› White Paper

Advantages of a Multivariate Longitudinal Approach to Educational Value-Added Assessment Without Imputation

Presented at the 2004 National Evaluation Institute, July 8-10, 2004, Colorado Springs, Colorado.

sas
THE POWER TO KNOW.

Contents

1. Introduction

Although the No Child Left Behind (NCLB) Act of 2001 specifies a particular "model" for using student test scores to evaluate the performance of schools and school districts, there continues to be debate about how best to perform educational evaluation. An attractive alternative to the proficiency level approach of NCLB is a value-added approach; but even among advocates of value-added assessment, there is debate about which statistical approach to value-added is most appropriate. A special issue of the *Journal of Educational and Behavioral Statistics* (Spring 2004) is devoted to this debate. Two papers in particular epitomize the debate between advocates of simple models and complex models. Tekwe, et al. (2004) present an empirical study using a fairly simple data set that suggests that a simple model is adequate. In contrast, McCaffrey, et al. (2004) present a general multivariate, longitudinal model, and they discuss the consequences of making various assumptions to simplify the model.

The purpose of this paper is to compare several different statistical approaches to value-added modeling of educational test scores. Specifically, the objective is to demonstrate the benefits of using a more complex, thus harder to explain, multivariate longitudinal approach rather than a more simplistic, easier to explain and easier to compute approach. The method used to address this question is simulation. That is, rather than analyzing actual test scores, computer generated scores are analyzed and the results from different analysis procedures are compared. One advantage of using simulation is that the true values of the parameters underlying the observed data are known; thus the results from statistical analyses of the data can be compared to The Truth. There are other advantages as well; these will be mentioned at the appropriate time. A disadvantage of using simulation is that one cannot simulate every possible situation that might arise in practice; thus simulation cannot prove that one method is universally superior to another. However, if realistic situations are simulated, it *can* provide credible evidence about the relative merits of various methods.

To help focus attention on the most relevant issues, the simulated data set was kept fairly simple. (This is another advantage of simulation; the data can be made as simple or as complex as desired.) Specifically, four consecutive years of data (test scores) from a single cohort of students were simulated and analyzed. The cohort had either 10, 25, 50, or 100 students. The educational parameter to be estimated was the average amount of growth ("gain") from year three to year four. Various patterns of gain, explained below, were simulated; and missing values were introduced into the data either randomly or non-randomly in such a way that lower performing students were more likely to have missing scores. Each simulation (data generation and data analysis) was repeated 5000 times.

Three methods were used to estimate the average gain from year three to year four: a "simple" method and two more complex methods. The simple method was the average observed gain: each student's year three observed score was subtracted from their year four observed score; the resulting observed gains were then averaged for all students in the cohort. Obviously, if either the year three or the year four score was missing, no observed gain could be calculated. The average uses only the non-missing observed gains. The two complex models were longitudinal models in which the yearly mean scores were estimated using generalized least squares with the covariance matrix among years being estimated by REML (Littell, *et al.*, 1996, especially Appendix 1 and Chapter 3). The estimated gain from year three to year four is simply the year four estimated mean minus the year three estimated mean. In this methodology, all non-missing scores are used; that is, students with incomplete data (fewer than all four years) are included in the analysis, but without "filling in" an imputed value for missing scores. In the first of the complex models, only years three and four were used; in the second complex model, all four years of data were used even though the parameter to be estimated involved only years three and four.

2. Simulation Details

The student test scores were generated using a classical test theory model; that is, each observed score was constructed from a true score plus random "error." The true score represents the student's actual level of attainment in a given year. The "error" is made up of factors which cause a student's score on a particular test administration to deviate from their true score. These factors include measurement error in the test (e.g., choice of which test items to include, error in estimation of item parameters, etc.) as well as student-specific factors (e.g., how much sleep the student got, etc.).

In the simulation, the true scores were generated from a multivariate normal distribution with means for the four years of 400, 500, 600, and 700. Thus the true gain from year three to year four (the value which is to be estimated) is 100 points. The standard deviation each year was 30. The pairwise correlation between any two years was 0.975. Figure 1 shows the scatterplot of year 4 versus year 3 true scores for one cohort of 50 students. The random errors were generated from a normal distribution with all means equal to zero, all standard deviations equal to 20, and all correlations equal to zero. The resulting observed scores were therefore normally distributed with means of 400, 500, 600, 700, standard deviations of 36, and correlations of 0.675. These values for the observed scores are typical of what we encounter in the actual test scores we analyze regularly. Figure 2 shows the scatterplot of the observed scores for year 4 versus year 3 for the same cohort of 50 students shown in Figure 1.

