



The Stability of Teacher Performance and Effectiveness: Implications for Policies Concerning Teacher Evaluation

Grant B. Morgan

Kari J. Hodge

Baylor University

Tonya M. Trepinski

Texas A & M International University



Lorin W. Anderson

University of South Carolina

United States

Citation: Morgan, G. B., Hodge, K. J., Trepinski, T. M., & Anderson, L. W. (2014). The stability of teacher performance and effectiveness: Implications for policies concerning teacher evaluation. *Education Policy Analysis Archives*, 22(95). <http://dx.doi.org/10.14507/epaa.v22n95.2014>

Abstract: The last five to ten years has seen a renewed interest in the stability of teacher behavior and effectiveness. Data on teacher performance and teacher effectiveness are being used increasingly as the basis for decisions about continued employment, tenure and promotion, and financial bonuses. The purpose of this study is to explore the stability of both teacher performance and effectiveness by determining the extent to which performances and effectiveness of individual teachers fluctuate over time. The sample consisted of 132 teachers for whom both observational and state standardized test data were available for five consecutive years. Neither teacher performance nor effectiveness were highly stable over multiple years of

the study. The observed relationship between teacher performance and teacher effectiveness was reasonably stable over time, but the magnitude of the relationship was quite small. Teacher performance was also likely to be inflated in low performing schools. We also discuss when different observed patterns may be acceptable based on the purpose for which the data are used. **Keywords:** teacher evaluation policy; effectiveness; teacher performance; teacher stability; value-added

La estabilidad del desempeño y eficacia docente: Implicaciones para las políticas de evaluación de maestros

Resumen: En los últimos cinco a diez años ha visto un renovado interés en la estabilidad de la conducta y la efectividad de los maestros. Los datos sobre el desempeño de los docentes y la eficacia docente se están utilizando cada vez más como la base para las decisiones sobre el mantenimiento del empleo, la tenencia y la promoción, y los bonos financieros. El propósito de este estudio es explorar la estabilidad de rendimiento y eficacia de los docentes analizando en que medida las actuaciones y la eficacia de los profesores individuales fluctúan con el tiempo. La muestra estuvo constituida estaban disponibles datos tanto observacionales y de pruebas estandarizadas durante cinco años consecutivos. Ni el desempeño docente ni la eficacia era muy estable a lo largo de varios años de estudio. La relación observada entre el desempeño de los docentes y la efectividad del maestro era razonablemente estable en el tiempo, pero la magnitud de la relación era bastante pequeña. Es probable que el desempeño de los docentes también se halla incrementado en escuelas de bajo rendimiento. También discutimos en que medida los diferentes patrones observados pueden ser aceptables como una base confiable para los fines propuestos por esas políticas.

Palabras clave: política de evaluación de los docentes; eficacia; desempeño de los docentes; estabilidad del profesorado; valor añadido

Estabilidade do desempenho e eficácia do ensino: Implicações para as políticas de avaliação de professores

Resumo: Nos últimos cinco a dez anos tem visto um renovado interesse na estabilidade do comportamento e da eficácia dos professores. Os dados sobre o desempenho e eficácia dos professores são cada vez mais usados como base para a toma de decisões sobre emprego retenção, posse e promoção, e incentivos. O objetivo deste estudo é explorar a estabilidade do desempenho e eficácia de professores analisando em que medida as ações e efetividade dos professores muda ao longo do tempo. A amostra foi composta por dados de 132 professores de exames e observações obtidos por cinco anos consecutivos. Nem o desempenho e a eficácia dos professores foram muito estáveis ao longo dos vários anos de estudo. A relação observada entre o desempenho do professor e eficácia do professor era razoavelmente estável ao longo do tempo, mas a magnitude da relação foi muito pequena. Desempenho dos professores também parece aumentar em escolas de baixo desempenho. Discutimos, também, em que medida os diferentes padrões observados podem ser aceitáveis para os objetivos das políticas.

Palavras-chave: política de avaliação de professores; eficiência; desempenho dos professores; estabilidade dos professores; valor

Introduction

Over the last decade, we have seen renewed desire on the part of an increasing number of educators, parents, and policy makers to ensure that every child is taught by a highly-qualified, competent teacher. This may be attributed in part to empirical evidence that teachers are the most

important school-based determinant of student achievement (Rivkin, Hanushek, & Kain, 2005; Hattie, 2009) as well as philanthropic efforts, such as those by the Bill and Melinda Gates Foundation, to improve teacher quality. Furthermore, state and federal laws (e.g., Race to the Top, No Child Left Behind) have created high-stakes situations for teachers by linking their classroom performance and effect on students to decisions about their employment, promotion, and compensation (Welsh, 2011). Not surprisingly, increased attention has been given to the most valid and reliable ways of measuring teacher performance (i.e., what they do) and effectiveness (i.e., what impact they have on their students) (Medley, 1982). At the same time, however, there is remarkably little research to guide such critical decisions as which teachers to hire, retain, remunerate, and promote (Rice, 2003).

Examining the validity of the effectiveness data is quite complex and requires multiple examinations and inferences. Classroom observations are the primary sources of data on teacher performance although document analysis (e.g., lesson plans, student assignments) and teacher surveys may also be used. Student test scores and, occasionally, student surveys are used to collect data on teacher effectiveness. Validity evidence must be collected and shown for teacher performance and effectiveness data.

The validity of the performance measures stems primarily from the connection between the items included on the instruments and results derived from research on teaching (Grossman et al., 2010; Hill et al., 2008) and/or frameworks developed by professional associations and organizations, such as the Interstate Teacher Assessment and Support Consortium (InTASC) (Danielson, 2007) and the California Commission on Teacher Credentialing (2009). The individual items are organized around standards or components which, in turn, are combined to form domains. Danielson's (2007) Framework for Teaching, for example, exemplifies this structure, with 76 elements, 22 components, and four domains. The four most common domains are Planning, Classroom Environment, Instruction, and Professional Responsibilities, although each of these domains can be, and has been, subdivided. Instruction, for example, has been broken down into Teaching Strategies and Assessment/Evaluation. Similarly, Professional Responsibilities has been parsed into Collegiality/ Professionalism and Communication/ Community Relations.

