Education Policy Analysis
Archives/Archivos Analíticos de Políticas
Educativas
ISSN: 1068-2341
epaa@alperin.ca
Arizona State University
Estados Unidos

Lohr, Sharon L.
Red Beads and Profound Knowledge: Deming and Quality of Education
Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas, vol. 23, 2015, pp. 1-21
Arizona State University
Arizona, Estados Unidos

education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

Arizona State University

Volume 23 Number 80 August 24th, 2015 ISSN 1068-2341

Red Beads and Profound Knowledge: Deming and Quality of Education

Sharon L. Lohr
Westat
United States

Citation: Lohr, S. (2015). Red beads and profound knowledge: Deming and quality of education. *Education Policy Analysis Archives, 23*(80). http://dx.doi.org/10.14507/epaa.v23.1974

Abstract: Value-added models are being implemented in many states in an attempt to measure the contributions of individual teachers and schools toward students' learning. Scores from these models are increasingly used for high-stakes purposes such as setting compensation, hiring or dismissing teachers, awarding tenure, and closing schools. The statistician W. Edwards Deming wrote extensively about improving quality in education and the damage caused by performance rankings. We examine uses and misuses of value-added models in the context of Deming's System of Profound Knowledge, and discuss contributions a Deming-based perspective and statistical science can make to improving education.
Keywords: educational improvement; educational quality; sampling; statistics; systems approach; teacher evaluation

Cuentas Rojas y Conocimientos Profundos: La Perspectiva de W.E. Deming y la Calidad de la Educación

Resumen: Los modelos de valor añadido se están implementando en muchos estados, en un intento de medir las contribuciones individuales de los profesores y escuelas en el aprendizaje de los estudiantes. Las puntuaciones de estos modelos se utilizan cada vez más para fines de consecuencias severa, tales como el establecimiento de una indemnización, contratación o despido de docentes, la concesión de una escuela a una firma, o el cierre de las escuelas. El estadístico W. Edwards Deming escribió extensamente acerca de la mejora de la calidad en la educación y los daños causados por los

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 1/7/2015
Revisions received: 8/6/2015
Accepted: 8/8/2015

rankings de desempeño. Examinamos usos y abusos de los modelos de valor agregado en el contexto del Sistema de Conocimiento Profundo de Deming, y discutimos las contribuciones de una perspectiva basada en Deming y como la estadística pueden contribuir para mejorar la educación.
Palabras clave: mejora educativa; calidad de la educación; muestreo; estadísticas; sistemas; evaluación docentes

Contas Vermelhas e Conhecimentos Profundos: A Perspectiva de W.E. Deming e a Qualidade da Educação

Resumo: Os modelos de valor agregado estão sendo implementadas em muitos estados, em uma tentativa de medir as contribuições individuais dos professores e das escolas no aprendizagem dos alunos. Dezenas de estes modelos são utilizados cada vez mais para fins de consequências severas, como a determinação da remuneração, contratação e demissão de professores, administração escolar cedidas a empresas, ou o fechamento de escolas. Estatístico W. Edwards Deming escreveu extensivamente sobre como melhorar a qualidade da educação e sobre os danos causados pelos rankings de desempenho. Nós examinamos o uso e abuso de modelos de valor agregado no contexto do Sistema de Deming do conhecimento profundo, e discutimos as contribuições de uma perspectiva baseada em Deming e como as estatísticas podem contribuir para melhorar a educação.
Palavras-chave: melhoria educacional; qualidade da educação; amostragem; Estatística; sistemas; avaliação de professores

Introduction

W. Edwards Deming was tremendously influential in quality improvement, management, survey sampling, and statistical practice. Deming's Fourteen Points for Management (Deming, 1986, Chapter 2; Neave, 1987) summarize his approach for transforming American business and industry. Deming emphasized that the principles in his philosophy for quality improvement, called the System of Profound Knowledge, apply to all organizations: small and large, manufacturing and service. In particular, Deming emphasized the importance of adopting the System of Profound Knowledge in education.

In his last book, *The New Economics for Industry, Government, Education* (Deming, 1994), Deming wrote about his ideal for a system of schools. It is one in which parents, school boards, policymakers, teachers, and students all work together to achieve the aims for the school: "growth and development of children, and preparation for them to contribute to the prosperity of society." In Deming's ideal system of education, "pupils from toddlers on up through the university take joy in learning, free from fear of grades and gold stars," and "teachers take joy in their work, free from fear in ranking." He wrote that "[s]uch a system of schools would be destroyed if some group of schools decided to band together to lobby for their own special interests" (Deming, 1994, pp. 62-63).

There are currently many initiatives for improving the quality of education. Some of these initiatives focus on holding teachers and schools accountable for the learning outcomes of their students. Statistical models called value-added models (VAMs) are commonly used to attempt to estimate the contributions of different teachers and schools toward student test scores. The results of these models are being used or proposed for "high stakes" purposes such as tenure decisions, merit raises, firing teachers, and closing schools. In many states, up to half of a teacher's evaluation depends on estimates from a VAM. There have also been proposals to use these models at the university level to make decisions about financial aid and other resources allocated to universities.

In this paper, I examine VAMs from the viewpoint of Deming's ideal system of education. My thesis is that VAMs need to be viewed within the context of improving the quality of the system. A VAM is a statistical model, and the validity of inferences drawn from it depends on the quality of the data used and on the variances and biases of estimates from the model.

The next section of this paper gives a brief overview of Deming's System of Profound Knowledge and the red bead experiment. The following section reviews the statistical methods in some of the VAMs in current use through an example. Finally, I discuss the contributions to improving education that could be made by adopting a Deming-based perspective and using statistical methods and expertise.

Deming's System of Profound Knowledge and The Red Bead Experiment

The main focus of Deming's System of Profound Knowledge is on understanding the entire system, rather than trying to optimize parts of it separately. The four components of the System of Profound Knowledge – Appreciation for a System, Knowledge about Variation, Theory of Knowledge, and Psychology – work together, and Deming (1994, chapters 3-4) wrote about how each is essential for improving quality.

In the education context, Appreciation for a System refers to viewing the education system as a whole, rather than treating teachers, students, curriculum, facilities, and other parts of the system as separate pieces. Deming, Crawford-Mason, and Dobyns (1993) displayed a flow diagram with the connections among students, teachers, administrators, parents, taxpayers, business, and society. This diagram showed how all of the components of the education system must act together to achieve the aims of the system. Knowledge about Variation relies on statistical theory to understand differences among people, interactions with the system, and the results of studies and experiments. Theory of Knowledge emphasizes the importance of stepping outside of the system to study it. Deming (1994, p. 101) argued that "management in any form is prediction," and that subject-matter experts are needed to make predictions and propose theories for improving the system. By including Psychology as a component of the System of Profound Knowledge, Deming recognized that people are the key to improving quality, and emphasized the importance of allowing people to take pride in their work and to create something of value. Deming's writings regarding the superiority of intrinsic to extrinsic motivation are in agreement with recent literature in psychology, education, and behavioral economics (see, for example, Pink, 2011).

Knowledge about Variation includes the distinction between common cause variation and special cause variation. Applied to the education system, common cause variation is that shared by all the schools or teachers, while special cause variation is that attributable to a specific school or teacher. Deming asked, "When will we understand variability? Children learn at different speeds" (Latzko & Saunders, 1995, p. 124), and he viewed children learning at different speeds as common cause variability. Deming emphasized that attempts to concentrate on one part of the system, without considering the system as a whole, were likely to be detrimental to quality, particularly if those attempts confused common and special causes of variation.

Deming frequently used the red bead experiment to explain sources of variability and the difference between common cause and special cause variation. In this experiment, seminar participants called "willing workers" were given rigid procedures for how to sample 50 beads from a bin containing 800 red and 3,200 white beads. The white beads represented acceptable product while the red beads represented unacceptable product. In one of his four-day seminars, Deming instructed, "You will then take the paddle. The paddle has 50 depressions in it. You will push it down into the beads, down into the beads. Gentle agitation. Are you watching? You will then raise

the paddle, axis horizontal. Tilt it 44 degrees. Any excess beads will roll off. I purposely made some red beads so that you may see what they look like" (Deming et al., 1993, vol. 7).

