	22	9	4	2

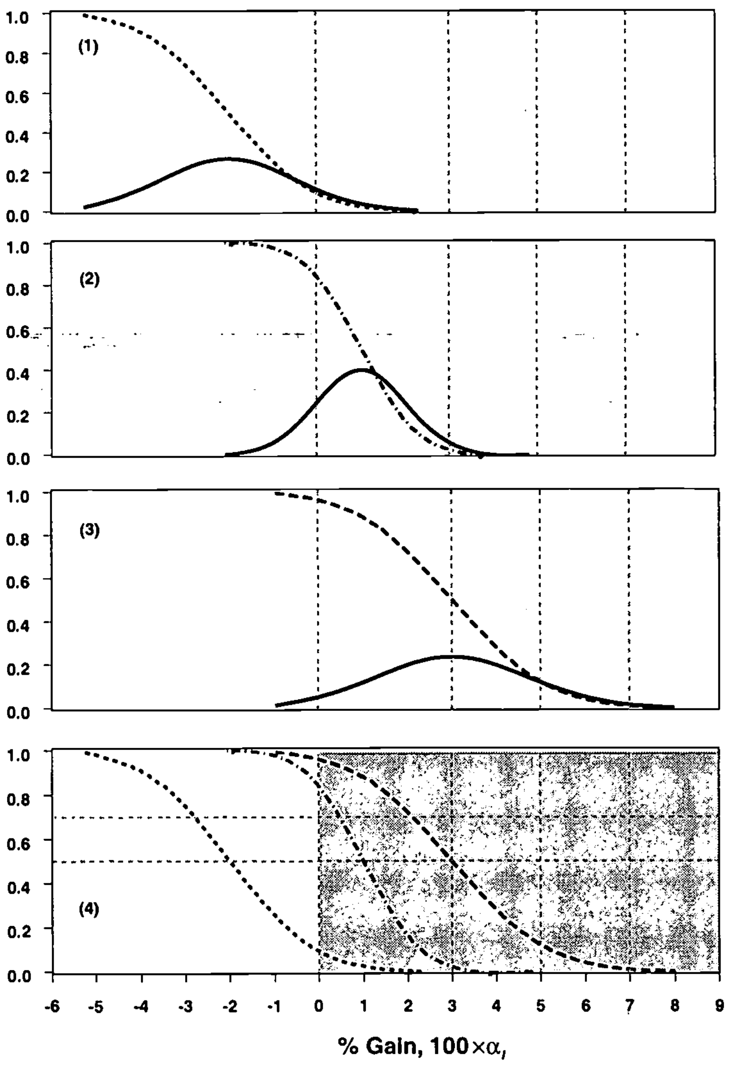

Fig. 5. Constructing school API productivity profiles: For set proportions of API gains, α_ℓ, simply replace the sampling distribution (solid line) of a gain estimate with its corresponding cumulative probability (dashed line) given by Equation (35). The height of the curve corresponding to an API gain of magnitude α_ℓ is the probability of observing an API gain of at least as large as α_ℓ. Panels (1), (2), and (3) give examples of three estimates with different location and spread, and Panel (4) overlays the results. We may easily determine the relative productivity gain for each school for fixed confidence levels (horizontal reference lines). *Reproduced from Thum (2002), and included for completeness.*