Occasionally, the construct validity evidence for the instruments is examined using factor analysis (see, for example, Cambridge Education, 2013; Fish & Dane, 2000; Piburn et al., 2000). In other cases, the criterion-related validity of instruments is explored using correlational and/or regression analysis. This can be done using the total score or overall rating on the performance measures as the independent variable and a measure of student achievement as the dependent variable. Increasingly, the student achievement measure is based on change in test scores over time, resulting in what are referred to as "value-added" scores (Polikoff, 2013). Regression analysis using value-added scores has shown that the variance in value-added scores that can be attributed to teacher performance rarely exceeds 10 percent (Daley & Kim, 2010; Mihaly, McCaffrey, Staiger, & Lockwood, 2013). Explanations for this relatively low degree of relationship include measurement error (Goldhaber & Hansen, 2010), restricted range of teacher performance scores (Ho & Kane, 2013), and context effects such as subject matter, grade level, and class size (Angrist & Lavy, 1999; Polikoff, 2013).

The first step in arguing for the validity of value-added ratings is to review the alignment of items on the achievement tests with the curriculum standards (be they state or federal) (Anderson & Krathwohl, 2001; McDonnell, 1995). The second step is to check the accuracy with which students are matched across multiple years, are matched with appropriate teachers, and have been enrolled in the teacher's classes for some minimum length of time (American Statistical Association, 2014; Bill & Melinda Gates Foundation, 2010; Reardon & Raudenbush, 2009). The third step is to determine

that the minimum number of students needed to ensure credible value-added data has been reached, examine the class composition and decide whether adjustments need to be made based on differences in class composition, and conduct other similar investigations (Amrein-Beardsley, 2008; Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012; Reardon & Raudenbush, 2009).

Once sufficient validity of the data and the inferences made based on the data have been established, concerns generally shift to the reliability of the data. The reliability of the document analysis and, particularly, the observational data is typically limited to the extent of agreement between trained observers (Daley & Kim, 2010). However, in a few studies reliability has been examined using statistical methods associated with generalizability theory (Goldhaber & Hansen, 2010; Hill, Charalambous, & Kraft, 2012; Shavelson & Dempsey-Atwood, 1976). The results of these studies suggest that teacher performance is inconsistent across grade levels and when classrooms have different student compositions (e.g., racial composition, class size). Furthermore, some components of teacher performance are more stable than others (Polikoff, 2013).

Why and When is Stability Important?

The last five to ten years has seen a renewed interest in the stability of teacher behavior and effectiveness (Darling-Hammond et al., 2012; Goldhaber & Hansen, 2010; Papay, 2011; Polikoff, 2013; Welsh, 2011). Data on teacher performance and teacher effectiveness are being used increasingly as the basis for decisions about continued employment, tenure and promotion, and financial bonuses. The awarding of financial bonuses can arguably tolerate a reasonable degree of instability because they are awarded primarily on evidence of effective teaching at one point in time; however, sound decisions about continued employment, tenure, and promotion are predicated on some degree of stability over time. It is imprudent to make such decisions of the performance and effectiveness of a teacher as “Excellent” one year and “Mediocre” the next.

In light of these considerations, the desirability of stability is largely a function of the purpose for which the data are to be used. Rogosa, Floden, and Willett (1984) identified four possible patterns that may be observed that remain relevant to current policy decisions: absolute invariance, trend, and scatter (including trend plus scatter). Invariance indicates that the same rating was observed for a teacher across time. Trend indicates that the observed ratings for a teacher increased for each time point. Scatter indicates random fluctuation, which may or may not be accompanied by an overall trend.

Different “patterns” are acceptable for different purposes. For employment/dismissal and promotion decisions, stability is very important (i.e., consistently poor and consistently high ratings, respectively). For evaluating professional development programs, one may only be interested in more “positive trends” (i.e., improvement over time). Conversely, professional development programs designed to provide remediation of some sort may benefit from identifying people with “negative trends.” For compensation, desirability of stability further depends on what pay increases are linked to. If the pay is definitely linked to one particular year, then consistency of ratings over time is not considered. If the pay is intended to retain excellent teachers, then stability is important (i.e., consistently high ratings). If the pay is intended to reward improvement over time, then positive trends are most appropriate. Regardless of the purpose, looking at score stability of individual teachers across time is the only way to obtain the information necessary to make these determinations.

Research Questions

In light of the previous discussion the primary purpose of this study is to explore the stability of both teacher performance (behavior) and effectiveness by determining the extent to

which performances and effectiveness of individual teachers are stable over time. Specifically, two questions guided the analysis of the data collected and presented here. They are:

- 1) How stable are teacher ratings based on expert observations and value-added student achievement over four years?
- 2) How stable are the differences between observational and value-added ratings across schools and, if instability exists, is it related to the overall school value-added rating?

Method

Population and Sample

The population consisted of 160 upper elementary (grades 4 and 5) and 128 middle grade (grades 6 through 8) teachers in 23 South Carolina schools that participated in a five-year program funded by the federal Teacher Incentive Fund (TIF). The poverty indices for the schools ranged from 71 to 99, with a median of 90. A poverty index of 90 means that 90% of the students in the school are eligible for Medicaid services, qualify for free or reduced-price lunches, or both.

The four primary goals of the program were to:

- Improve student achievement by increasing teacher and principal effectiveness;
- Reform teacher and principal compensation systems so that teachers and principals are rewarded for increases in student achievement;
- Increase the number of effective teachers teaching poor, minority, and disadvantaged students in hard-to-staff subjects; and
- Create sustainable performance-based compensation systems.