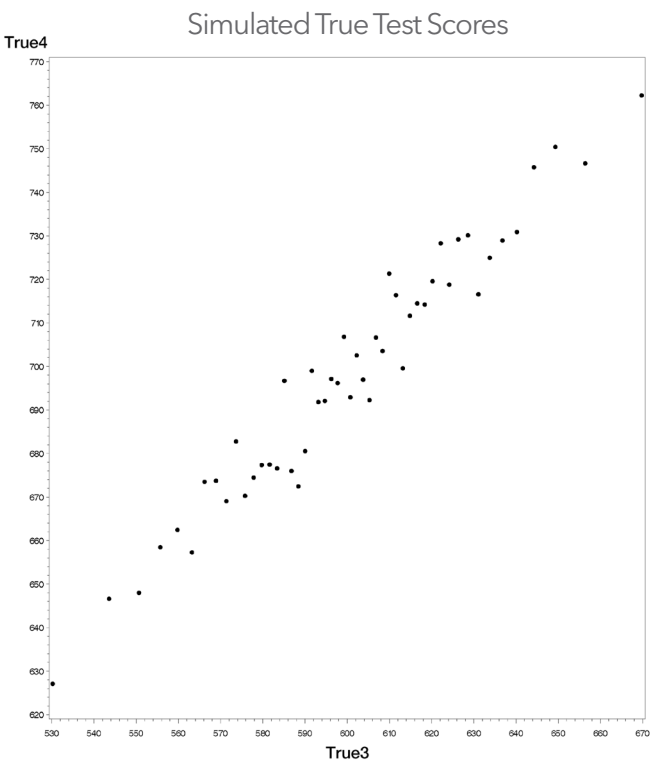

Figure 1. Year 4 versus Year 3 True Scores for 50 Students

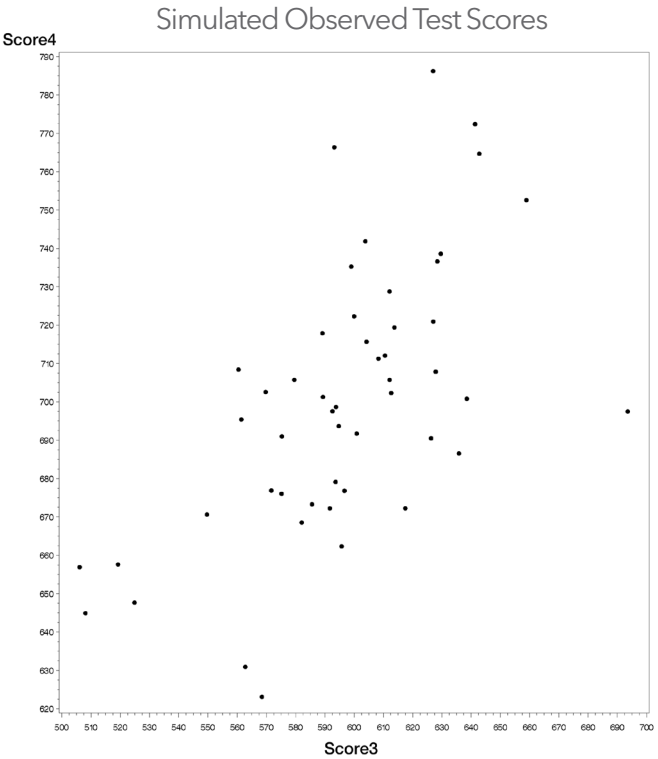

Figure 2. Year 4 versus Year 3 Observed Scores for 50 Students

In the true scores and observed scores just described, students at all levels of attainment make the same gain on average, namely 100 points. This is described as a "flat" gain pattern. In practice (in the actual scores we encounter) this is often not the case. In many cases, students of lower attainment may tend to make more gain, on average, than students of higher attainment; in other cases the reverse may be true, with students of higher attainment making more gain than students of lower attainment, on average. Many other patterns are possible, depending on where teachers choose to focus their instruction, but these two "non-flat" patterns seem to occur most often. These two patterns are denoted the "downward shed" (or "downshed") and the "upward shed" (or "upshed"). In the simulations, these two patterns were created by adjusting the year four true scores (and the corresponding observed scores) to produce a linear relationship between the true gain and the year three true score. The difference in true gain between the high end of the "shed" and the low end was 50 points, which corresponds to one-half year's worth of growth (a year's worth being 100 points). Note that the average true gain for the entire cohort (the value which is to be estimated) is still 100 points.

Figure 3 shows the true gains versus the year 3 true scores for a cohort of 50 students with a flat gain pattern. Figure 4 shows the downward shed pattern. An upward shed pattern (not shown) would be a mirror image of Figure 4, with the higher end of the "shed" to the right rather than to the left.