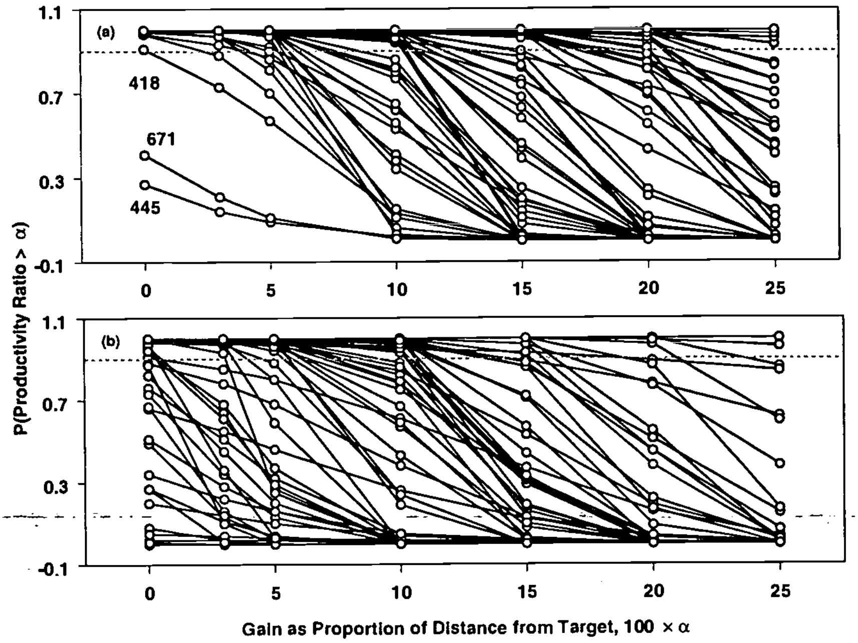

Fig. 6. (a): Productivity profiles for API gains schools made in 1999 against how far in 1998 each school was from the statewide API target of 800. **(b)**: Productivity profiles for API gains schools made in 2000 against how far in 1999 each school was from the statewide API target of 800. Schools 427, 428, 442, 458, and 628 are excluded because in 1999 they are either close to or have exceeded the 800 mark and therefore do not have meaningful productivity profiles.

Figure 5, taken from Thum (2002), shows how to represent a school's gain with its productivity profile. For our LBUSD schools, their productivity profiles for 1999 gain are given in columns 4–10 of Table 6 while columns 13-19 contain the estimates for 2000. Profiles for schools with reasonable estimates for 1999 and 2000 are plotted in Panel (a) and Panel (b) of Figure 6, respectively. In 1999, only School 418 and School 671 failed to show any growth at the 90% confidence level while four schools did not register productivity improvement of at least 3% at the 90% confidence level. Six of the 69 schools failed to achieve the 5% goal at the 90% confidence level. Productivity in 2000 is however quite mixed, with only 31 schools attaining the 5% relative improvement target with 90% confidence.[19] The display allows the user complete flexibility to set both the improvement target as well as the level of confidence of any judgement. In contrast, the current PSAA ratio has none of these essential properties to make it a useful starting point for commenting on school improvement.

8 Summary and Discussion

In this paper, we set out to provide a coherent statistical framework for thinking about educational performance indicators such as the API, and for its improvement measure in the PSAA productivity ratio. With the help of several computational examples and sample code segments, we reason aloud the steps needed for an initial analysis. We consider this foundation a useful first stride towards

[19] The last row in Table 6 counts the number of schools gaining at least the corresponding α_ℓ at the 90% confidence level.

a deeper appreciation for the conceptual underpinnings of a class of growth indicators. We hope that, in turn, this would foster their proper use.

We begin by pointing to the lack of a realistic assessment of the precision of published results without which we have scant statistical basis for detecting school progress over time. We show how conventional computations of the school API and the productivity ratio may be easily represented in the form of the standard weighted linear regression model. We note in particular that our formulation provides standard inferences for year-to-year gains and, by employing the $\delta-$method, deduces approximate standard errors for the PSAA productivity ratio. The immediate result of this preliminary step is that we may now begin to "model" rather than "compute" these indicators at the school-level.

If the productivity result for a school is to represent how its students perform, we find that the current school API definition is silent about student progress. We see a clear indication of a conceptual incongruity between school and student performance when we try to accommodate students with missing scores on one or several of the tests by re-distributing the student weights as explained in Section 3.4. This is a mistaken effort in our judgement, although we do not suggest a remedy here. Until some resolution is found, we are content to show that for students without any missing test scores, the school API can reflect aggregate student performance when we employ a multivariate-multilevel model (tests over time within student, students within school). In this approach, the school API estimates are the average of the API estimates for each student.