To achieve these goals, the program emphasized ongoing professional development opportunities and activities led by master and mentor teachers with content determined primary from classroom observations and students' test scores. In terms of compensation, teachers received bonuses based in part on (1) their observed teaching performance, (2) improved test scores of their students, and (3) overall school effectiveness.

Teachers were observed four times per year with the results of each observation summarized on a structured, standardized rubric (see discussion below). Prior research has suggested that from three to four observations per year yields relatively stable estimates of teacher performance (Hill, Charalambous, & Kraft, 2012; Smolkowski & Gunn, 2012). Each year students in grades 3 through 8 were administered the Palmetto Assessment of State Standards (PASS) in mid-March (writing) and early May (reading and mathematics). Students in grades 4 and 7 also were administered science and social studies tests in early-May. Students in grades 3, 5, 6, and 8 were administered either science or social studies tests in alternate years, also in early May. Students in the majority of schools were administered the Measures of Academic Progress (MAP) periodically each year. MAP results were used primarily for professional development, whereas PASS scores were used (along with observation ratings) to determine teacher bonuses.

The final analytic sample consisted of 132 teachers for whom both observational and PASS data were available for five consecutive years. Of this sample, 79 taught students in grades 4 and 5 and 53 taught students in grades 6, 7, or 8. Sixty-one teachers were lost over the five year period because they (1) changed grade levels, (2) changed schools and/or school districts, or (3) retired. Because Year 1 was primarily a planning year, the Year 1 data were considered baseline data. Therefore, four years of data are included in this study (Years 2 through 5).

Variables

There were three primary variables: observation ratings, teacher value-added ratings, and overall school ratings. Each is discussed in this section.

Observational ratings. The rubric used to record the observations was developed by researchers at the National Institute for Excellence in Teaching, who operate the TAP program. The instrument contains 19 indicators that “provide sufficient breadth to ensure that evaluation ratings reflect the kind of effective instructional practices that predict positive learning outcomes” (Jerald & Van Hook, 2011, p. 4). Each of the 19 indicators measured one of three domains: Designing and Planning Instruction, Learning Environment, and Instruction. Designing and Planning Instruction, for example, was associated with three indicators: Instructional Plans, Student Work, and Assessment. Learning Environment was associated with four indicators and Instruction was associated with 12. Figure 1 contains an example of the rating scale for one indicator associated with the domain, Learning Environment.


	Exemplary (5)	Proficient (3)	Emerging (1)
Respectful Culture 	<ul style="list-style-type: none"> • Teacher-student interactions demonstrate caring and respect for one another. • Students exhibit caring and respect for one another. • Teacher seeks out, and is receptive to the interests and opinions of all students. • Positive relationships and interdependence characterize the classroom. 	<ul style="list-style-type: none"> • Teacher-student interactions are generally friendly, but may reflect occasional inconsistencies, favoritism, or disregard for students' cultures. • Students exhibit respect for the teacher, and are generally polite to each other. • Teacher is sometimes receptive to the interests and opinions of students. 	<ul style="list-style-type: none"> • Teacher-student interactions are sometimes authoritarian, negative, or inappropriate. • Students exhibit disrespect for the teacher. • Student interaction is characterized by conflict, sarcasm, or put-downs. • Teacher is not receptive to interests and opinions of students.

Figure 1. Example of the rating scale for one indicator associated with Learning Environment domain. Adapted from National Institute for Excellence in Teaching (2013, p. 64).

Each teacher was observed four times each year by a school administrator, master teacher, or mentor teacher. For each of the 19 indicators, the observer assigned a rating from 1 to 5 based on what was observed. Narrative descriptions were available for three of the rating categories: Exemplary (5), Proficient (3), and Emerging (1). Although narrative descriptions were not available for the other two rating categories (2 and 4), observers were trained to assign these intermediate ratings when what they observed did not fit clearly in any of the categories with narrative descriptions.

For each teacher, the results of each observation can be summarized as a vector with 19 digits, with each digit representing a particular indicator and ranging from 1 to 5. Over four observations, the results can be summarized as a 19 × 4 matrix. In part because there are a greater number of indicators associated with the Instruction domain, ratings on that domain received a

greater weight (75%) in a computational formula. Using the computational formula, a single rating from 1 to 5 was assigned to each teacher for each observation. The ratings for the four observations were then averaged and the result rounded to the nearest half point. For any given year, nine possible observational ratings could be assigned: 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, and 5.0.

Teacher value-added ratings. The Palmetto Assessment of State Standards (PASS) is a battery of tests that are administered to students in grades 3 through 8. The tests have been carefully aligned with the state academic standards. PASS tests in English language arts and mathematics are administered to students enrolled in all six-grade levels every year. PASS tests in science and social studies are administered to students enrolled in grades 4 and 7 every year. For students in grades 3, 5, 6, and 8 PASS tests in science or social studies are administered in alternate years.

To support the TIF program, schools contracted with SAS data services in Cary, North Carolina to compute value-added ratings based on two-year, longitudinally matched sets of PASS scores. SAS uses the Education Value-Added Accountability System (EVAAS) to compute the ratings. Using mixed model equations, EVAAS uses the covariance matrix from the longitudinal data sets to estimate student progress *vis-à-vis* state-level normative data. In the model, each student acts as his or her own control and no other covariates are used. Because two years of data are needed, third grade teachers are not included in the sample since PASS tests are first administered in grade 3. Once the EVAAS scores have been computed, they are converted to a five-point scale, with 5 being the most positive. Unlike the observation ratings, half points are not assigned.

School effectiveness ratings. School effectiveness ratings are also based on EVAAS scores. The scores used to determine the school effectiveness ratings are aggregated directly to the school level. Like the teacher value-added ratings, the school effectiveness ratings are on a five-point scale, with 5 being the most positive.

Relationships between teacher ratings. The estimated polychoric correlation coefficients for years 2 through 5 of the study were respectively .19, .22, .29, and .42. Although the correlation coefficients increase across time, the highest correlation observed was .42. In other words, the variability explained between value-added and observational ratings was at best about 18% in this study.