One by one, each willing worker dipped the paddle into the bin and drew a sample of 50 beads. The worker presented his or her sample to two inspectors, who counted the number of red beads, and a recorder wrote the result on a chart next to the worker's name. After each worker had drawn one sample, representing the first day's production, Deming assessed the results in his role as foreman. He praised and gave merit raises to workers who had produced few red beads, and he placed workers who had produced 12 or more red beads on probation. The workers then proceeded to the second "day" of production, drawing samples and presenting them to the inspectors. Deming again scrutinized the results and noted that some of the workers had improved while others were still on probation. The process repeated for the third and fourth "days." Latzko and Saunders (1995, p. 89) reported Deming's comments about willing worker Sue, who had produced 7 red beads on day 1, 8 on day 2, and 13 on day 3: "Sue just got careless. It sometimes happens after a merit raise. But I like her attitude and am sure she will get better." Deming ascribed the improvement of a worker who had been on probation to "[o]ur progressive discipline." At the end of the third or fourth day, he laid off the worst-performing workers but the change resulted in no lasting reduction of the number of red beads produced and, sadly, the bead-producing company went out of business.

In the red bead experiment, all of the variation was due to the system (i.e., common cause variation), and the workers could not get around that variation no matter what they did. So why did Deming need two hours of his seminar to make this point? It was clear to everyone, on an intellectual level, that the workers receiving merit raises were chosen solely on the basis of random variation. The two-hour format, however, gave the experiment a visual and emotional impact that could not have been conveyed in a lecture about variation. Deming would instill fear in the workers and talk about losing jobs, saying, "If he makes any red beads, see how he did it. Make sure you don't do it" (Walton, 1986, p. 43). The willing workers all knew that the variability was due to the system and not to them, but they were frustrated nonetheless. This emphasized the need for understanding psychology in the System of Profound Knowledge.

The System of Profound Knowledge advocates using data to improve quality. It also emphasizes the importance of having a deep knowledge of the connections among components of the system, understanding the psychology behind people's actions, and considering long-term potential consequences of actions. When using data to improve quality, the Knowledge about Variation component of the System emphasizes the importance of distinguishing between special cause variation and underlying randomness. VAMs should be viewed within the larger picture of improving quality. To see what types of information VAMs can reliably provide for quality improvement, in the next section I review some of the VAMs in common use.

An Introduction to Value-added Models

A VAM attempts to model changes in a student's test scores over time as a function of the teachers who have taught that student.[1] VAMs originated because it was recognized that simply using the most recent test scores of a teacher's students to evaluate that teacher is unfair because students have widely varying backgrounds. Using growth relative to prior scores, or including other covariates in a statistical model, are attempts to account for those backgrounds.

The estimates of value added from the different models are commonly termed "teacher effects," but I would like to emphasize that these are not necessarily causal. In this paper, the term

[1] Some VAMs are used to estimate school effects instead of, or in addition to, teacher effects. In this paper, I discuss only teacher-level estimates and note that the same issues apply to school-level estimates.

"teacher effect" is shorthand for differences at the classroom level that are not explained by other terms in the model. Thus, if a specific teacher usually teaches students with special learning needs, and there is no information about these special needs in the database used for data analysis, then we cannot distinguish the effects of the teacher from the characteristics of the students instructed. Similarly, if a teacher regularly teaches in a classroom that is in poor physical condition or that has insufficient supplies or resources for teaching, then the estimated teacher effect is confounded with the effect of the classroom condition.

Four basic types of VAMs are illustrated in this paper using data published by McCaffrey and Lockwood (2011). This data set contains up to five years of test scores from more than 9,000 students, with each score linked to the teacher who instructed the student in that year. Value-added (VA) scores from different models are calculated for a specific teacher in year 3, called Teacher A, who was selected at random from the set of all 306 year-3 teachers. The mathematical formulas for these models, and more detailed discussions of the model properties, are presented in Guarino, Reckase, and Wooldridge (2015), Lohr (2012), McCaffrey, Lockwood, Koretz, and Hamilton (2003), and McCaffrey, Lockwood, Koretz, Louis, and Hamilton (2004). In this paper, I illustrate the simplest forms of the models graphically, and note that many variations of the models exist.

Regression-Type Models

The first three models are based on a regression analysis relating year-3 test scores to year-2 test scores. Figure 1(a) shows the scatterplot of the year-3 test scores vs. the year-2 test scores for the set of all students who have scores for both years, together with the least squares (LS) linear regression line. The VA score for a teacher is computed by comparing the test scores of the teacher's students in year 3 with the scores that are predicted from the regression model adopted.

Figure 1(b) shows the scores of students of Teacher A, along with the residuals for those students from the LS regression line. The VA score for Teacher A using the average residual method is the average of the residuals for the students of Teacher A. For all but one of the students of Teacher A, the actual test score in year 3 is not as high as the prediction from the LS line; thus, Teacher A has a negative VA score. The residuals or the VA scores for this method may be scaled or normalized if desired.

The second regression-type model uses analysis of covariance (ANCOVA), with a common slope for the previous year test score and an indicator variable for each teacher in the data set. The VA score for teacher A is calculated from the coefficient of teacher A's indicator variable. Figure 1(c) shows the VA score as the difference between the line for Teacher A and the average of the lines for all of the teachers. For Teacher A, the VA score from the ANCOVA method is again negative.

The student growth percentiles (SGP) (Betebenner, 2009) model is sometimes referred to as a "growth" model rather than a VAM, but it, like the average residual and ANCOVA models, relates the current year score to the prior year score. Quantile regression models are fit to predict the qth percentile of the year-3 score from the year-2 score, for $q = 1, \ldots, 99$. Either linear or nonparametric quantile regression models can be used. Figure 1(d) displays the test scores for the students of Teacher A along with the quantile regression curves for $q = 95, 75, 50, 25$, and 5, fit using cubic splines. For this data set, the median regression line is practically identical to the LS line from Figure 1(a, b). In other data sets, however, the parametric and nonparametric regressions may differ.

The quantile regression curves are used to find the closest predicted percentile for each student's year-3 score, conditionally on his or her year-2 score. The first three scores for students of Teacher A in Figure 1(d), starting from the left, fall at the 19th, 75th, and 41st predicted percentiles, respectively. The VA score for Teacher A equals the median (some states use the mean) of the

predicted percentiles for the students of Teacher A. A teacher with many students who score at above-median levels in year 3, adjusted for their year-2 scores, will have a high VA score from the SGP model; a teacher such as Teacher A whose students mostly score below the median in year 3 will have a low VA score.