Even when we employ students with a full set of scores, we have not skirted the issue. First, we alluded to the cross-sectional nature of its design. Second, there is no assessment of student variability in the school API. If student variability is taken to be null, this amounts to the untenable presumption that students perform comparably when in reality assessment results vary considerably among students in a school. The problem is clearly demonstrated, with a numerical example in Section 4, when we showed how the change in the school API from one year to the next may sharply contradict the pattern of change as conveyed in an average of student growth (also see Meyer, 1996).

With the individual school results as input, we next introduce a meta-analysis model to aggregate over the results from our set of schools. The primary reason for preferring a meta-analysis over a larger multilevel model (*e.g.*, tests over time within student, students within school, schools within school system) is that sample sizes for schools are typically large and their estimates are quite likely to be well estimated. We do not expect large differences between the separate school estimates and those obtained from a larger model for the system. Secondly, with good school estimates as input, a meta-analysis is far less demanding computationally. Third, the model also provides better estimates of the true performance of individual schools. We expect that this combination of school-specific multivariate multilevel modelling and a meta-analysis for the collection of schools can easily accommodate the nearly 5000 elementary schools in California utilizing only modest computing resources.

We have further illustrated several significant advantages with a Bayesian formulation of our meta-analysis model as implemented in WINBUGS. We obtain the marginal posterior distributions of all unknown parameters, in our case, school API status and gain estimates as well as the between school components of variance. In addition, it is a straightforward matter to simulate the posterior marginal distribution of any function (linear or otherwise) of unknown parameters, such as the PSAA school productivity ratio, reliability estimates for gains, and contrasts between schools on any of these indicators, as well as the rank of a school. When we predicted school gains from prior school initial status, we have also illustrated a simple application of latent variable regression within the multilevel framework. For reasons of scope and space, we have not employed any covariates in our discussion although the reader should be well aware that they are critical to a fuller analysis and that covariates are also easily accommodated within our modelling framework.

Contrary to present practice, we argue that any judgement about how much a school has improved in terms of an estimate should also take into account the uncertainty of that estimate. We provided an adequate strategy in the school productivity profile. In a display which juxtaposes the amount of gain with its corresponding confidence level, the user is given all the information for making a proper statistical determination. Not illustrated here are extensions of this procedure to describe the productivity of different grade levels, significant subgroups, school clusters, and districts, and to include adjustments on appropriate covariates (Thum, 2002).

Like many researchers (*e.g.*, Goldstein & Spiegelhalter, 1996; Thum & Bryk, 1997; Linn, 2000), we recognize that sound accountability decisions require more than a set of modelling procedures and guidelines on its use, particularly in the present climate of school reform. Both the purveyors of school indicators and their consumers alike understand that the validity of our results is contingent upon careful analysis, identifying the appropriate contextual factors and a good theory about their roles in impacting how students, parents, teachers, and schools perform. Beyond the need to contextualize comparisons, the absence of precision estimates is at the core of many of the criticisms aimed at the so-called "league tables" and the now numerous public postings of school report cards. In this study, we have detailed the beginnings of a workable modelling approach for the API that, by making explicit the inherent uncertainties in our estimates, would greatly enhance the credibility of the evaluations we make using student assessment data. When implemented within a responsible program that includes technical audits and open reviews (Thum, 1999), the core tools we have introduced above should help to produce diagnostics more trustworthy than most now in circulation.

For more information please contact

Yeow Meng Thum
Department of Education
Graduate School of Education and Information Studies
University of California, Los Angeles
2019A Moore Hall, Box 951521
Los Angeles, CA 90095-1521
thum@ucla.edu

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (1999). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Bishop, Y. M. M., Fienberg, S. and Holland, P. (1975). *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, MA: MIT Press.