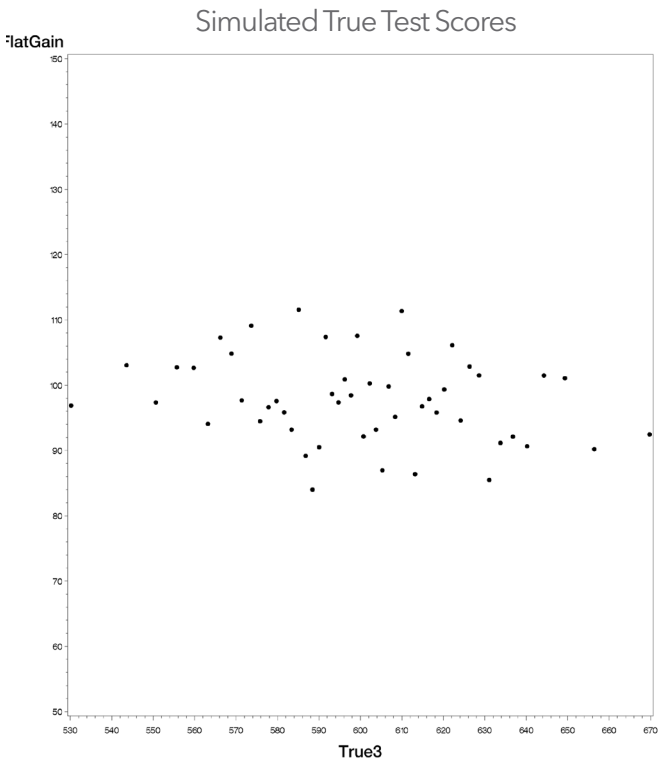

Figure 3. True Gain versus Year 3 True Score for 50 students, flat gain pattern.

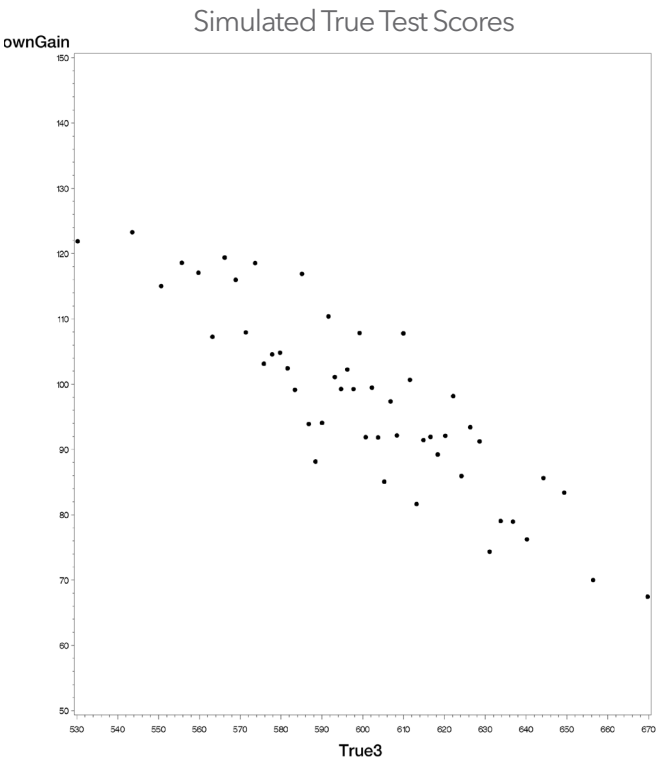

Figure 4. True Gain versus Year 3 True Score for 50 students, downward shed pattern.

Simulations were done both with and without missing values. In those simulations that included missing values, 20% of the values each year were set to missing. The values to be set to missing were chosen either (1) completely at random, (2) partially non-randomly, or (3) mostly non-randomly. Non-random missingness was produced, for each year, as follows. The true scores were ranked from 1 (lowest) to N (highest); a random "rank" from 1 to N was generated; a weighted average of the true rank and the random "rank" was calculated (with the sum of the weights equal to one); the students with the lowest 20% on the weighted average had their scores set to missing. If the weight given to the true rank is zero, random missingness results. A weight of 1.0 (for the true rank) results in completely non-random missingness, with the lowest 20% of true scores being missing. In the simulations, weights (for the true ranks) were 0.35 for "partially non-random" and 0.65 for "mostly non-random."

Figure 5 shows the complete set of true and observed year 4 scores ("True4" and "Score4") for a cohort of 50 students along with three columns having missing values. The rows of the table are in order of increasing true score. The "Miss4_CAR" column has values missing completely at random: the missing values are scattered throughout the column. In the "Miss4_PNR" column, missing values are partly non-random such that the missing values are somewhat more concentrated in the lower true scores. In the "Miss4_MNR" column, missing values are mostly nonrandom, and the missing values are very heavily concentrated in the lower true scores.