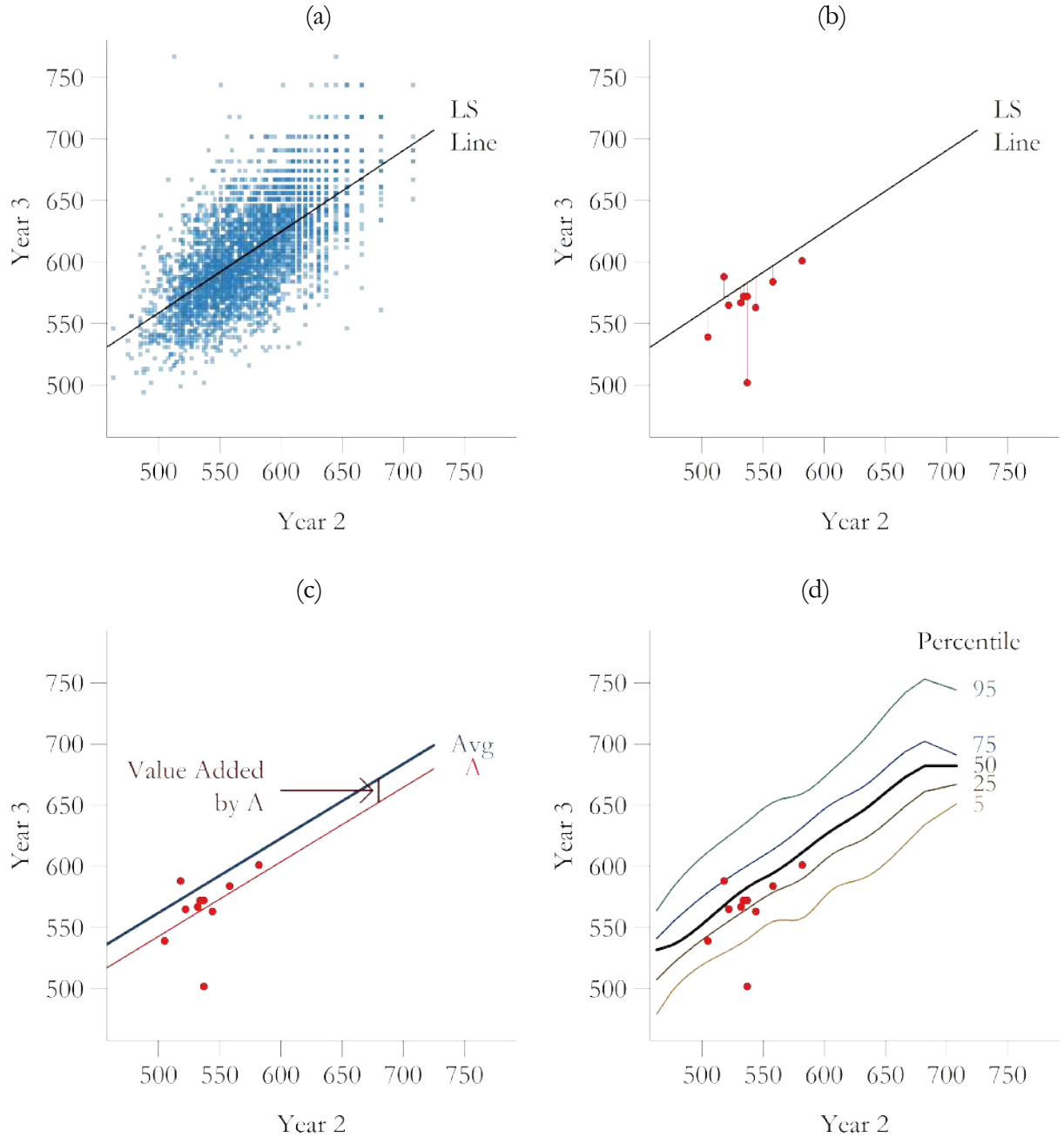

Figure 1. (a) Scatterplot of year-3 test scores vs. year-2 test scores for all students, with the least squares (LS) regression line. The shading is proportional to the number of students with test scores at that particular value. (b) Scores of the students of Teacher A, with residuals (shown by the vertical lines). (c) Analysis of covariance (ANCOVA) line for Teacher A, together with the average of all teachers' ANCOVA lines. (d) Quantile regression lines, used for student growth percentiles (SGP).

The description above presented the simplest forms of each model. In practice, the models can include additional previous test scores as well as other covariates that are thought to adjust for other aspects of a student's background. The choice of covariates varies across locations. A model that has been used in Chicago, for example, included covariates for gender, race/ethnicity, and low-income status (Chicago Public Schools, 2014) while the model used in Florida in 2013 included different covariates (American Institutes for Research, 2013).

All of the values of the covariates need to be known (or imputed) for a student in order to include that student in a regression-type model. Thus, if the model uses the year-1 scores as well as the year-2 scores as covariates, the complete record of test scores for years 1, 2, and 3 is needed for a student to be included in the model-fitting. A student who transfers to the school district during year 2, for example, might not have enough information to be included in the calculations of the VA score. If students who transfer have different test score trajectories, then omitting them from the model may lead to bias in the VA score estimates.

For the average residual and ANCOVA models, the teacher effects can be treated as fixed or random. Usually random effects are used for teachers, and the VA scores for teachers of small classes are "shrunk" toward the mean.

Persistence Models

The fourth type of model, called a persistence model, is different. Instead of regressing the year-3 score on previous scores, the analyst considers the vector of a student's test scores for all years as a multivariate response. The effect of a teacher in year 3 is assumed to persist over time. In other words, the year-3 teacher not only is associated with the gains in the year-3 score but also has a carryover effect to year 4, year 5, and subsequent years. Both regression-type and persistence models use the test scores of the students of year-3 teacher A when calculating teacher A's VA score. While regression-type models use only the history of the students (their background covariates and scores in year 2 and possibly year 1), persistence models also consider the future test scores of these students. The motivation behind persistence models is the idea that an excellent teacher, or an awful teacher, will have a lasting effect on his or her students. That is, the students of an excellent teacher will perform better than expected on the year-3 test, and that superior performance will persist into the future.

Figure 2 illustrates a different view of data that is used for a persistence model, in which the horizontal axis represents the year or grade, and the vertical axis gives the test score for each year. The scores are generally assumed to be vertically equated and scaled (Briggs & Domingue, 2013). The difference between the student score and the overall mean at each year is modeled as the sum of three latent components. The first component is an overall latent effect for that student, assumed to be constant over time. That component includes home environment, poverty, motivation, extracurricular opportunities, and other unmeasured factors that might lead to the student being above or below the overall mean. The second component is the sum of the effects of the teachers who have taught the student to date. Finally, there is a component of random variation, which includes everything that is not explained by the other terms. Thus, (student score, year 1) = (mean score, year 1) + latent student effect + effect of year-1 teacher + noise. The student score for year 2 is modeled similarly, with terms for the effect of the year-1 teacher as well as the year-2 teacher. With the accumulating teacher effects, (student score, year 3) = (mean score, year 3) + latent student effect + effect of year-1 teacher + effect of year-2 teacher + effect of year-3 teacher + noise, and the student scores for subsequent years are modeled similarly.

Several types of persistence models are discussed in the literature. The complete persistence (CP) model (Sanders, Saxton, & Horn, 1997) assumes that the effect of a teacher in year 3 is the

same for all subsequent years. Other models, such as the variable persistence model (McCaffrey et al., 2004), allow the effect of a teacher to diminish over time. The generalized persistence (GP) model (Mariano, McCaffrey, & Lockwood, 2010) estimates multiple effects for a teacher in year 3: one effect for the teacher on the student scores in year 3, a different effect on the scores of the teacher's students in year 4, and yet another effect on the scores in year 5. For the illustration in this paper, the GP model was fit with one effect of the year-3 teacher on year-3 student scores, and a separate common effect for years 4 and 5.

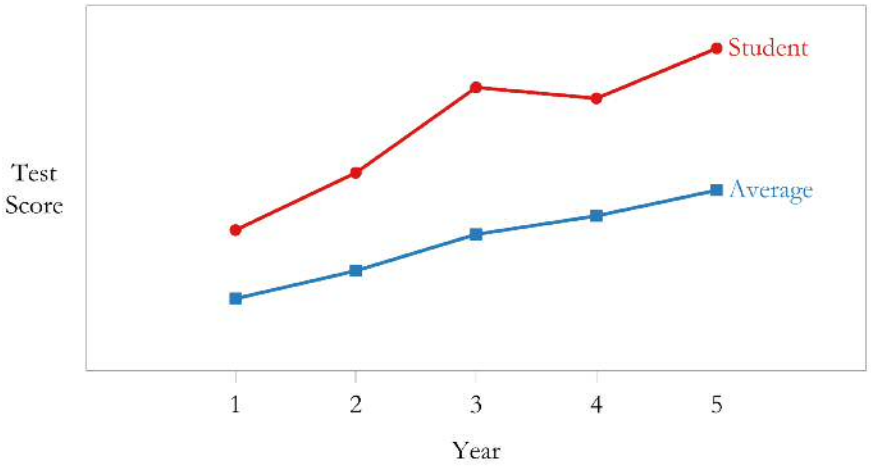

Figure 2. Average score of all students for each year (blue line) and score for hypothetical student for each year (red line).