Bock, R. D. (1975). Multivariate Statistical Methods in Behavioral Research. McGraw-Hill: New York, NY.

Brooks, S. P. and Roberts, G. O. (1998). Assessing convergence of Markov Chain Monte Carlo algorithms. *Statistics and Computing*, 8, 319–335.

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

Bryk, A. S., Thum, Y. M., Easton, J. Q., and Luppescu, S. (1998). Assessing School Academic Productivity: The Case of Chicago School Reform. *Social Psychology of Education*, 2,103–142.

Burstein, L. (1980a). The analysis of multilevel data in educational research and evaluation. *Review of Research in Education*, 8, 158–233.

Burstein, L. (1980b). The role of levels of analysis in the specification of educational effects. In R. Dreeben and J. A. Thomas (Eds.) *The analysis of educational productivity, Vol. 1: Issues in microanalysis.* Cambridge, MA: Ballinger.

Cochran, W.J. (1977). Sampling Techniques, Second Edition, New York: John Wiley & Sons, Inc.

Collins, L. M. (1996). Is reliability obselete? A commentary on "Are simple gain scores Obselete?" *Applied Psychological Measurement*, 20, 289–292.

Congdon, P. (2001). *Bayesian Statistical Modelling.* London, UK: John Wiley & Sons.

Cowles, M. K. and Carlin, B. P. (1996). Markov Chain Monte Carlo diagnostics: A comparative review. *Journal of the American Statistical Association*, 91, 883–904.

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulating questions, design, and analysis (Occasional Paper).* Stanford, CA: Stanford University Consortium.

Cronbach, L. J. and Furby, L. (1970). How should we measure "change" – or should we? *Psychological Bulletin*, 74, 68–80.

deGroot, M. H. (1975) *Probability and Statistics.* Reading, MA: Addison-Wesley.

Dyer, H. S., Linn, R. L., and Patton, M. J. (1969). A comparison of four methods of obtaining discrepancy measures based on observed and predicted school system means on achievement tests. *American Educational Research Journal*, 6, 591–605.

Goldstein, H. and Healy, M. J. R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society, A*, 158, 175–177.

Goldstein, H. and Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society, Series A*, 159, 384–443.

Gray, J., Jesson, D., Goldstein, H., Hedger, K., and Rabash, J. (1995). A multi-level analysis of school improvement: Changes in school's performance over time. *School Effectiveness and School Improvement*, 6, 97–114.

Harker, R. and Nash, R. (1996). Acedemic outcomes and school effectiveness: Type "A" and Type "B" effects. *New Zealand Journal of Educational Studies*, 32, 143–170.

Indurkhya, A., Gardiner, J. C., and Luo, Z. H. (2001). The effect of outliers on confidence interval procedures for cost-effectiveness ratios. *Statistics in Medicine*, 20, 1469–1477.

Laird, N. M. and Louis, T. A. (1989). Empirical Bayes ranking methods. *Journal of Educational Statistics*, 14, 29–46.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, 29, 4–16.

Linn, R. L. and Slinde, J. A. (1977). The determination of the significance of change betweenpre- and postting periods. *Review of Educational Research*, 47, 121–150.

Maris, E. (1998). Covariance adjustment versus gain score – revisited. *Psychological Methods*, 3, 309–327.

Mellenbergh, G. J. and van de Brink, W. P. (1998). The measurement of individual change. *Psychological Methods*, 3, 470–485.

Mellenbergh, G. J. (1999). A note on simple gain score precision. *Applied Psychological Measurement*, 23, 87–89.

Meyer, R. H. (1996). Value-added indicators of school performance. In E. A. Hanushek & D. W. Jorgenson (Eds.), *Improving America's Schools: The role of incentives (pp. 197–223).* Washington, DC: National Academic Press.

Normand, S. L. T. (1999). Meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in Medicine*, 18, 321–359.