Simulated Test Scores

True4	Score4	Miss4_CAR	Miss4_PNR	Miss4_MNR
627	645	.	.	.
647	671	671	671	.
648	657	657	.	.
657	658	658	658	658
658	648	648	648	648
662	631	631	631	.
669	623	.	.	.
670	708	708	708	708
672	662	662	662	.
673	703	.	.	.
674	691	691	691	691
674	669	.	.	.
676	695	695	695	695
677	716	716	.	.
677	676	676	676	676
677	677	677	677	677
681	679	.	.	.
683	706	706	706	706
692	698	698	698	698
692	677	677	677	677
692	699	699	.	699
693	706	706	706	706
696	691	691	691	691
697	672	672	672	672
697	702	702	702	702
697	694	694	694	694
699	673	673	673	673
700	712	712	712	712
703	701	.	701	701
704	672	672	672	672
707	766	766	766	766
707	692	.	.	692
712	722	722	722	722
714	687	.	687	687
714	735	735	735	735
716	719	.	.	719
717	739	739	739	739
719	729	729	729	729
720	701	701	701	701
721	711	711	711	711
725	737	737	737	737
728	721	721	721	721
729	753	753	753	753
729	718	718	718	718
730	765	.	765	765
731	742	742	742	742
746	772	772	772	772
747	697	697	697	697
750	708	708	708	708
762	786	786	786	786

Figure 5. Year four scores without and with missing values. (CAR=completely at random, PNR=partially non–random, MNR=mostly non–random).

3. Simulation Results

In describing how well a method of estimation performs in uncovering The Truth, statisticians commonly use three concepts: bias, variance, and mean square error. These concepts involve the sometimes confusing idea of multiple "realizations" of a set of scores. The idea is that, because of the "random error" component of the observed scores, the actual set of observed scores from a particular test administration is only one of an infinitely large number of sets of scores that might have occurred (if a certain student had eaten breakfast instead of skipping it, if another student had guessed differently on a particular question , etc.). In real life, only one "realization" is available. Another advantage of simulation is that multiple realizations can be generated (in the present case, 5000 realizations were generated) and results can be averaged over multiple realizations to see how well a method of estimation performs "on average" rather than for just one idiosyncratic set of observed scores.

Semi-formal definitions of bias, variance, and mean square error are given below. The "Avg" in these formulas refers to averaging over multiple realizations. Theoretically, the average should be over an infinite number of realizations; in the simulations "infinity" was set to 5000.

Bias = Avg{Estimate – Truth}.

Variance = Avg{[Estimate – Avg(Estimate)] ^{2}}.

Mean square error = Avg{(Estimate – Truth) 2} = **Variance** + **Bias** ^{2}.

In words, **bias** indicates how far, on average, an estimate is from the truth. More colloquially, it measures how far "off target" an estimate tends to be. Understandably, in general an unbiased estimator (one with bias equal to zero) is preferred. **Variance** is the average (squared) distance of the individual estimates from their average value. It measures how tightly clustered the estimates are around their average value, ignoring whether or not the estimates are "off target" (biased). The **standard error**, which is the square root of the variance, may be more familiar; but the variance is more convenient mathematically. The **mean square error** measures the (squared) distance of the estimates from their *true value* (rather than from their average value), combining the information contained in the bias and the variance. Consequently, the mean square error provides an overall indicator of how successful an estimator is at recovering The Truth; the smaller the mean square error, the better. Results from the simulations will be compared primarily based on the mean square error while also noting the extent to which the magnitude of the mean square error results from the bias component or from the variance component.

Results from the simulations are presented as bar charts. The total length of each bar is the mean square error, and each bar is subdivided into a variance component and a bias component. Bars occur in sets of three, corresponding to the three methods used to estimate the gain from year three to year four. The first bar, for the simple average of the non-missing gains, is labeled SMG for simple mean gain. The second bar is labeled LM2; it represents the longitudinal model using two years of data. The third bar, for the longitudinal model with four years of data, is labeled LM4.

Results with No Missing Data. Figure 6 displays the simulation results for cohorts of size 10, 25, 50, and 100 students with no missing data and with a flat gain pattern. This is an ideal situation which rarely occurs in practice. Two conclusions are immediately apparent: (1) all three methods produce exactly the same results, and (2) the estimates are unbiased: the bars are made up entirely of the variance component with no bias component. Also note how dramatically the mean square error decreases as the cohort size increases. The longitudinal models used in the simulations are models that we use in practice for school-level and district-level analyses involving hundreds, or even thousands, of students. With large numbers of students, the estimates will be very close to The Truth. Although the gain pattern in Figure 6 is flat, when there are no missing values, all gain patterns produce the same result.