We also tested the equality of the four correlation coefficients by computing simultaneous 95% confidence intervals using 1,000 bootstrapped samples. To control the Type I error rate at 5%, we used Bonferroni adjustment such that each confidence interval was based on the .00625 and .99375 quantiles of the bootstrapped sampling distribution. The confidence interval for the year 2 was (-.05, .42), year 3 was (.00, .44), year 4 was (.03, .50), and year 5 was (.22, .60). Due to the overlap between the intervals, we could not conclude that the correlation between value-added and observational ratings at the teacher-level in any given year was statistically different from any other year.

Data Analysis

In this study, we conceptualized stability as the change in scores and/or observational ratings across time, and there are multiple aspects of change. For example, change can be shown by the number of different scores across time and/or by the magnitude of the differences in scores. For each research question, we conducted multiple types of analyses in an effort to capture different aspects of score stability and ultimately provide a more robust examination of the fluctuation of scores and ratings across time. The analysis was primarily descriptive in nature and did not require distributional assumptions about the errors. Furthermore, all 132 records were complete.

To answer the first research question, we first generated frequency distributions of the value-added and observational rating range for each teacher. Second, we generated frequency distributions

for the number of adjacent years for which teachers had discrepant value-added and observational ratings. Third, we constructed a panel plot for a random subsample of nine teachers in order to examine the variation in individual value-added and observational rating patterns. Last, we categorized each of the teachers into one of the four longitudinal patterns discussed above (i.e., invariant, trend, scatter, trend plus scatter).

For the second research question, we computed the mean school-level value-added rating across the years of the study and then classified the schools on the basis on the computed means. For teachers within each school, we subtracted the value-added rating from the observational rating in each year of the study. Next, we estimated the mean difference for all teachers within each school. Finally we examined the distribution of mean difference scores with each school-level value-added category.

Results

For each research question, we conducted multiple types of analysis in an effort to capture different indicators of score stability. The results from this study are organized by research question.

Research Question 1

First, we computed the range of scores for each teacher across the four years of the study. For example, a teacher with a score pattern of 1-1-1-5 has a range of four, and a teacher with a score pattern of 3-4-3-4 has a range of one. Among the value-added ratings, eight teachers (6.1%) had the same score in all four years, 38 teachers (28.8%) had a range of one, 68 teachers (51.5%) had a range of two, 16 teachers (12.1%) had a range of three, and two (1.5%) teachers had a range of four. Nearly two-thirds of the teachers had value-added ratings that differed two or more points over the course of the study. This suggests that the value-added rating may not be stable across time.

For the observational rating distribution, 14 teachers (10.6%) had a range of zero, 63 teachers (47.7%) had a range of 0.5, 43 teachers (32.6%) had a range of 1.0, 11 (8.3%) teachers had a range of 1.5, and one teacher (0.8%) had a range of 2.0. It should be noted that it would be very difficult to observe observational ratings at the extremes (i.e., near 1 or 5) given that the observational rating for each year is based on the mean of the teacher observations across multiple measurements within that academic year. The observational ratings observed in the study ranged from 2.0 ($n = 5$) to 5.0 ($n = 1$). In light of this distribution, that nearly 90% of different observational ratings across the study and over 40% had observational ratings that differed by at least 1.0 point suggests that the observational ratings may not be stable.

Third, we reported the number of year-to-year score discrepancies per teacher. For example, a teacher with a score pattern of 1-1-1-5 would receive a one, and a teacher with a score pattern of 3-4-3-4 would receive a three. A discrepancy value of zero indicates that ratings were the same across the years of the study, and the maximum discrepancy score is three, which indicates that a teacher's score was never the same in adjacent years. The discrepancy scores for value-added and observational ratings and their associated frequencies are presented in Table 1 below. Nearly 94% of the teachers had a different value-added rating in at least two years, and one out of four teachers had a different value-added rating in *every* year of the study. Similarly, almost 90% of the teachers had different observational ratings in two or more of the years of the study, and half of the teachers had different observational ratings in three or four of the years. This analysis provides strong evidence for the instability of the value-added and observational ratings of teachers.

Fourth, we generated line graphs of small subsamples to demonstrate the variation in teacher value-added and observational rating patterns. The panel plot of value-added ratings for five teachers is presented in Figure 2, and the panel plot of observational ratings for six teachers is

presented in Figure 3. Ideally one of two patterns would be observed: 1) scores would generally increase as time progressed (i.e., left-to-right in the plot), or 2) scores that were high at in Year 2 and remained high across all years of the study. As indicated in the plots, there is considerable variation the patterns amongst the selected teachers. For both plots, the scale of the vertical axis should be noted. That is, the plot of value-added ratings (Figure 2) shows values from the entire range (i.e., one to five), and the plot of observational ratings (Figure 3) only shows values between 2.5 and 4.5. The restriction in range is discussed in a previous section.

Table 1

Observed Discrepancy Frequency of Value-Added and Observational Ratings Between Adjacent Years

Adjacent Year Discrepancies	Value-Added Ratings		Observational Ratings	
	<i>N</i>	%	<i>N</i>	%
0	8	6.1	14	10.6
1	28	21.2	52	39.4
2	62	47.0	51	38.6
3	32	24.2	15	11.4