O'Hagan, A., Stevens, J. W., and Montmartin, J. (2000). Inference for the cost-effectiveness acceptability curve and the cost-effectiveness ratio. *Pharmocoeconomics*,17, 339–349.

O'Hagan, A., Stevens, J. W., and Montmartin, J. (2001). Bayesian cost-effectiveness analysis from clinical trial data. *Statistics in Medicine*, 20, 733–753.

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

Raudenbush, S. W. and Bryk, A. S. (1986). A hierarchical linear model for school effects. *Sociology of Education*, 59, 1–17.

Raudenbush, S. W. and Bryk, A. S. (2001). *Hierarchical linear models: applications and data analysis methods, 2^{nd} Edition.* Newbury Park, CA: Sage.

Raudenbush, S. W., Fotiu, R. P. and Cheong, Y. F. (1999). Synthesizing results from the trial state assessment. *Journal of Educational and Behavioral Statistics*, 24, 413–438.

Raykov, T. (1993). A Structural Equation Model for Measuring Residualized Change and Discerning Patterns of Growth or Decline. *Applied Psychological Measurement*, 17, 53-71.

Raykov, T. (1999). Are Simple Gain Scores Obsolete? On an Approach to the Study of Correlates and Predictors of Ability Growth. *Applied Psychological Measurement*, 23, 120-126.

Rothstein, R. (2000). Toward a composite index of school performance. *The Elementary School Journal*, 100, 409–441.

Rogosa, D. R., Brand, D. and Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, 90, 726–748.

Rogosa, D. R. and Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika*, 50, 203–228.

Rogosa, D. R. (1995). Myths and methods: "Myths about longitudinal research" plus supplemental questions. In J. M. Gottman (ed.), *The Analysis of Change.* Mahwah, NJ: LEA.

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

Seltzer, M., Choi, K. and Thum, Y. (2001). *Examining relationships between where students start and how rapidly they progress: Implications for conducting analyses that help illuminate the distribution of achievement within schools..* Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing, UCLA. (Submitted.)

Singer, J. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, 24, 323–355.

Smith, T. C., Spiegelhalter, D. J. and Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in Medicine*, 14, 2685–2699.

Spiegelhalter, D. J. and Marshall, E. C. (1998). Comparing institutional performance using Markov chain Monte Carlo methods. In B. Everitt and G. Dunn (Eds.) *Recent Advances in the Statistical Analysis of Medical Data, pp. 229–250.* London, UK: Edward Arnold.

Spiegelhalter, D. J., Thomas, A., and Best, N. G.(1999). *WINBUGS: Bayesian inference Using Gibbs Sampling*, Version 1.3, Cambridge, U. K.: MRC Biostatistics Unit.

Thum, Y. M. (1997). Hierarchical linear models for multivariate outcomes. *Journal of Educational and Behavioral Statistics*, 22, 77–108.

Thum, Y. M. (1999). Assessing Academic Performance: Some Notes on New Directions. *Presented at the CRESST Annual Conference, September, 1999, at the Univerity of California, Los Angeles.*

Thum, Y. M. (2002). Measuring Progress towards a Goal: Estimating Teacher Productivity using a Multivariate Multilevel Model for Value-Added Analysis. (*Submitted.*)

Thum, Y. M. and Bryk, A. S. (1997). Value-added Productivity Indicators: The Dallas System. In J. Millman (Ed.), *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure? (pp. 100–109)*, Thousand Oaks, CA: Corwin.

Wasserman, L. A (2000). Asymptotic inference for mixture models using data dependent priors. *Journal of the Royal Statistical Society, Ser. B*, 62, 159–180.

Webster, W. J. and 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.

Willett, J. B. (1988). Questions and answers in the measurement of change. *Review of Research in Education*, 15, 345–422.

Williams, R. H. and Zimmerman, D. W. (1996). Are gain scores obselete? *Applied Psychological Measurement*, 20, 59–69.