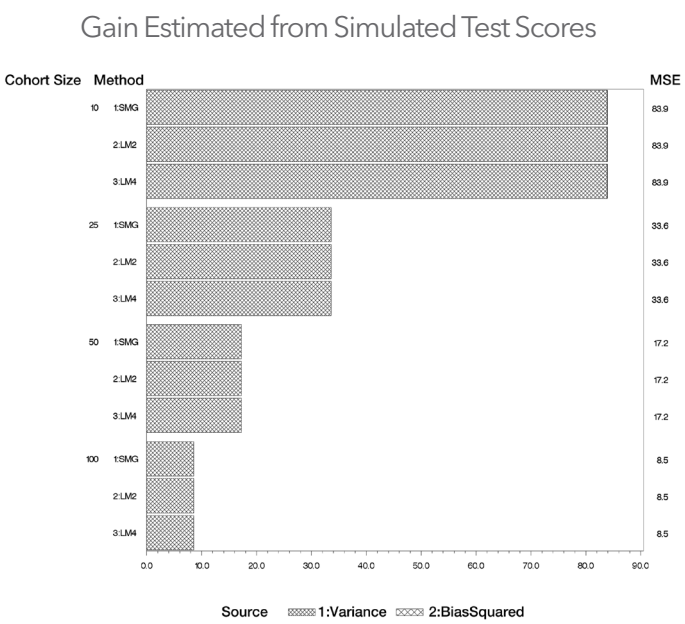

Figure 6. Mean square errors with flat gain pattern and no missing data.

Results with 20% Missing at Random. Figure 7 shows results for cohorts of size 25, 50, and 100 students with a flat gain pattern and 20% of the data missing completely at random. The cohort with 10 students was omitted because, with only 10 students and 20% of the values missing, the 4-year longitudinal model frequently failed to obtain a solution (and the 2-year longitudinal model occasionally failed): there was insufficient data to estimate all the parameters (the longitudinal models must estimate variances and covariances as well as the yearly means). The estimates from all three methods are unbiased, just as they were with no missing data (whenever data are missing at random, the estimates will be unbiased). But in this case, the more complex the model the better the performance (smaller mean square error, i.e., estimates that tend to be closer to The Truth). It is particularly worth noting that the 4-year longitudinal model performs best, even though the parameter of interest (the year 3 to year 4 gain) involves only the last two years. By taking advantage of the correlations among all four years, the 4-year model is able to provide a more precise estimate.

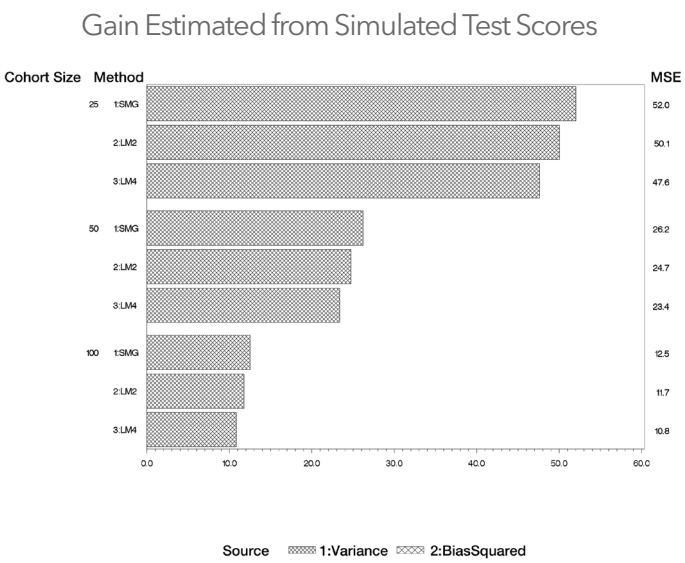

Figure 7. Mean square errors with flat gain pattern and 20% missing at random.

Figures 8 and 9 show comparable results with a downward shed and an upward shed pattern, respectively. The results are essentially identical to Figure 7. When the data are missing at random, the gain pattern has no effect on the performance of the estimates.

Another noteworthy feature in Figure 7, though not immediately obvious from looking at the figure, is that the *relative* advantage of the more complex models over the simple model increases with sample size. For example, in the 25 student cohort, the simple model's mean square error is about 9% larger than that of the 4-year longitudinal model. (Remember that a larger mean square error represents poorer performance.) In the 50 student cohort, the simple model's mean square error is about 12% larger; in the 100 student cohort, it is about 16% larger. A similar pattern occurs in Figure 9; however, the pattern is not evident in Figure 8.

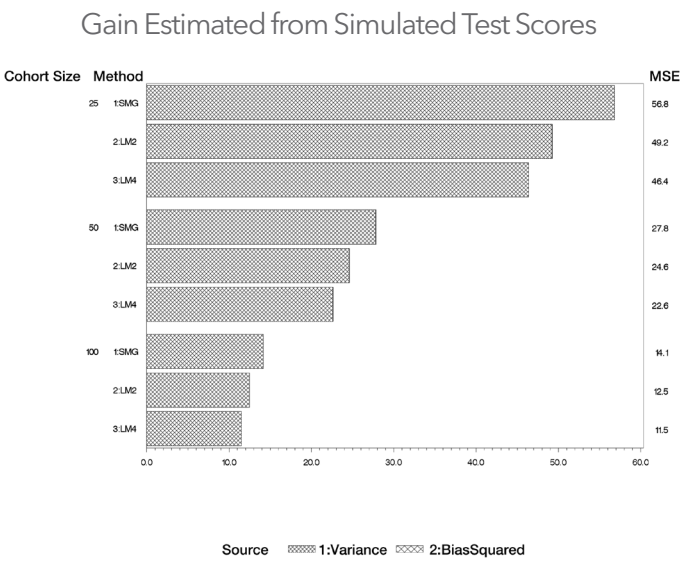

Figure 8. Mean square errors with downward shed pattern and 20% missing at random.