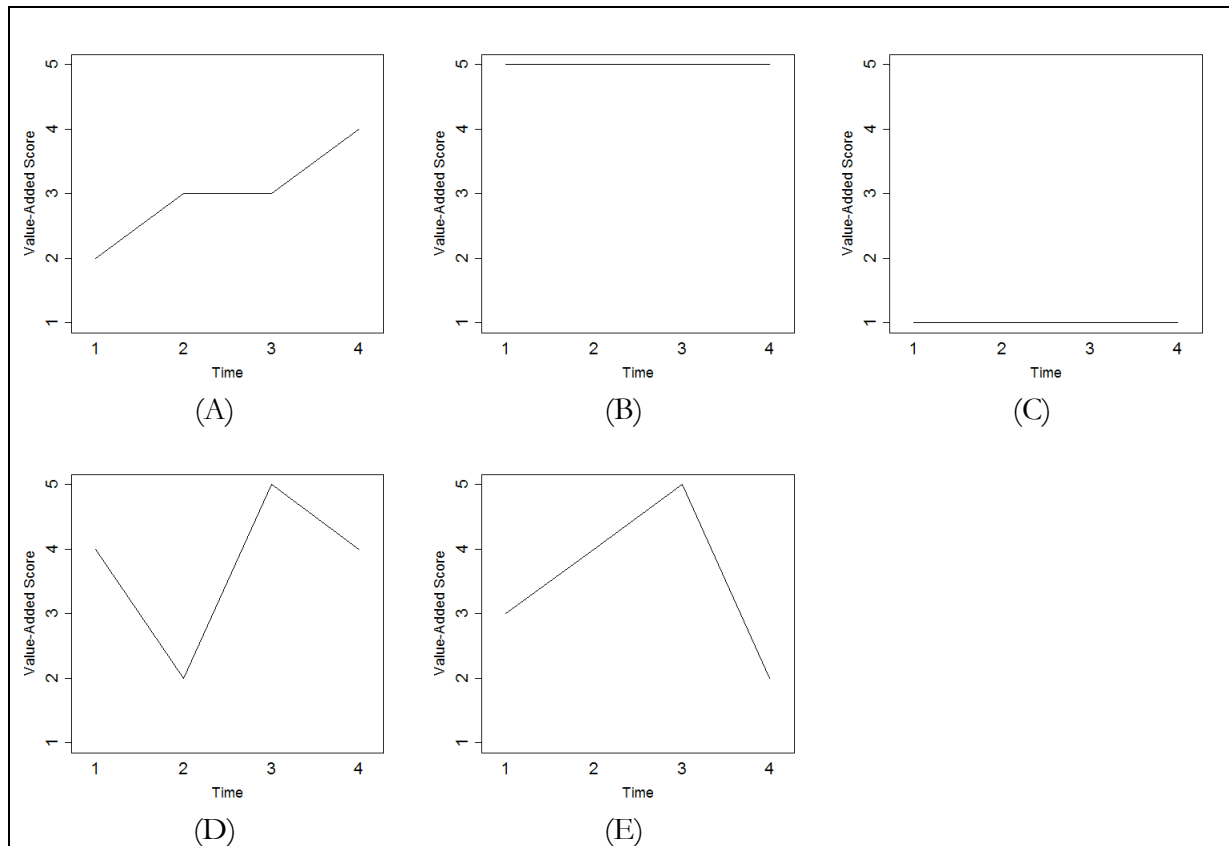


Figure 2. Panel plot of value-added scores from five teachers.

Panel A shows a general upward trend from Times 1 to 4 although no change was observed between Times 2 and 3. Panels B and C shows no change in observational rating score from Times 1 to 4. Panel D shows sizeable changes in value-added scores of adjacent years. Panel E shows an upward trend from Times 1 to 3 but then shows a drastic decline at Time 4. Panel A is an example of a positive trend. Panels B and C are examples of invariant patterns. Panels C and D are examples of scatter.

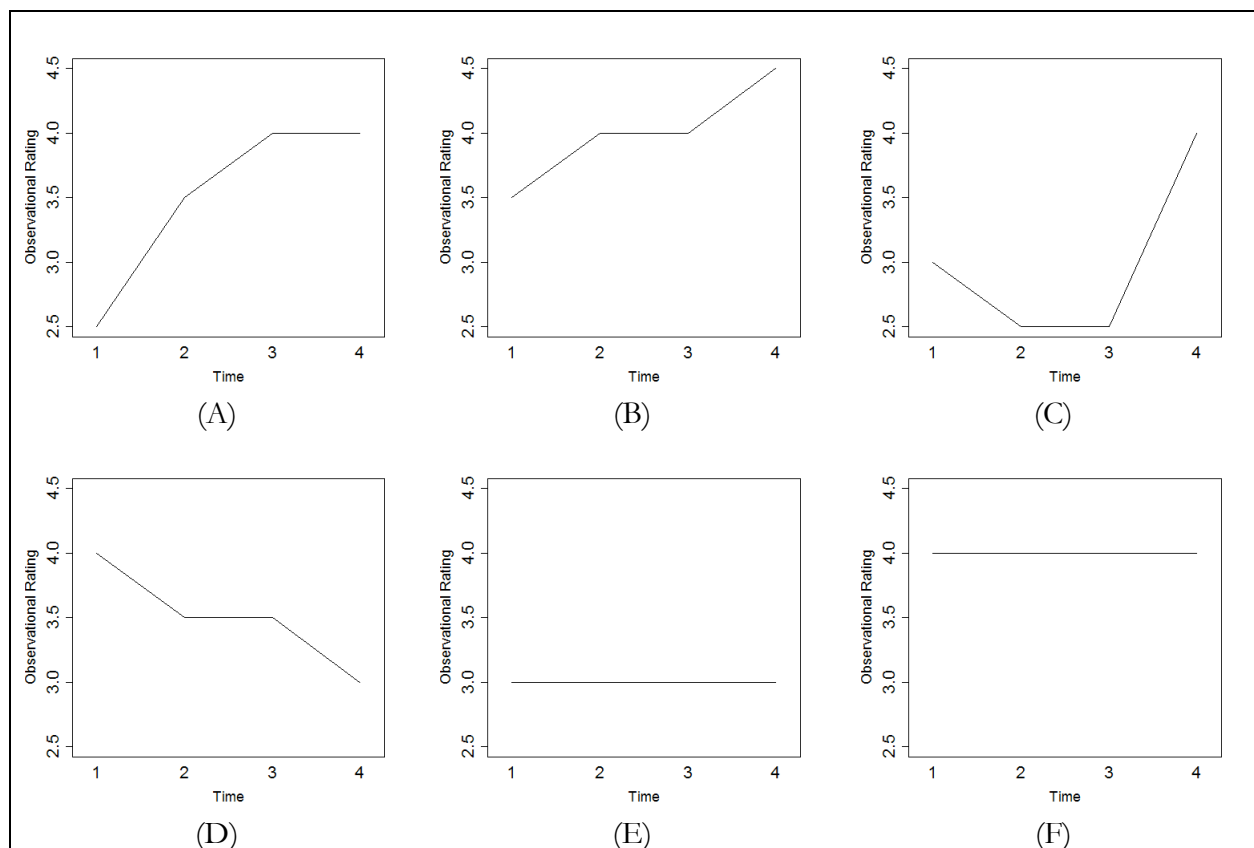


Figure 3. Panel plot of observational ratings from six teachers.