Figure 3(a) presents a line plot of all of the students of Teacher A, for all five years of data. Each red line in Figure 3(a) displays the test scores available for one of the students who was instructed by Teacher A in year 3, so that each student's pattern over time can be seen. Some students have missing data (e.g., the student with the lowest score in year 3 is missing the scores for years 4 and 5). Persistence models, unlike the regression-type models, allow students with partial data to be included in the analysis.

All of the students in Figure 3(a) have the same teacher – Teacher A – in year 3. But the students come from different classes in years 1 and 2, and they may be dispersed among different teachers in subsequent years. Although the scores of Teacher A's students are a bit below the average in years 2 and 3, the scores are again approximately evenly distributed about the average in years 4 and 5. Because a persistence model allows each teacher's effect to carry over to subsequent years, it is difficult to tell from the graph whether the improvement in year 4 would be credited to the year-3 teacher, to the teachers of these students in year 4, or perhaps to teachers who instructed these students in year 1 or 2.

The teacher effects depend in a complex way on the assignments of students to teachers for all five years, and estimating which parts of student changes in test scores are attributed to different teachers is challenging. The model typically used to apportion the test scores among different teachers is a multivariate response mixed effects model, with random effects for the teachers. The covariance structure is more complicated than in a hierarchical model, and special software is needed to compute VA scores. Karl, Yang, and Lohr (2013) and Wright, White, Sanders, and Rivers (2010)

describe software packages that may be used to compute parameter and VA estimates from persistence models.

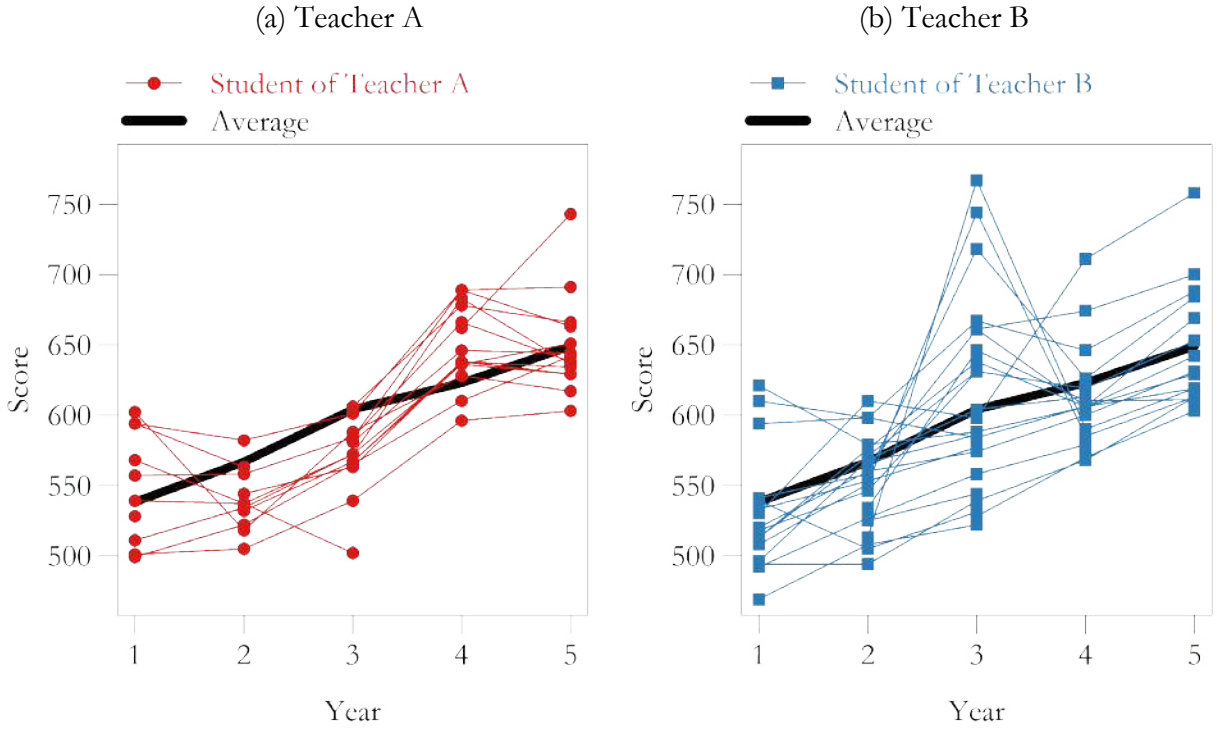

Figure 3. (a) Scores of the students of Teacher A, for all 5 years. (b) Scores of the students of Teacher B, for all 5 years.
Note: Teachers A and B taught their respective students during the period between the year-2 assessment and the year-3 assessment. The dark solid line connects the average scores of all students at each year.

Figure 3(b) displays a similar line plot of the scores for students of Teacher B, who also teaches in year 3. There is a twist here, though. In the earlier discussion of these models, it was assumed that the students disperse to different teachers in subsequent years. In this case, however, almost all of the students of Teacher B in year 3 go on to have the same teacher in year 4. It is thus difficult to separate the effect of Teacher B from that of the year-4 teacher of these students. Several of the high scores for Teacher B's students in year 3 decrease substantially in year 4. Because almost all of Teacher B's students who have data for year 4 are instructed by the same year-4 teacher, that year-4 teacher ends up with one of the lowest VA scores among the year-4 teachers. If the year-3 test scores for the students of Teacher B were inflated through cheating or one of the manipulation schemes described in Lohr (2012), the year-4 teacher following Teacher B would be penalized for that inflation.

The VA scores are highly (but not perfectly) correlated across the different models, but, as Goldhaber, Walch, and Gabele (2014) demonstrated, those high correlations do not prevent some teachers from being ranked differently by different models. The models all produce estimates that are on different scales; to compare them, Table 1 displays the percentile ranks for Teachers A and B from the different models. For these teachers, some methods give very different rankings. In particular, Teacher B with the outlier students is ranked at about the 80th percentile if the outliers are given their full value (as in the residual, ANCOVA, and persistence models) but at the 30th percentile with the SGP method.

Table 1
Percentile Ranks for Teachers A and B from Different Models

Model	Residual	ANCOVA	Student Growth Percentiles	Complete Persistence	Generalized Persistence, year 3	Generalized Persistence, years 4-5
Teacher A	21	19	28	42	7	50
Teacher B	85	83	30	94	74	59

The four types of models – average residual, ANCOVA, SGP, and persistence – have several features in common. First, as usually implemented, the measures of student growth are usually scores on a standardized test. The models do not directly address teacher contributions to other outcomes, and also do not directly measure long-term outcomes. Second, most analyses are based on observational data. The models attempt to control for other factors through covariates and the use of student growth, but it is possible that the teacher effects are actually unmeasured characteristics of the students who are in that class or other classroom environmental factors.

