Racing to the Top: A Treadmill to Nowhere

Walter A. Rosenkrantz (Professor Emeritus)*
Department of Mathematics University of Massachusetts-Amherst
Email: rkrantz@math.umass.edu

June 8, 2012

Abstract

The high stakes testing regime in grades 3-8, required by the No Child Left Behind (NCLB) law generates massive amounts of data that are used to measure not only the performance of the students, but their teachers as well. One study found that across five large urban districts, among teachers who were ranked in the top 20% of effectiveness in the first year, fewer than a third were in that top group the next year, and another third moved all the way down to the bottom 40%, (Baker et.al,[1]). Empirical data, however, are never definitive. A simple statistical model is proposed to explain why the use of student test scores to evaluate teacher effectiveness exhibits the unusual volatility observed in empirical data. Statistical theory suggests that current methods for interpreting the data, based on more complex data models − including more socioeconomic factors− are more likely to be infested with multicollinearity, which makes it difficult to draw reliable conclusions from the data. In particular, as one adds more variables to the model the margin of error associated with estimating them increases, so they become less reliable as a measure of teacher performance .

1 Introduction

The high stakes testing regime in grades 3-8, required by the No Child Left Behind (NCLB) law, measures not only the performance of the students, but their teachers

*W. A. Rosenkrantz is currently a Professorial lecturer in the Department of Statistics at George Washington University

as well. A novel feature of the legislation is that if all students are not proficient in reading and mathematics by 2014 then the schools and their teachers would be punished. Until then schools must demonstrate Adequate Yearly Progress (AYP) according to the standards of the NCLB law. How many standards are required? According to a report in the Washington Post (Outlook, March 12, 2006) one of the Principals interviewed mentioned meeting standards for 36 sub-categories , and another, mentioned 29. This is absurd. Why? Because it is obvious that as the number of categories for which the AYP standard is measured increases-regardless of how they are defined- so do the proportion of schools that fail to meet them. In particular, the Post reported that in 2006 more than 200 Washington-area schools failed to meet the Adequate Yearly Progress (AYP) standards.

In the meantime, with the 2014 deadline looming, a serious problem arose: How does one use the massive amounts of data generated by the standardized tests taken by the students to evaluate their teachers? The Washington Post recently reported that, "... Maryland won a $250 million federal grant [from the Race to the Top fund] with a promise to build a model to evaluate teachers and principals that would be transparent and fair and tie their success for the first time to student test scores and learning" (Chandler, [3]).

The article goes on to note that these efforts have been "bogged down by political infighting". Although it does not mention what teacher evaluation methods have been proposed there's a good chance that Value-added models (VAM) are among them. VAM refers to an increasingly popular methodology for measuring teacher effectiveness, based on a variety of complex statistical techniques originally developed to analyze complex data sets arising in agricultural and industrial quality control. The problems with VAM as a teacher evaluation tool have been skillfully summarized by John Ewing, who writes, " ... it is essential to ask whether the results [from VAM] are consistent year to year. Are the computed teacher effects comparable over successive years for individual teachers? Are value-added models consistent?" (Ewing, [5]). Empirical data in a paper recently published by the Economic Policy Institute (EPI), "Problems with the Use of Student Test Scores to Evaluate Teachers" suggests that the answer is no.

> For a variety of reasons, analyses of VAM results have led researchers to doubt whether the methodology can accurately identify more and less effective teachers. VAM estimates have proven to be unstable across statistical models, years, and classes that teachers teach. One study found that across five large urban districts, among teachers who were ranked in the top 20% of effectiveness in the first year, fewer than a third were in that top group the next year, and another third moved all the way down to the bottom 40%. Another found that teachers effectiveness ratings in one

> year could only predict from 4% to 16% of the variation in such ratings in the following year. Thus, a teacher who appears to be very ineffective in one year might have a dramatically different result the following year. The same dramatic fluctuations were found for teachers ranked at the bottom in the first year of analysis. This runs counter to most peoples notions that the true quality of a teacher is likely to change very little over time and raises questions about whether what is measured is largely a teacher effect or the effect of a wide variety of other factors (Baker et. al, [1]).