Panel A shows sizeable increases from Time 1 to Time 3 before leveling off in Time 3. Panel B shows a general upward trend from Times 1 to 4 although no change was observed between Times 2 and 3. Panel C shows initial decrease from Time 1 to Time 2, then a no change between Times 2 and 3, and then a sharp increase in Time 4. Panel D shows a general downward trend from Times 1 to 4 although no change was observed between Times 2 and 3. Panels E and F show no change in observational rating score from Times 1 to 4. Panels A and B are examples of positive trend. Panel C is an example of scatter. Panel D is an example of negative trend. Panels E and F are examples of invariant patterns.

Last, we categorized each teacher's change pattern into one of the four patterns described by Rogosa et al. (1984). The frequency with which each of these patterns was observed in the present study is provided in Table 2. Invariant patterns were those that did not fluctuate at all. Positive trend (with and without scatter) patterns were those that showed (a) positive changes between at least two consecutive years and (b) no negative changes. Negative trend (with and without scatter) patterns were those that showed (a) negative changes between at least two consecutive years and (b) no positive changes. Scatter patterns were those that showed (a) change between only one pair of consecutive years or (b) positive and negative changes within the same pattern. Examples of these patterns are shown in Figure 2 and/or 3. Among the value-added ratings, scatter was the most commonly observed pattern (62.9%), and the most commonly observed pattern of observational ratings was positive trend (43.9%).

Table 2
Frequency of Value-Added and Observational Rating Patterns

Rogosa et al. Categories	Value-Added Ratings	Observational Ratings
Invariant	8 (6.1%)	14 (10.6%)
Positive Trend	25 (18.9%)	58 (43.9%)
Negative Trend	16 (12.1%)	19 (14.4%)
Scatter	83 (62.9%)	41 (31.1%)

Note. A “trend” does not have to be “perfect.” For example, a value-added ratings pattern of 2-3-4-5 is clearly a positive trend. A value-added rating pattern of 2-3-3-4 was also considered as a positive trend as well since the overall change across time was positive.

Research Question 2

To explore the potential relationship between the teacher-level value-added and observational ratings based on the school-level value-added rating, we first computed the mean school value-added rating across the four years. Next, we classified each of the 23 schools into one of three categories on the basis of the school-level value-added means. Schools with means between 1.0 and 2.99 were assigned to group 1 ($n = 4$), between 3.0 and 3.99 were assigned to group 2 ($n = 10$), and between 4 and 5 were assigned to group 3 ($n = 9$). To examine possible discrepancy between teacher-level value-added and observational ratings, we computed the difference between the two ratings such that positive difference scores indicated that the observational rating was higher than the value-added ratings, negative difference scores indicated that the observational rating was lower than the value-added ratings, and a zero difference score indicated that the observation and value-added ratings were the same. The mean difference score was computed for each school, which allowed us to generate a scatterplot of the school-level value-added ratings against the computed mean difference scores of teachers within each school. The scatterplot is shown in Figure 5. The mean difference scores for groups 1 through 3 were respectively 0.70, -0.07, and 0.14. Of the schools that were classified into groups 2 or 3, the distribution of differences scores was centered near zero. This suggests that on average the observational ratings were the same as the value-added ratings in schools performing average or above average. On the other hand, schools that were poor performing (i.e., group 1) tended to have higher mean observational ratings than value-added ratings for their teachers. This indicates that on average the observational ratings may overestimate teacher effectiveness in lower performing schools.

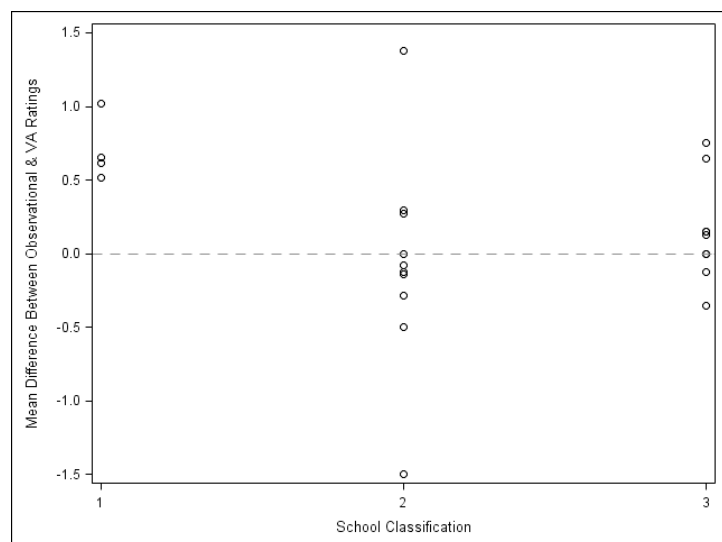


Figure 5. Distribution of mean difference scores by school classification

Discussion

This study focused on the stability of individual teacher performance and effectiveness ratings across time. The importance of why individual teacher stability should be addressed can be exemplified by the work of Ho and Kane (2013) that examined administrator ratings of teachers. In their study, administrators rated the performance of their own teachers higher than administrators from other schools; however, the correlation of administrators' ratings of teachers was 0.87. Thus, even though there was bias in the performance data, the bias did not impact the reliability of the differences between teachers. Thus, the absence of within-teacher examination prevents one from having a complete set of information on which to draw conclusions.