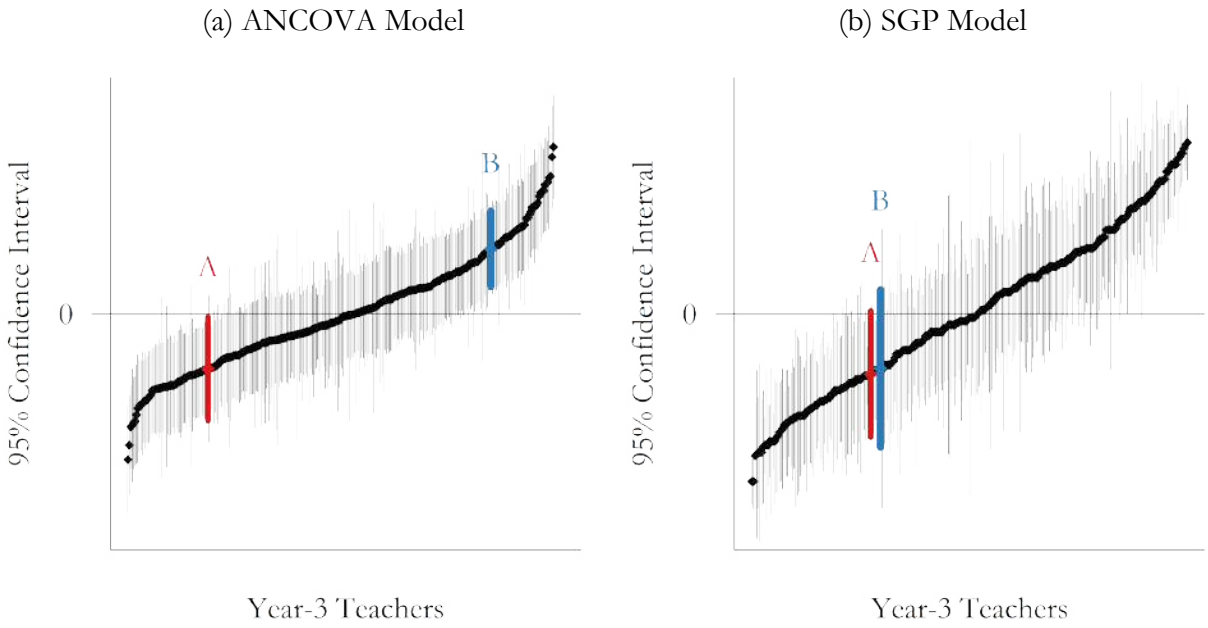

Figure 4. Value-added scores (dots) and 95% confidence intervals (vertical lines) for the 306 year-3 teachers, using (a) analysis of covariance (ANCOVA) model with random effects for the teachers and (b) student growth percentiles (SGP) model.

The third common feature is that most of these methods have low precision for the estimated teacher effects. Goldstein and Spiegelhalter (1996) emphasized the need to report standard errors or interval estimates for measures used to compare institutions. Figure 4 displays interval estimates for the teacher effects from the ANCOVA model with random effects and the SGP model. Each vertical line represents the confidence interval for one teacher, with the teachers arranged from lowest to highest VA score. Many of the confidence intervals include zero; they are wide, and would be even wider if an adjustment were made for multiple testing. Because confidence

intervals are always interpreted within the framework of the model, they can change if a different model is used. And, in fact, the confidence interval changes greatly for Teacher B, who is near the top for the ANCOVA model but is practically indistinguishable from Teacher A for the SGP model. I think this demonstrates that great care is needed in using these models, with consideration of their statistical properties and assumptions.

The confidence intervals in Figure 4 include the uncertainty due to the variance of the estimate within the context of the model and observed data. They do not include uncertainty from bias, which is the difference between the expected value of the estimated VA score for a teacher and the "true" measure of VA for that teacher. As stated by the American Statistical Association (ASA, 2014), the "validity of the VAM scores as a measure of teacher contributions depends on how well the particular regression model adopted adjusts for other factors that might systematically affect, or bias, a teacher's VAM score." Standard errors can be reduced by using multiple years of data, but this will not reduce potential bias that arises from attributing other sources of classroom variation to teachers.

Profound Knowledge and Value-added Models

Deming's System of Profound Knowledge has direct implications for the use of VAMs. I first briefly discuss Deming's background and work in sampling and quality improvement, and then describe how Deming's theories apply to education and VAMs.

Biographical Influences

Deming's System of Profound Knowledge resulted from his many years of experience as a physicist, statistician, and teacher. Mann (1994, p. 366) wrote that "throughout his [Deming's] life, he championed the belief that statistical theory shows how mathematics, judgment, and substantive knowledge work together to the best advantage." Austenfeld (2001), Gabor (1990), and Kilian (1992) described Deming's early years and training in electrical engineering, mathematics, and physics. During the summers while studying for his PhD in physics, Deming worked at the Western Electric Hawthorne plant, where he became familiar firsthand with the damage caused by numerical quotas and piecework pay. The women employed at the Hawthorne plant were paid by the piece and their pay was reduced if the piece failed inspection (Walton, 1986, p. 6). He also became familiar with the quality improvement ideas of Walter Shewhart during his time at Western Electric. Deming wrote, "What I learned at the Hawthorne plant made an impression for the rest of my life" (Kilian, 1992, p. 174). He later included "eliminate numerical quotas for the work force and numerical goals for management" among his Fourteen Points for Management (Deming, 1986, outside back cover).

In 1939, based on his work in sampling at the U.S. Department of Agriculture, Deming was invited to lead the U.S. Census Bureau's program of using sampling with the 1940 census to obtain information from a 5% sample of persons on additional topics. This resulted in the first census long form data, which ultimately led to the American Community Survey. Duncan and Shelton (1978) described the shift to probability sampling in U.S. government data collection, led by Deming and his colleagues at the Census Bureau, as a "revolution." Deming's (1950) groundbreaking book on sampling emphasized that the statistician needs to be aware of *all* sources of error and not just sampling errors. Deming recommended probability sampling to obtain more accurate data at lower cost. In one of the many quality improvement procedures that Deming instituted at the Census Bureau, he replaced 100% inspection of punch cards by inspection of a sample along with statistical process control techniques. The procedure resulted in fewer errors, lower costs, and faster completion (Deming & Geoffrey, 1941). During this period, Deming also taught intensive courses on Shewhart's methods for quality improvement in the War Training Program (Mann, 1995).

Deming is famous for his lectures in Japan on statistical quality control, which Mann (1995, p. 53) credited with providing "the critical impetus for changing the image of Japanese products." The original purpose of his 1950 trip to Japan, however, was to provide advice on sampling techniques for the 1951 Japanese census. While Deming was planning his 1950 trip as a consultant on sampling, the Union of Japanese Scientists and Engineers invited him to talk about statistical quality control during that visit. In the 1950 lectures, Deming taught what later became known as the Plan-Do-Study-Act (PDSA) cycle, based on Shewhart's cycle for learning and improvement.[2] Afterwards, he continued his work as a statistical consultant, both in Japan and the United States. He became a household name in the United States after NBC featured him in a television program in 1980 called "If Japan Can … Why Can't We?" Deming was 80 years old at the time, and he continued presenting seminars and consulting on statistics and quality improvement until his death in 1993.

System of Profound Knowledge and Value-added Models

The themes from Deming's career and the System of Profound Knowledge apply directly to education and provide guidance for using VAMs. Deming of course did not write about today's accountability initiatives in education, but we do have his views on a program proposed in 1991 called "America 2000." This program called for a 90% high school graduation rate by 2000, a merit schools program in which "individual schools that make notable progress toward the national education goals deserve to be rewarded," and a "public reporting system on the performance of education institutions and systems" (U.S. Department of Education, 1991, pp. 23, 37). The report stated, "There's no place for a no-fault attitude in our schools. It's time we held our schools – and ourselves – accountable for results" (pp. 3-4).

Deming (1994, p. 45) called the America 2000 program a "horrible example of numerical goals in public places." First, there was a numerical goal of increasing the graduation rate to 90%, but no one provided a method for how this was to be accomplished. He asked, "Why stop at 90? If you don't have to have a method, why not make it 95? 98?" (Orsini, 2013, p. 56). Second, the program relied almost exclusively on ranking schools and publishing results so that citizens would demand improvement from the "underperformers." One common theme in all of Deming's writings on quality improvement is that people should not be ranked.

Deming (1994, p. 169) used the red bead experiment to demonstrate that ranking people in a stable system "is wrong and demoralizing, as it is actually merely ranking the effect of the process on people." The worker producing the fewest red beads did so purely as a result of the common-cause variation due to system-level characteristics beyond the worker's control. Deming also wrote about the connection between understanding variation and the merit rating system: "If psychologists understood variation, as learned in the experiment on the Red Beads, they could no longer participate in continual refinements of instruments for rating people" (Orsini, 2013, p. 70). Ranking induces competition, rather than cooperation, among people to the detriment of the organization (Deming, 1994, p. 148).