In other words, the predictive power of teacher rankings that are based heavily on their students' performance on high stakes tests is quite poor. In particular, these methods of evaluating teachers' effectiveness give little or no weight to socioeconomic factors. Consider, for example, the 2011 test score data of the Washington DC Comprehensive Assessment System exams given annually to students in grades 3 through 8. In Ward 3, the city's wealthiest area, where the median household income is $97,960 the reading proficiency pass rate for the elementary schools is 84%, while in Ward 8, the city's poorest area, where the median household income is $31,188, the pass rate is 28%. Is it any surprise, then, that 35% of the teachers in Ward 3, were rated highly effective, and only 5 % in Ward 8?

Empirical data, however, are never definitive; what is needed is a simple statistical model that explains the surprising variation in rankings of teachers reported in the EPI paper, and shows, with near mathematical certainty, that all VAM models are vulnerable to the unavoidable statistical variation embedded in the raw data and the statistical methods used to analyze them. It is the purpose of this note to illustrate this by adapting a model due to W. Edwards Deming, (Deming, [4]).

2 A Statistical Model for the Observed Variation in Student/Teacher Performance

Reading Ewing's description of VAM, I was reminded of that curious phenomenon in modern finance known as the "performance chasing investor"; these are the investors who pile into hot stocks, mutual funds, and IPOs, while ignoring the well meaning and accurate advice that "past performance is no guarantee of future results." Not surprisingly, this strategy frequently fails because there is a strong element of chance in the performance of the stocks and bonds selected by the mutual fund manager; which explains why some mutual fund managers ranked in the top quarter of their peers in one year rank in the bottom half the next. "Standard & Poor's regular surveys document that most actively managed mutual funds trail their benchmarks

virtually every year, nullifying the statistical value of short term outperformance." (Barrons Business and Financial Weekly, [2])

Unlike a mutual fund manager, however, who can select the stocks on which his fund's performance will be computed, the classroom teacher must accept the students assigned to his class, many of whom come from families with a wide variety special needs that are unknown to the administration, as well as the teacher. In view of its well known deficiencies as an investment strategy, why then would we then use it to evaluate teacher performance? But this is more or less one of the implications of VAM. In the DC Public Schools system, for example, teachers are graded as "highly effective," " effective," or "minimally effective." Teachers classified as minimally effective for two consecutive years are automatically fired.

The unreliability, lack of consistency, and surprising variability of evaluating teachers mainly on the basis of their students' performance on high stakes tests would not have surprised W.E. Deming (1900–1993), originally trained as a physicist but employed as a statistician, and who became a highly respected, if not always welcome, consultant to America's largest corporations. "The basic cause of sickness in American industry and resulting unemployment," he argued, "is failure of top management to manage." This was, and still is, a sharp departure from the prevalent custom of blaming workers first when poor product quality leads to decline in sales, profits and dividends. Among the poor management practices he criticizes most sharply is performance evaluation of employees. *In particular, the most important question is to determine how much of the variation in the teachers' performance is due to the teachers themselves and how much is attributable to the system they work in.* "Fair rating is impossible", writes Deming. "A common fallacy," he continues, "is the supposition that it is possible to rate people; to put them in rank order of performance next year, based on performance last year." To illustrate this point he devised a simple experiment that illustrates "the unbelievable differences between people that must be attributed to action of the system, not to the people."

Example 2.1 *Adapted from W. Edwards Deming ([4], pp. 109–112).*

In some urban school districts it is not unusual, unfortunately, to have 55% of the students non proficient in reading and/or math. Consider, then, a group of 20 fifth grade teachers each of whom is randomly assigned 25 students. Following Deming we model the system as follows: Consider a bowl filled with 550 red beads and 450 white beads. In the industrial context considered by Deming the beads in the bowl represent a shipment of parts from a supplier: the white beads represent the parts that are conforming, that is, they meet the specifications, and the red beads represent the parts that are non-conforming, that is, they fail to meet the specifications. For