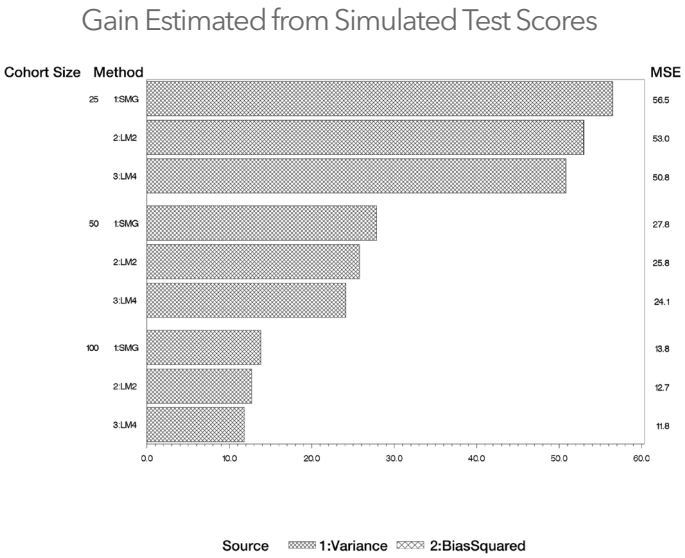

Figure 9. Mean square errors with upward shed pattern and 20% missing at random.

Results with a Flat Pattern and 20% Missing Non-Randomly. Figure10 displays results for cohorts of size 25, 50 and 100 with a flat gain pattern and 20% of the data missing partly non-randomly. The pattern of mean square errors is similar to that in Figures 7, 8 and 9. That is, all estimates are again unbiased (this is always the case with a flat gain pattern), but the more complex models perform better than the simple model. In this case, the advantage of using all four years rather than just two years in the longitudinal model is less clear. The lack of advantage for the 4-year model is more apparent in Figure 11 which shows results with a flat gain pattern 20% missing mostly non-randomly. In this case, the 4-year longitudinal model is comparable in performance to the simple model. The 2-year longitudinal model continues to outperform the simple model, however. A possible explanation for the relatively poorer performance of the 4- year model is that, when nearly all of the lower performing students have missing values, in effect there is a restriction in the range of observed test scores, resulting in correlations among years that are biased toward zero.

The relative advantage of a more complex model compared to the simple model again increases with cohort size, though not as dramatically as in the earlier cases in Figures 7 and 9. In Figures 10 and 11, the simple model's mean square error is about 5% larger than that of the 2-year longitudinal model with 25 students; it is about 7% larger (Figure 10) or 6% larger (Figure 11) when there are 50 or 100 students.

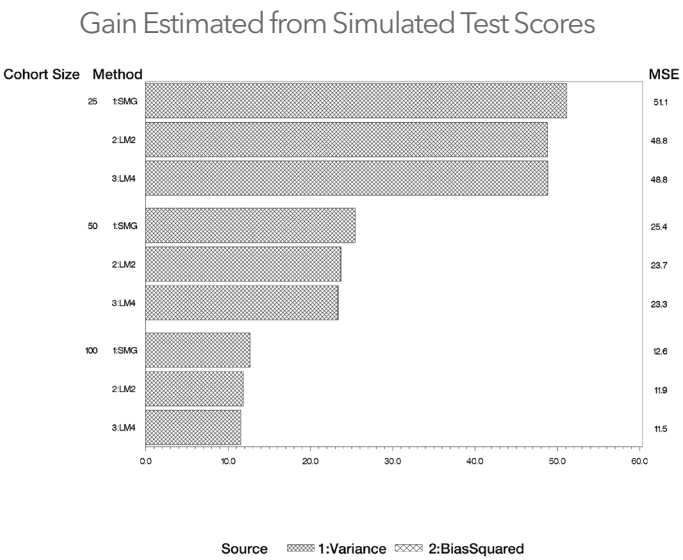

Figure 10. Mean square errors with flat pattern and 20% missing partly non-randomly.