Deming's views of ranking arose from statistical principles. He argued that one can identify and give special attention to people who fall outside of the control limits, but people between the control limits should not be ranked. He acknowledged that there is a natural desire to assign blame

[2] The PDSA cycle presented a view of quality improvement as a wheel in which the steps of Plan (propose an idea for improvement), Do (carry out an experiment to test the idea), Study (analyze the results), and Act (adopt or abandon the idea, or repeat the experiment with modifications) are repeated over and over. The PDSA cycle depends on the Theory of Knowledge because theory is needed to propose and study modifications to the system.

when something goes wrong: "Anything bad that happens, it might seem, is somebody's fault, and it wouldn't have happened if he had done his job right" (Orsini 2013, p. 247). This view of blame, however, is not supported by the statistical evidence, because in many cases people are ranked on the basis of random variation: "There is no harm in a lottery, so far as I know, provided it is called a lottery. To call it an award of merit when the selection is merely a lottery, however, is to demoralize the whole force, prize winners included" (Deming, 1986, p. 275).

Much has been written about possible unintended consequences of test-based accountability systems (see, for example, Amrein-Beardsley, Berliner, & Rideau, 2010; Baker et al., 2010; Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012; Johnson, 2015; Koretz, 2008), including increased time and resources spent on testing, teaching to the test, cheating, effects on teacher workforce or motivation, and effects on long-term student outcomes such as love of learning. These potential consequences result from focusing on one part of the system instead of viewing the system as a whole as called for in Appreciation for a System. A discussion of these potential consequences is beyond the scope of this paper; instead, I will focus on statistical reasons for why ranking has harmful effects.

We know that reacting to random variation as though it were a special cause can result in increased variability and thus decrease quality (Deming, 1986, pp. 327-329). Colleagues may see that teacher B (from the example in Table 1 and Figures 3 and 4) is ranked highly, and try to adopt teacher B's methods, even though those methods might have nothing to do with the students' test scores. Deming referred to the practice of reacting to common cause variation as if it were from a special cause as "tampering." Dismissing teachers who by chance have low VA scores is an example of tampering. Deming wrote that it is inevitable that someone will receive a low rating in any rating system (Orsini, 2013, p. 173).

A major reason Deming opposed ranking was because "[f]ear invites wrong figures…. To keep his job, anyone may present to his boss only good news" (Deming, 1994, p. 94). Moore (2010) performed a variation of the red bead experiment in which he told the willing workers that they would be fired if they produced any batch with 5 or more red beads, and then moved the bin out of sight. In each repetition of this experiment, the willing workers came back with fewer than 5 red beads but reported no violation of the procedures. Moore concluded that "willing workers, when faced with the need to preserve their livelihood, have three choices: improve the system, distort the system, or distort the data." We need good data to improve the system, and we cannot get those data if they are being distorted.

I am not talking about cheating primarily, although there have been numerous instances in which states or districts have nullified test results because of suspected cheating (Government Accountability Office, 2013). No one condones cheating. Even without cheating, though, it is very easy to take advantage of the unexplained variability in the system to manipulate one's ranking through such activities as carefully choosing students, encouraging students to drop out or be excluded from tests, exploiting the shape of the functional form or the covariates, or other methods (Lohr, 2012). And there are cascading effects of these false figures, especially on people who do not try to game the system. A grade-4 teacher with 20% or more of students coming from grade-3 teachers who cheated or manipulated the system can easily end up near the bottom of the rankings.

I think that one of the biggest costs of the fear created by ranking and performance standards is distrust of statistical methods. This data dread leads some people to not want to collect data or use statistical methods at all, for fear the data will be misused. Some people have attacked the statistical methods themselves. A frequent objection is that the models were developed for other applications – Lee, Sridharan, and Sung (2014), for example, questioned the use of CP models

because they were "originally developed to study genetic trends in cattle" – and therefore, how can they be applied to education?[3]

This argument ignores the true value of the statistical discipline and the reason that statistical science can make such immense contributions to all fields: Statistical methods transfer from one application to another, as long as the data structures are similar (Deming, 1965). Mixed models may have been developed for genetics applications, but they can be applied in education, medicine, agriculture, small area estimation, or any area in which observations have a clustered structure. Statisticians can draw on this wealth of knowledge and theory across application areas when working with data, and that is why contributions from the statistical profession are needed for the challenges in improving education.

Contributions from Statistical Science

Statistical Research and Practice

Statistics is the only profession that specializes in the study of variation. The whole training and ethos of statistical science is in designing studies and extracting information from data, without having a vested interest in a specific outcome (Deming, 1965). Statisticians do not have a stake in a specific research hypothesis – their interest is that the statistical methods used for reaching the conclusion are employed properly.

One of the most valuable contributions of VAMs is the knowledge they provide about variation. Many empirical studies (see, for example, Haertel, 2013; Nye, Kanstantopoulos, & Hedges, 2004; and Schochet & Chiang, 2013) find that teachers account for less than 15% of the variability in test score gains.[4] Remember, by teachers I mean unexplained variability at the classroom level, so we cannot necessarily parse out the portion due to teachers and the portion due to other classroom-level factors. Nevertheless, these models give us an idea of where we might want to concentrate efforts if the goal is to reduce variability.

The various models used in value-added are complicated and require a deep level of statistical expertise to use and interpret. Some of the statistical issues that arise are whether to use fixed or random effects for the teachers, choice of regression or persistence model, whether to use parametric or nonparametric regression, and how to account for missing data. There is also debate over which covariates should be included, if any. Typically, the CP model does not include other covariates, and there is debate about whether a student's set of test scores captures everything one would need to know about the latent student effect. A number of studies have shown that the teacher effect estimates can change when different covariates or models are used (see, for example, Briggs & Domingue, 2011; Newton, Darling-Hammond, Haertel, & Thomas, 2010), and that teacher effect estimates for different outcomes can be negatively correlated (Broatch & Lohr, 2012).

The statistics discipline can make great contributions in assessing the appropriateness of VAMs for different purposes. These contributions include explicating the assumptions for the models and the implications of the statistical properties of the estimates. VAMs, in my opinion, should not be used to rank teachers. As argued earlier and by the ASA (2014), the estimates of

---

[3] If this argument were valid, then no other fields could use the *t* statistic because it was originally developed to study beer.

[4] The ASA (2014) noted that this is not the same thing as saying that teachers do not make a difference. The analysis only looks at variability among teachers with respect to the outcome of test score gains. It does not consider any of the other ways that a teacher might affect a student's life, and discounts possible effects of cooperation and improving the system. Cooperation and mentorship programs often reduce variability among teachers.

individual teachers' value added from these models can be unstable and subject to bias; Deming argued that using data to rank people frequently results in the data being unreliable for the purpose of monitoring the system.[5] However, this does not mean that the statistical models underlying VAMs should not be used at all. These models are quite good at providing information about system-level conditions, and the different models generally provide consistent information on those conditions. If we have uncorrupted data that accurately measure the outcomes of interest, VAMs can provide valuable information for quality improvement.

There is also a huge amount of statistical research that is needed for models that can be used with education data. To mention just one of these research areas, note the stripes along the top and the right side of the scatterplot in Figure 1(a). These stripes are the result of the scaling of the data. The raw data from test scores are typically scaled and transformed so that the transformed values follow an approximate normal distribution. But in many cases, the raw scores have a negatively skewed distribution. Thus, the distance between the two stripes on the right or at the top of Figure 1(a) is from a student getting one more question correct on the test. This scaling can have a large effect on estimates and their properties.

Nonparametric mixed models are another promising avenue for more statistical research, because there are often ample amounts of data to allow nonparametric function estimation. As designed experiments become more common in education, research is needed on experimental designs that can be used as part of an ongoing quality improvement effort. Table 2 lists these and a few of the other areas in which there is a need for statistical work. Some of these topics are being researched right now, but much more needs to be explored.