In the current study, neither teacher performance nor teacher effectiveness ratings was highly stable. Teacher performance is somewhat more stable than teacher effectiveness, but this may be the result of the restricted range of the teacher performance ratings (see also Ho & Kane, 2013). Additional discussion is warranted regarding the desirability of stability. On the one hand, economic theory commonly models unobserved worker quality as a given parameter that is relatively fixed over time (Goldhaber & Hansen, 2010). With this theory, stability is a requisite for many of the policy implications offered by economists. For example, the feasibility of Hanushek's (2011) suggestion that replacing the bottom 5–8 percent of teachers with average teachers could move the U.S. near the top of international math and science rankings would first require that teacher quality be a stable characteristic. On the other hand, many educators argue that good teaching requires adaptations and accommodations that require some degree of instability. Berliner (1976), for example, stated that the standard of excellence in teaching commonly held implies a teacher whose behavior is inherently unstable.

In our study, we showed the frequency with which each of the patterns identified by Rogosa et al. (1984) was observed in Table 2. The most commonly observed patterns for value-added ratings and observational ratings were scatter and positive trend, respectively. These data suggest that while teachers in the study tended to improve their performance ratings over time, the improvement in classroom performance did not translate into improved effectiveness ratings, on average. Though not a part of the guiding research questions, this finding is consistent with our analysis of the

relationship between the value-added and observational ratings. We observed a generally weak positive relationship between the two sets of ratings. These findings support the use of multiple measures of teacher competence, teacher performance, and teacher effectiveness in all teacher evaluation systems, which has been widely recommended (American Statistical Association, 2014; Bill & Melinda Gates Foundation, 2010; Steele, Hamilton, & Stecher, 2010)

There are several possible explanations for the inconsistency between teacher performance and teacher effectiveness ratings. As suggested by Goldhaber and Hansen (2010), the first source is measurement error. Correcting correlations for attenuation can increase percent of variation accounted for by 10 to 15 percent (see Polikoff, 2013). Variables have also been proposed that may mediate relationship between teacher behavior and student achievement, such as task perception, self-regulation, motivation, teacher efficacy, and curriculum alignment (Doyle, 1978; Gallagher, 2004; Winne, 1987). Finally, there are contextual differences, such as grade level, subject matter, and class size and composition. It is possible that different observation protocols are necessary for observing teachers of different grade levels. Subject-specific observation forms may also be necessary to accurately reflect the differences in practices employed by teachers of different subjects as has been recommended by Bill & Melinda Gates Foundation (2010).

Finally, our analysis indicated that teacher performance was inflated in low performing schools on average. This suggests that performance ratings may be school-specific. Performance ratings are based on the performance of teachers relative to other teachers in that school, which are independent of where the school stands in terms of their overall effectiveness (e.g., value-added test scores). This must also be taken into account in teacher quality evaluation programs. The use of independent, external observers may mitigate this school effect.

The stability of teacher performance and effectiveness has important implications for the formation of teacher policies. If, as Goldhaber and Hansen (2010) suggest, the assertion that quality is relatively fixed over time is valid, then perhaps the best way to improve our educational system is to weed out poor performers. The validity of this assumption lies at the heart of Hanushek's (2011) assertion mentioned previously. If, however, teacher performance and/or effectiveness tend to be an unstable characteristic, then, it may be necessary to "radically re-think the direction of teacher-based accountability" (Goldhaber & Hansen, 2010, p. 2).

References

- Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher*, 37(2), 65-75. <http://dx.doi.org/10.3102/0013189X08316420>
- Angrist, J. D. & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, 114, 533-575. <http://dx.doi.org/10.1162/003355399556061>
- Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). A taxonomy for learning, teaching and assessing: A revision of Bloom's Taxonomy of educational objectives. New York: Longman.
- Berliner, D. C. (1976). A status report on the study of teacher effectiveness. *Journal of Research in Science Teaching*, 13(4), 369-382. <http://dx.doi.org/10.1002/tea.3660130415>
- Bill & Melinda Gates Foundation. (2010). Empowering effective teachers: Readiness for reform. Seattle, WA: Author. Retrieved May 22, 2013, from <http://docs.gatesfoundation.org/united-states/documents/empowering-effective-teachers-readiness-for-reform.pdf>
- Cambridge Education. (2013). Tripod Survey Assessments. Westwood, MA: Author. Retrieved May 21, 2013, from <http://tripodproject.org/wp-content/uploads/2012/03/Flyer-Tripod-2012.pdf>