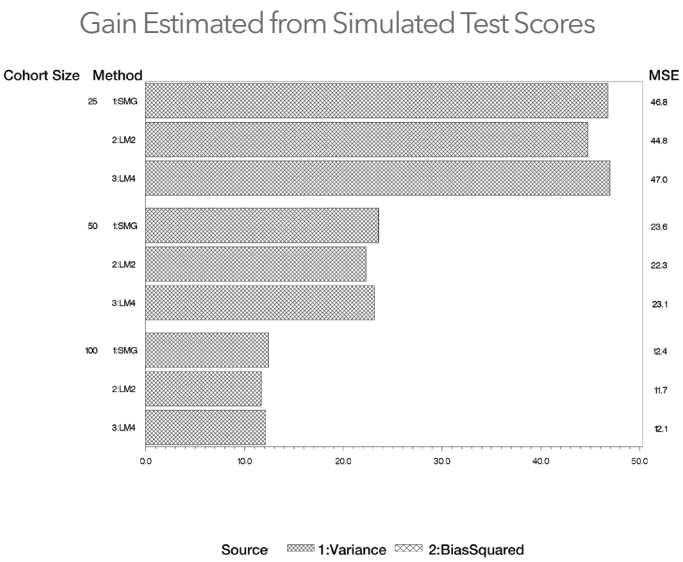

Figure 11. Mean square errors with flat pattern and 20% missing mostly non-randomly.

Results with a Shed Pattern and 20% Missing Non-Randomly. Figures 12 and 13 show results for cohorts of size 25, 50 and 100 students with a downward shed pattern and 20% of the data missing partly non-randomly (Figure 12) or mostly non-randomly (Figure 13). Figures 14 and 15 show comparable results with an upward shed pattern. All four figures tell the same story. That story is that the estimates from the simple model perform *much more poorly* (have much larger mean square error) than those from the more complex models, mostly because of bias. The amount of bias is roughly the same whether the gain pattern is a downward shed or an upward shed, but the direction of the bias is reversed. In the case of a downward shed, students with lower true scores tend to have higher gains, but these students are also more likely to have missing values. Consequently, the estimated gain is too low (negatively biased) because of the loss of the higher gaining students from the data. The district, school, or teacher being evaluated will be judged to be doing a poorer job than they actually are doing. In the case of an upward shed, just the opposite happens. The lower gaining students are more likely to be missing from the data, so the estimated gain is too high, and the district, school, or teacher will get credit for doing a better job than is justified. The longitudinal models, especially the 4-year model, do an impressive job of removing the bias, thus providing a fairer evaluation of the performance of the district, school, or teacher responsible for these students.

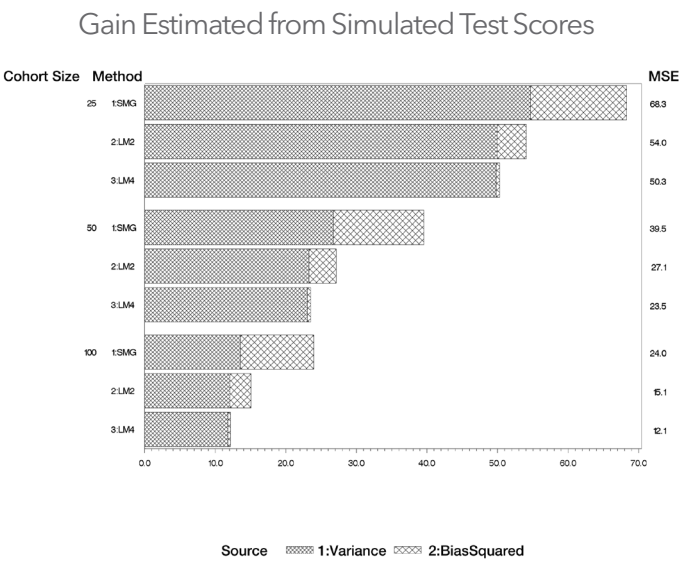

Figure 12. Mean square errors with downward shed and 20% missing partly non-randomly.

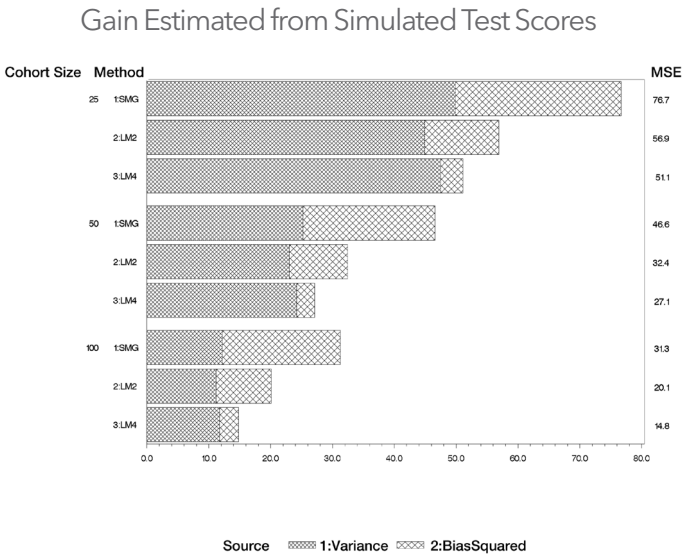

Figure 13. Mean square errors with downward shed and 20% missing partly non-randomly.