Table 2
A Few Statistical Research Problems in VAMs

Design and Analysis of Studies	**Analysis of Data**	**Analysis and Interpretation of Results**
Design of Experiments	Nonparametric Methods	Visualization of Data
Measurement Error	Computational Methods	Confidentiality Protection
Missing Data	Network Models	Multiple Responses
Data Structure	Scaling	Evaluating Effects of Interventions on Mean *and* Variance
Mixing of Students	Nonnormal Data	Robustness of Pedagogical Methods

VAMs and similar models are powerful tools for use in education. They can capture many of the features of the data structure (although not all, which is why more statistical research is needed). I think one of the most valuable features of these models is their ability to investigate the effects of programs at the student, teacher, and school levels. Typically, most evaluation of educational innovations focuses on the impact on mean scores of students, but these models also allow for studying impacts on variances and robustness: What methods work well even when they are not implemented under ideal conditions with experienced teachers?

[5] A similar idea is commonly referred to in the social sciences as Campbell's Law (Campbell, 1976).

Probability Sampling

As argued earlier, VAMs can be used to provide information for improving education. Almost all of the benefits I have discussed from using VAMs as part of a quality improvement effort could be realized by collecting data on a probability sample of schools or districts instead of doing a census. Deming (1947), in the article in which he coined the term "probability sample," argued that not only is a probability sample less expensive but it also can avoid many of the biases associated with haphazardly collected large data sets. Deming frequently wrote about the need to cease 100% inspection because it is wasteful and does nothing to improve quality because it is not accompanied by a method telling how to improve quality. He quoted Harold Dodge's aphorism that "You can not inspect quality into a product" (Deming, 1986, p. 29). He advocated using samples to monitor the process, then using the resources that had previously been devoted to mass inspection to improve the system through designed experiments and studies. The PDSA cycle provides a model for planning and learning from experiments or studies on teaching innovations, mentoring or peer assistance programs, school reorganizations, teacher training methods, incentives, or other innovations suggested by theory.

Fitting VAMs through a probability sample would decrease costs dramatically, because only a sample of schools would need to have yearly assessments, and that sample could be rotated to reduce burden. This was the motivation behind using probability sampling to collect information on additional questions in the 1940 census. As Deming (1950, p. 3) said, "The statistician's aim in designing surveys and experiments is to meet a desired degree of reliability at the lowest possible cost under the existing budgetary, administrative, and physical limitations within which the work must be conducted. In other words, the aim is *efficiency* – the most information (smallest error) for the money." He emphasized (1950, Chapter 2) that in many cases, a sample produces more accurate estimates than a complete census because a sample can be carefully designed to reduce biases. Measurement on a sample, instead of on everyone, would shift the use of the models from the current emphasis on accountability and evaluation to an emphasis on how to improve the quality of the system as a whole. The models would be used to estimate system-level parameters and identify potential areas for improvement rather than estimating a VA score for every individual teacher.[6]

Use of probability samples to monitor the system also fits better with Deming's idea to improve the system constantly and forever. Many people (see, for example, Thomas, Wingert, Conant, & Register, 2010) have suggested that the main way to improve quality is to fire the bad teachers, and then the education system will be fixed. But in a Deming-based view, quality improvement never stops. It is a process that goes on forever and ever, and the system can always be improved more.

Continual Improvement

Deming emphasized the importance of what we now call lifelong learning, long before the term was even coined (Schwinn & Schwinn, 1995, p. 19). He expounded on the goal of lifelong learning in one of the Fourteen Points: "Point 13 is self-improvement. A program of self-improvement: education, improvement in other ways, helping people to live better. Education in whatever one's fancy might take him into. History, music, archaeology, anything whatever. Keep

[6] In addition to the research problems mentioned in Table 2, statistical research is needed to develop probability sampling designs and models that allow estimation of system-level parameters. The VAMs discussed in this paper require at least two years of data from each student in the sample, and the probability sample would need to be carefully designed to allow measures of student change while not overburdening the schools selected to be in the sample.

people's minds developing. Education need not be connected with the work, it may be better if it's not. Point number 6 is training and re-training for the job. Point number 13 is elevating people's minds. No organization can survive with just good people. You need people that are improving. The problem isn't people at the top. In any profession, the good ones are hard to find. Help people be good ones" (Deming et al., 1993, vol. 2). He did not think of people as static, but thought everyone should keep learning and keep improving.

Discussion

Deming' System of Profound Knowledge has many lessons for the use of VAMs. VAMs are statistical models, and their results need to be interpreted within the context of the statistical properties, such as standard errors and biases, of the models and estimates. Deming was adamant that results from statistical models such as VAMs should not be used rank individual workers, arguing that such ranking ends up rewarding or punishing people based on random variation or system-level features out of their control. He argued that variability occurs in any system, and knowledge of the structure of that variation is needed to improve quality. VAMs can provide important knowledge about the different sources of variation. The models can be valuable as part of an ongoing cycle of quality improvement, as they can be used to study the results of innovations on teachers as well as on students. But they need to be viewed from the context of the whole education system, with an understanding of psychology and an appreciation of the effects on other parts of the system.

Hansen and Deming (1950, p. 215) made an important point about the limitations of data and statistical models, stating that "too often the standard error of a result obtained from sampling has been confused with the standard error of a forecast that is based, partially at least, on this result." Data collected for VAMs may provide an accurate picture of the system at the time the data are collected. A forecast, however, depends on assumptions that the conditions in effect at the time the data are collected will continue into the future. While retrospective analyses may suggest methods that might be considered for improving the system, these actions must be tried out experimentally through the PDSA cycle in order to know their effects. Johnson (2015) argued that more experimental and comparative research is needed on long-term consequences of expanding the use of VAMs for teacher evaluation.

Deming advocated recognizing the natural variability among people and changing the system to fit the people. He wanted to replace a system of fear and ranking with a system in which students and teachers can enjoy learning. Accountability, in Deming's world, is replaced by pride in one's work as a teacher. The system recognizes that students learn at different speeds and in different ways, and that teachers have different strengths. It exploits that variability, not to rank people, but to promote learning and continual improvement for all. Deming viewed appropriate and responsible use of statistical methods as key for improving the education system.

References

American Institutes for Research. (2013). *Florida Comprehensive Assessment Test (FCAT) 2.0 value-added model: Technical report 2012-13*. Retrieved June 28, 2014 from http://www.fldoe.org/committees/doc/Value-Added-Model-Technical-Report.docx.

American Statistical Association. (ASA, 2014). *ASA statement on using value-added models for educational assessment*. Retrieved June 14, 2014 from https://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf

Amrein-Beardsley, A., Berliner, D. C., & Rideau, S. (2010). Cheating in the first, second, and third degree: Educators' responses to high-stakes testing. *Educational Policy Analysis Archives*, *18*(14), 1–74. doi:http://dx.doi.org/10.14507/epaa.v18n14.2010

Austenfeld, R. B., Jr. (2001). W. Edwards Deming: The story of a remarkable person. *Papers of the Research Society of Commerce and Economics, 42*(1), 49–102.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., . . . Shepard, L. A. (2010). Problems with the use of student test scores to evaluate teachers. *Economic Policy Institute Briefing Paper*, *278*. Retrieved July 5, 2014 from www.epi.org/publication/bp278.

Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practices*, *28*(4), 42–51. doi:10.1111/j.1745-3992.2009.00161.x

Briggs, D. C., & Domingue, B. (2011). *Due diligence and the evaluation of teachers*. Boulder, CO: National Education Policy Center. Retrieved June 14, 2014 from http://nepc.colorado.edu/publication/due-diligence.

Briggs, D. C., & Domingue, B. (2013). The gains from vertical scaling. *Journal of Educational and Behavioral Statistics, 38*, 551–576. doi:10.3102/1076998613508317

Broatch, J., & Lohr, S. (2012). Multidimensional assessment of value added by teachers to real-world outcomes. *Journal of Educational and Behavioral Statistics*, *37*, 256–277. doi:10.3102/1076998610396900

Campbell, D. T. (1976). *Assessing the impact of planned social change* (Paper #8 of the Occasional Paper Series). Hanover, NH: Public Affairs Center, Dartmouth College.