our purposes the red beads represent the non proficient students and the white beads the proficient ones. We return to Deming's "experiment": Each worker stirs the beads and then, while blindfolded, inserts a special tool into the mixture with which he draws out (randomly) exactly 25 beads. The variable of interest to the manager is the number of red beads (non proficient students) produced by the employee (teacher). The aim of this process is to produce white beads; equivalently, to produce as few red beads as possible. Employees are rated by the number of red beads produced; the greater the number of red beads produced the lower his ranking. Table 1 displays the result of a statistical simulation (using Mathematica) of this experiment . [**Technical note:** I simulated a binomial random variable with parameters $n = 25 =$ class size, and $p = 0.55 =$ proportion of red beads in the bowl. Statistical details appear in Section 3.]

The first column lists the 20 teachers identified by an integer $i = 1, 2, \ldots, 20$. The second column list the number of students who were labeled non proficient and the third column lists the teacher's rank for year 1. Columns 3 and 4 lists the data for the same set of teachers for year 2. If two, or more ranks, are tied then we assign to each of the ties the average of their ranks. Thus, in year 1 four teachers (2,5,8,10) each had 14 non proficient students. Their corresponding ranks are: 12, 13, 14, 15, with average rank=13.5.

It might be helpful to think of the beads in the bowl as voters divided into two political parties: the "reds" and the "whites". Each entry in columns 2 (Year 1) and 4 (Year 2) represents the number of voters in the sample who identify themselves as "reds." Using standard statistical theory we can compute the margin of error from the polling data and then a confidence interval for the number of red beads in each sample. The total number of red beads in column 2 is 264, so the average number of red beads produced by each worker is $264/20 = 13.2$. Statistical theory-briefly summarized in Section 3- tells us that with probability 0.95 the number of red beads produced by each worker lies within the interval [8.3, 18.1]; this is an example of what statisticians call a 95% confidence interval. Looking at Table 1 we see that almost all the observed variation in the teachers' performance can be accounted for by chance alone. Only the number of non proficient students in Teacher 19's class lies outside the 95% confidence interval; which is what the theory predicts. Returning to the quality control context of Deming's experiment, it is clear that the workers' skills in stirring the beads are irrelevant to the final results, the observed variation between them due solely to chance; or, to paraphrase (slightly) Deming: "It would be difficult to construct physical circumstances so nearly equal for 20 people, yet to the eye, the people vary greatly in performance."

Looking at Table 1 we see that the top 7 teachers (out of 20 and ranked in decreasing order) are: Teachers 11, 9, 18, (6, 12, 13, 14); the last four teachers are tied

for fourth place. Of these seven teachers in Year 1 note that, in Year 2, Teachers 9, 18, 12, and 13 each produced at least 15 non proficient students (red beads); consequently, each of them rank no higher than the median (which equals 14.5) in Year 2. That is, 57% of the teachers ranked above the top half in Year 1, ranked in the bottom half in Year 2.

Lessons Learned: This statistical "thought experiment" helps explain why rating teachers according to the performance of their students on standardized tests of the sort mandated by NCLB is unlikely to reliably distinguish among highly effective, effective, or minimally effective teachers. In particular, "test scores can be affected by many factors−the incoming levels of achievement, the influence of previous teachers, the attitudes of peers and parental support," (Strauss, [8]). Defenders of VAM assert that it takes into account "all the factors that might influence test results," including all those just cited, (Ewing, [5]). Unfortunately, these factors are highly correlated, which implies that the data are very likely infested with *multicollinearity*. Consequently, the estimates of the factor effects of the model are unstable, and sometimes of the wrong sign, (Myers, [7]). Another possible source of difficulty is a statistical analogue of Heisenberg's uncertainty principle: As one adds more factors to the model, the margin of error of their estimates can increase, so they become less reliable. It is worth noting, however, that there are alternative methods for improving teacher and student performance in mathematics. For readers interested in learning about some of them, Liping Ma's penetrating study, "Knowing and Teaching Mathematics," (Ma, [6]), is highly recommended.