As in the earlier examples, the relative advantage of the more complex models increases with cohort size. In this case the increases are very dramatic. In Figure 12 the mean square error of the simple model is 36% higher than that of the 4-year longitudinal model at n=25; it is 68% higher at n=50 and 95% higher (essentially twice as large) at n=100. In Figure 14, the percentages are 32%, 59%, and 101%. In Figure 13, they are 50%, 72%, and 111%. In Figure 15, they are 47%, 75%, and 110%.

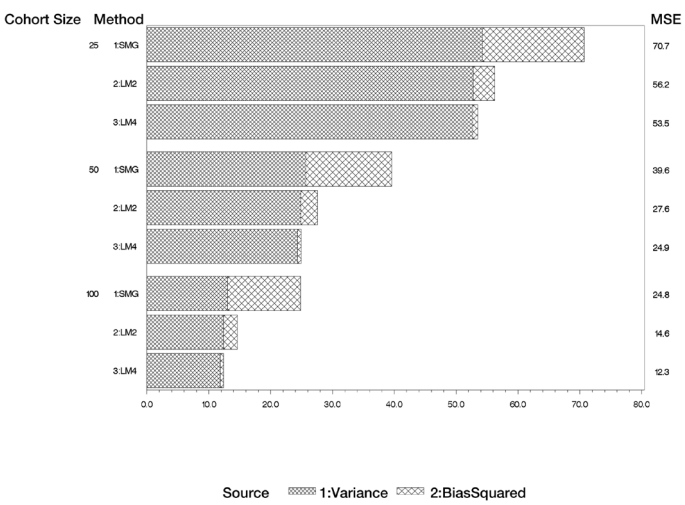

Figure 14. Mean square errors with upward shed and 20% missing partly non-randomly.

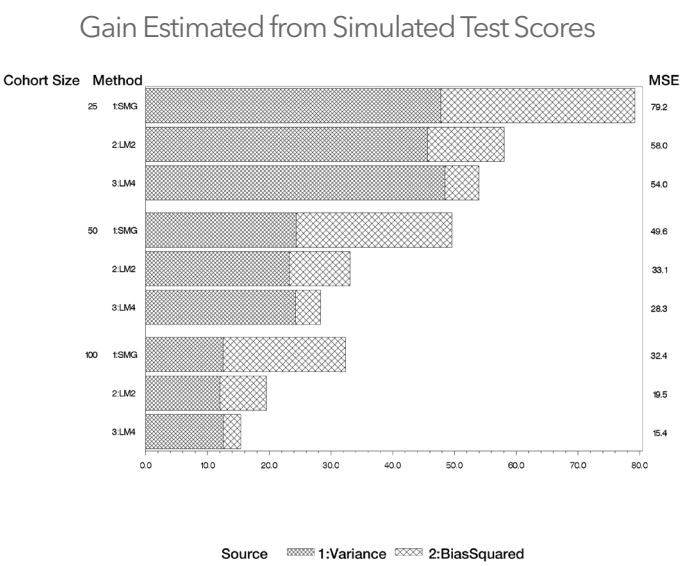

Figure 15. Mean square errors with upward shed and 20% missing partly non-randomly.

4. Conclusions

The important message to be remembered from these results is this: *In a perfect world, things are simple and simple analyses suffice; but in the real world, things are complicated and complex analyses are necessary to obtain results that are fair and reliable*. The "perfect world" is one in which there are no missing data values and all students have the same opportunity to grow academically (a flat gain pattern). In the real world, students miss tests, usually in a non-random fashion, and not all students may have the same opportunity for academic growth. Using a simple model with such real world data can have disastrous consequences, producing biased and imprecise estimates that result in unfair evaluations of district, school, or teacher performance. The 4-year longitudinal model largely eliminates the bias, and even in the less likely scenarios where the simple model estimate is unbiased, the longitudinal model estimates are generally more precise. Finally, the advantage of the complex models over the simple model increases with sample size. Given that, in the real world, sample sizes are likely to be larger than those used in the simulations (especially in the case of district-level analyses), it is likely that the results reported here understate the advantage of using a longitudinal model rather than a simple average-of-observed-gains model.

References

Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. (1996), *SAS® System for Mixed Models*, Cary, NC: SAS Institute Inc.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., and Hamilton, L. (2004), "Models for Value-Added Modeling of Teacher Effects," *Journal of Educational and Behavioral Statistics*, Vol. 29, No. 1, pp. 67-101.

Tekwe, C. D., Carter, R. L., Ma, C-X., Algina, J., Lucas, M. E., Roth, J., Ariet, M., Fisher, T., and Resnick, M. B., (2004), "An Empirical Comparison of Statistical Models for Value-Added Assessment of School Performance," *Journal of Educational and Behavioral Statistics*, Vol. 29, No. 1, pp. 11-36.

To contact your local SAS office, please visit: sas.com/offices

107157_S126217.0614