Chicago Public Schools. (2014). *Information on the value-added metric*. Retrieved August 26, 2014 from http://www.cps.edu/Pages/valueadded.aspx.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E. H., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan, 93*, 8-15. Retrieved August 1, 2015 from http://www.edweek.org/ew/articles/2012/03/01/kappan_hammond.html

Deming, W. E. (1947). Some criteria for judging the quality of surveys. *Journal of Marketing, 12*(2), 145–157. doi:10.2307/1245354

Deming, W. E. (1950). *Some theory of sampling*. New York, NY: Dover.

Deming, W. E. (1965). Principles of professional statistical practice. *Annals of Mathematical Statistics, 36*, 1883–1900. doi:10.1002/0471667196.ess2058.pub2

Deming, W. E. (1986). *Out of the crisis*. Cambridge, MA: MIT Press.

Deming, W. E. (1994). *The new economics for industry, government, education*. Boston, MA: MIT Press.

Deming, W. E., Crawford-Mason, C., & Dobyns, L. (1993). *The Deming video library*. Washington, DC: CC-M Productions.

Deming, W. E., & Geoffrey, L. (1941). On sample inspection in the processing of census returns. *Journal of the American Statistical Association, 36*(215), 351–360. doi:10.1080/01621459.1941.10500570

Duncan, J., & Shelton, W. (1978). *Revolution in United States Government statistics, 1926–1976*. Washington, DC: U.S. Department of Commerce, Office of Federal Statistical Policy and Standards.

Gabor, A. (1990). *The man who discovered quality*. New York, NY: Times Books.

Goldhaber, D., Walch, J., & Gabele, B. (2014). Does the model matter? Exploring the relationship between different student achievement-based teacher assessments. *Statistics and Public Policy, 1*(1), 28–39. doi:10.1080/2330443X.2013.856169

Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of international performance. *Journal of the Royal Statistical Society, Series A, 159*, 385–443. doi:10.2307/2983325

Government Accountability Office. (2013). *K-12 education: States' test security policies and procedures varied.* GAO report GAO-13-495R. Retrieved June 28, 2014 from http://www.gao.gov/assets/660/654721.pdf.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015). Can value-added measures of teacher performance be trusted? *Education Finance and Policy*, 10, 117–156. doi:10.1162/EDFP_a_00153

Haertel, E. H. (2013). *Reliability and validity of inferences about teachers based on student test scores*. Princeton, NJ: Educational Testing Service. Retrieved July 4, 2014 from https://www.ets.org/Media/Research/pdf/PICANG14.pdf.

Hansen, M. H., & Deming, W. E. (1950). On an important limitation to the use of data from samples. *Bulletin de L'Institut International de Statistique, 32,* 214–219.

Johnson, S. M. (2015). Will VAMS reinforce the walls of the egg-crate school? *Educational Researcher, 44,* 117–126. doi:10.3102/0013189X15573351

Karl, A. T., Yang, Y., & Lohr, S. L. (2013). Efficient maximum likelihood estimation of multiple membership linear mixed models, with an application to educational value-added assessments. *Computational Statistics and Data Analysis, 59*, 13–27. doi:10.1016/j.csda.2012.10.004

Kilian, C. (1992). *The world of W. Edwards Deming* (2nd ed.). Knoxville, TN: SPC Press.

Koretz, D. (2008). *Measuring up: What educational testing really tells us.* Cambridge, MA: Harvard University Press.

Latzko, W. J., & Saunders, D. M. (1995): *Four days with Dr. Deming: A strategy for modern methods of management*. Reading, MA: Addison-Wesley.

Lee, J., Sridharan, A., & Sung, A. (2014, May 7). HISD gets a failing grade on its teacher evaluations. *Houston Chronicle*. Retrieved May 10, 2014 from http://www.houstonchronicle.com/opinion/outlook/article/Lee-Sridharan-Sung-HISD-gets-a-failing-grade-5460933.php.

Lohr, S. L. (2012). The value Deming's ideas can add to educational evaluation. *Statistics and Public Policy, 3*(2), 1–40. doi:http://dx.doi.org/10.1515/2151-7509.1057

Mann, N. R. (1994). W. Edwards Deming 1990–1993. *Journal of the American Statistical Association*, *89*, 365–366. doi:10.1080/01621459.1994.10476752

Mann, N. (1995). Dr. Deming: Holder of the keys to excellence. In F. Voehl (Ed.), *Deming: The way we knew him* (pp. 49–58). Delray Beach, FL: St. Lucie Press.

Mariano, L. T., McCaffrey, D. F., & Lockwood, J. R. (2010). A model for teacher effects from longitudinal data without assuming vertical scaling. *Journal of Educational and Behavioral Statistics*, *35*(3), 253–279. doi:10.3102/1076998609346967

McCaffrey, D. F., & Lockwood, J. R. (2011). Missing data in value-added modeling of teacher effects. *The Annals of Applied Statistics*, *5*(2A), 773–797. doi:http://dx.doi.org/10.1214/10-AOAS405

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability.* Santa Monica, CA: Rand.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A. & Hamilton, L. S. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*, 67–101. doi:10.3102/10769986029001067

Moore, S. (2010). Deming's famous red bead experiment with a twist: Improve the system, distort the system, or distort the data [blog post]. Retrieved March 8, 2014 from http://www.qualitydigest.com/inside/twitter-ed/deming-s-famous-red-bead-experiment-twist.html#.

Neave, H. (1987). Deming's 14 points for management: Framework for success. *Journal of the Royal Statistical Society, Series D, 36*, 561–570. doi:10.2307/2348667

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives*, *18*(23), 1–27. doi:http://dx.doi.org/10.14507/epaa.v18n23.2010

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237–257. doi:10.3102/01623737026003237

Orsini, J. N. (2013). *The essential Deming: Leadership principles from the father of quality*. New York, NY: McGraw-Hill.

Pink, D. (2011). *Drive: The surprising truth about what motivates us*. New York, NY: Riverhead Books.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee Value-Added Assessment System: A quantitative, outcomes-based approach to educational assessment. In J. Millman, (Ed.), *Grading teachers, grading schools: Is student achievement a valid educational measure?* (pp. 137–162). Thousand Oaks, CA: Corwin Press.

Schochet, P. Z., & Chiang, H. S. (2013). What are error rates for classifying teacher and school performance using value-added models? *Journal of Educational and Behavioral Statistics*, *38*(2), 142–171. doi: 10.3102/1076998611432174

Schwinn, C., & Schwinn, D. (1995). W. Edwards Deming: Personal remembrances of leadership. In F. Voehl (Ed.), *Deming: The way we knew him* (pp. 17-22). Delray Beach, FL: St. Lucie Press.

Thomas, E., Wingert, P., Conant, E., & Register, S. (2010). Why we can't get rid of failing teachers. *Newsweek*, *155*(11), 24–27.

U.S. Department of Education (1991). *America 2000: An education strategy*. Washington, DC: Author.

Walton, M. (1986). *The Deming management method*. New York, NY: Dodd, Mead & Co.

Wright, S. P., White, J. T., Sanders, W. L., & Rivers, J. C. (2010). *SAS® EVAAS® Statistical Models*. Cary, NC: SAS Institute, Inc. Retrieved June 27, 2014 from http://www.sas.com/resources/asset/SAS-EVAAS-Statistical-Models.pdf.

About the Author

Sharon Lohr
Westat
sharonlohr@westat.com
Sharon Lohr is Vice President at Westat and Professor Emerita of Statistics at Arizona State University. She received her Ph.D. in Statistics from the University of Wisconsin-Madison. She is the author of the book *Sampling: Design and Analysis* and has published numerous articles on survey sampling, hierarchical models, missing data, design of experiments, and applications of statistics in the social sciences and education. She is a Fellow of the American Statistical Association, an elected member of the International Statistical Institute, and a recipient of the Gertrude M. Cox Statistics Award and the Morris Hansen Lectureship Award. This paper is based on Dr. Lohr's 2014 Deming Lecture, which was presented at the Joint Statistical Meetings in Boston.

education policy analysis archives

Volume 23 Number 80 August 24th, 2015 ISSN 1068-2341