Willms, D. (1992). *Monitoring School Performance: A Guide for Educators.* Washington, D. C.: Falmer Press.

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

Yang, M., Goldstein, H., Browne, W. J. and Woodhouse, G. (2001). Multivariate multilevel analysis of examination results. *To appear in Journal of the Royal Statistical Society, Series A.*

Appendix

The $\delta-$Method for Inference on the PSAA Ratio

The $\delta-$method provides an argument for approximating the asymptotic variances (and covariances) of functions of parameters. Here, we will only provide a simple motivation for a more general result.

Suppose we have an estimate, $\hat{\boldsymbol{\theta}}_n$, based on n observations for the population parameters $\boldsymbol{\theta}$. For large n, we may assume that $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$ converges in distribution to $\mathcal{N}\left(\mathbf{0}, \Sigma(\boldsymbol{\theta})\right)$. Thus, $\hat{\boldsymbol{\theta}}_n$ is distributed asymptotic normal with mean vector $\boldsymbol{\theta}$ and asymptotic variance-covariance matrix of $\Sigma(\boldsymbol{\theta})/n$. If $\mathrm{f}(\hat{\boldsymbol{\theta}}_n)$ is a scalar function for which its first derivatives $\partial \mathrm{f}/\partial \boldsymbol{\theta}$ exist, then for large n, an approximation of $\mathrm{f}(\hat{\boldsymbol{\theta}}_n)$ can be obtained by a Taylor-series expansion of $\hat{\boldsymbol{\theta}}_n$ about $\boldsymbol{\theta}$:

$$\mathrm{f}(\hat{\boldsymbol{\theta}}_n) - \mathrm{f}(\boldsymbol{\theta}) \simeq (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \left(\frac{\partial \mathrm{f}'}{\partial \boldsymbol{\theta}}\right) .$$

Theorem 1. *Given the above and, taking expectations, the asymptotic distribution of* $\mathrm{f}(\hat{\boldsymbol{\theta}}_n)$ *is*

$$\mathcal{L}\left[\sqrt{n}\left(\mathrm{f}(\hat{\boldsymbol{\theta}}_n) - \mathrm{f}(\boldsymbol{\theta})\right)\right] \rightarrow \mathcal{N}\left[0, \left(\frac{\partial \mathrm{f}}{\partial \boldsymbol{\theta}}\right) \Sigma(\boldsymbol{\theta}) \left(\frac{\partial \mathrm{f}}{\partial \boldsymbol{\theta}}\right)\right]$$

Bishop, Fienberg, and Holland (1975, pp. 492) also gave a version of Theorem (1) in their Theorem 14.6-2 for a vector-valued function $\mathbf{f}(\hat{\boldsymbol{\theta}}_n)$.

This approach is especially useful in the case of non-linear functions, such as, in our case, a ratio of parameters, as in Equation (21) in Section 3.3,

$$\lambda_j^{(t)} = \frac{\left(\mu_j^{*(t)} - \mu_j^{*(t-1)}\right)}{\left(800 - \mu_j^{*(t-1)}\right)} .$$

The approximation is obtained by deducing the first partial derivatives of $\lambda_j^{(t)}$ with respect to $\mu_j^{*(t-1)}$ and $\mu_j^{*(t)}$. They are

$$\frac{\left(800 - \mu_j^{*(t)}\right)}{\left(800 - \mu_j^{*(t-1)}\right)^2} \quad \text{and} \quad \frac{1}{\left(800 - \mu_j^{*(t-1)}\right)}$$

respectively. Alternative results for functions of other time points include, for example, a comparison of yearly gains relative to performance at the initial time point,

$$\frac{\left(\mu_j^{*(t)} - \mu_j^{*(t-1)}\right)}{\left(800 - \mu_j^{*(1)}\right)} .$$

Given $\Sigma(\boldsymbol{\theta})$, the standard errors for $\mathrm{f}(\hat{\boldsymbol{\theta}}_n)$ are easily obtained. □