Table 1 *Data from Deming's red bead experiment*

Teacher	Number Non-proficient Year 1	Teacher's Rank Year 1	Number Non-proficient Year 2	Teacher's Rank Year 2
1	18	19	17	19.5
2	14	13.5	12	5
3	13	9.5	14	8.5
4	13	9.5	16	16.5
5	14	13.5	11	2.5
6	11	5.5	14	8.5
7	13	9.5	12	5
8	14	13.5	17	19.5
9	10	2.5	15	12.5
10	14	13.5	16	16.5
11	8	1	11	2.5
12	11	5.5	16	16.5
13	11	5.5	16	16.5
14	11	5.5	15	12.5
15	13	9.5	15	12.5
16	15	16	9	1
17	16	17.5	14	8.5
18	10	2.5	15	12.5
19	19	20	12	5
20	16	17.5	14	8.5

3 Some Statistical Theory

Calculation of the limits of variation attributable to the system

Following Deming we now calculate how much of the observed variation is attributable to the system. Let X denote the number of red beads produced by a worker. The exact distribution of X is hypergeometric with parameters n, N, D given by

$$n = 25,\ N = 1,000,\ D = 550,\ p = \frac{550}{1,000} = 0.55.$$

Note that

$$\frac{n}{N} = \frac{25}{1,000} = 0.025 < 0.05,$$

consequently, we can apply the binomial approximation to the hypergeometric and assume that X is $B(x; 25, p)$ distributed. Applying the normal approximation to the binomial, with $n = 25$, yields

$$P(25p - 1.96\sqrt{25p(1-p)} \leq X \leq 25p + 1.96\sqrt{25p(1-p)}) \approx \Phi(1.96) - \Phi(-1.96) = 0.95. \quad (1)$$

This yields a 95% confidence interval $[L, U]$ for the expected number of red beads produced by each worker, where L, U are given by

$$L = 25p - 1.96\sqrt{25p(1-p)},\ U = 25p + 1.96\sqrt{25p(1-p)}. \quad (2)$$

Deming calls confidence intervals of the form $\{L, U\}$ the *limits of variation attributable to the system*. (Instead of the 95% confidence interval used here Deming uses a 99% confidence interval, which is obtained by replacing 1.96 with 2.58 in Equation 2). Let us now apply these theoretical results to the actual data from Table 1. Since management does not know the true proportion of red beads in the mixture it uses the sample proportion denoted $\hat{p}$, which, in Year 1 equals $264/500 = 0.528 = \hat{p}_1$ and in Year 2 equals $280/500 = 0.56 = \hat{p}_2$. With $n = 25$, $\hat{p}_1 = 0.528$

The estimated limits of variation attributed to the system for Year 1 data are given by

$$\begin{aligned} L = 25\hat{p} - 1.96\sqrt{25\hat{p}(1-\hat{p})} &= 13.2 - 4.9 = 8.3; \\ U = 25\hat{p} + 1.96\sqrt{25\hat{p}(1-\hat{p})} &= 13.2 + 4.9 = 18.1. \end{aligned}$$

The estimated limits of variation for the Year 2 data are calculated in the same way yielding the values $L = 9.1$, $U = 18.9$.

References

[1] Eva L. Baker, et. al, Problems with the Use of Student Test Scores to Evaluate Teachers, Economic Policy Institute Briefing Paper#278, August 29, 2010, Washington, DC. http://www.epi.org/publications/entry/bp278

[2] Mike Hogan, The Electronic Investor, Barrons Business and Financial Weekly, August 8, 2011

[3] Michael Alison Chandler, Md. *Teacher evaluation redesign bogs down*, Washington Post, June 4, 2011

[4] W. Edwards Deming, Out of the Crisis, MIT Press, (1982)

[5] John Ewing, Mathematical Intimidation: Driven by the Data, Notices of the American Mathematical Society, vol. 58, number 5, May 2011, pp.667-673

[6] Liping Ma, Knowing and Teaching Mathematics, Lawrence Erlbaum Associates, 1999

[7] R.H. Myers, Classical and Modern Regression With Applications, (2e), PWS-Kent Publishing Co. (1990)

[8] V. Strauss, The Answer Sheet, Washington Post, June 27, 2011