- Commission on Teacher Credentialing. (2009). *California standards for the teaching profession*. Sacramento, CA: Author. Retrieved May 22, 2013, from <http://www.ctc.ca.gov/educator-prep/standards/CSTP-2009.pdf>
- Daley, G., & Kim, L. (2010). *A teacher evaluation system that works*. Lecture presented at National Institute for Excellence in Teaching, Santa Monica, CA.
- Danielson, C. (2007). *Enhancing professional practice: A framework for teaching, 2nd edition*. Washington, D. C.: ASCD.
- Darling-Hammond, L. (2000). Teacher quality and student achievement. *Educational Policy Analysis Archives*, 8(1). Retrieved from <http://epaa.asu.edu/ojs/article/view/392>
- Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012, March). Evaluating teacher evaluation: What we know about value-added models and other methods. *Phi Delta Kappan*, 93(6), 8-15. <http://dx.doi.org/10.1177/003172171209300603>
- Doyle, W. (1978). Paradigms for research on teacher effectiveness. In L. S. Shulman (Ed.), *Review of Research in Education*, 5 (pp. 163-98). Itasca, IL: F. E. Peacock.
- Fish, M. C. & Dane, E. (2000). The Classroom Systems Observation Scale: Development of an instrument to assess classrooms using a systems perspective. *Learning Environment Research*, 3(1), 67-92. <http://dx.doi.org/10.1023/A:1009979122896>
- Gallagher, H. A. (2004). Vaughn Elementary's innovative teacher evaluation system: Are teacher evaluation scores related to growth in student achievement? *Peabody Journal of Education*, 79(4), 79-107. http://dx.doi.org/10.1207/s15327930pje7904_5
- Goldhaber, D., & Hansen, M. (2010). *Is it just a bad class?: Assessing the stability of measured teacher performance (CEDR Working Paper 2010-3)*. Seattle, WA: University of Washington.
- Grossman, P., Loeb, S., Cohen, J., Hammerness, K., Wycoff, J., Boyd, D., & Lankford, K. (2010, May). *Measures for measure: The relationship between measures of instructional practice in middle school English Language Arts and teachers' value-added scores*. NBER Working Paper No. 16015.
- Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review* 30, 466-479. <http://dx.doi.org/10.1016/j.econedurev.2010.12.006>
- Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.
- Hill, H. C., Blunk, M., Charalambous, C. Y., Lewis, J., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, 26, 430-511. <http://dx.doi.org/10.1080/07370000802177235>
- Hill, H. C., Charalambous, C. Y., Blazar, D., McGinn, D., Kraft, M. A., Beisiegel, M.,...Lynch, K. (2012). Validating arguments for observational instruments: Attending to multiple sources of variation. *Educational Assessment*, 17, 1-19. <http://dx.doi.org/10.1080/10627197.2012.715019>
- Hill, H. C., Charalambous, C. Y., & Kraft, M. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, 41(2), 56-64. <http://dx.doi.org/10.3102/0013189X12437203>
- Ho, A. D. & Kane, T. J. (2013). *The reliability of classroom observation by school personnel*. Seattle, WA: Bill & Melinda Gates Foundation. Retrieved May 22, 2013, from http://metproject.org/downloads/MET_Reliability_of_Classroom_Observations_Research_Paper.pdf
- McDonnell, L. M. (1995). Opportunity to learn as a research concept and a policy instrument. *Educational Evaluation and Policy Analysis*, 17, 305-322. <http://dx.doi.org/10.3102/01623737017003305>
- Medley, D. M. (1982). *Teacher competency testing and the teacher educators*. Charlottesville, VA: Association

- of Teacher Educators and the Bureau of Educational Research, University of Virginia.
- Mihaly, K., McCaffrey, D. F., Staiger, D., & Lockwood, J. R. (2013). *A composite estimator of effective teaching*. Santa Monica, CA: RAND Corporation.
- National Institute for Excellence in Teaching (2013). *Evaluator handbook: South Carolina teaching standards*. Retrieved June 15, 2014, from https://ed.sc.gov/agency/se/Educator-Evaluation/documents/SC_Teaching_Standards_Manual.pdf
- Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal*, 48(1), 163-193. <http://dx.doi.org/10.3102/0002831210362589>
- Piburn, M., Sawada, D., Turley, J., Falconer, K., Benford, R., Bloom, I., & Judson, E. (2000). *Reformed Teaching Observation Protocol (RTOP): Reference Manual*. (ACEPT Technical Report No. IN00-3). Tempe, AZ: Arizona Collaborative for Excellence in the Preparation of Teachers.
- Polikoff, M. S. (2013). *The stability of observational and student survey measures of teaching effectiveness*. Paper presented at the 2013 Annual Conference of the Association for Education Finance and Policy, New Orleans, LA.
- Race to the Top Fund. (2013, February 1). *Race to the Top Fund*. Retrieved May 6, 2013, from <http://www2.ed.gov/programs/racetothetop/index.html>
- Reardon, S. F. & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy*, 4, 492-529. <http://dx.doi.org/10.1162/edfp.2009.4.4.492>
- Rice, J. K. (2003). *Teacher quality: Understanding the effectiveness of teacher attributes*. Washington, DC: Economic Policy Institute.
- Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73, 417-458. <http://dx.doi.org/10.1111/j.1468-0262.2005.00584.x>
- Rogosa, D., Floden, R., & Willett, J. B. (1984). Assessing the stability of teacher behavior. *Journal of Educational Psychology*, 76(6), 1000-1027. <http://dx.doi.org/10.1037/0022-0663.76.6.1000>
- Sanders, W. L., Wright, S. P., Rivers, J. C., & Leandro, J. G. (2009). *A response to criticisms of SAS EVAAS*. Cary, NC: SAS.
- Shavelson, R., & Dempsey-Atwood, N. (1976). Generalizability of measures of teaching behavior. *Review of Educational Research*, 46(4), 553. <http://dx.doi.org/10.3102/00346543046004553>
- Smolkowski, K. & Gunn, B. (2012). Reliability and validity of the Classroom Observation of Student-Teacher Interactions (COSTI) for kindergarten reading instruction. *Early Childhood Research Quarterly*, 27(2), 316-328. <http://dx.doi.org/10.1016/j.ecresq.2011.09.004>
- Steele, J. L., Hamilton, L. S., & Stecher, B. M. (2010). *Incorporating student performance measures into teacher evaluation systems*. Santa Monica, CA: RAND Corporation.
- Welsh, M. E. (2011). Measuring teacher effectiveness in gifted education: Some challenges and suggestions. *Journal of Advanced Academics*, 22(5), 750-770. <http://dx.doi.org/10.1177/1932202X11424882>
- Winne, P. H. (1987). Why process-product research cannot explain process-product findings and a proposed remedy: The cognitive mediational paradigm. *Teaching and teacher education*, 3(4), 333-356. [http://dx.doi.org/10.1016/0742-051X\(87\)90025-4](http://dx.doi.org/10.1016/0742-051X(87)90025-4)

About the Authors

Grant B. Morgan

Baylor University

grant_morgan@baylor.edu

Dr. Grant Morgan is an assistant professor in the Department of Educational Psychology at Baylor University. His research interests include psychometrics, advanced quantitative methods, and classification/clustering using Monte Carlo methods.

Kari J. Hodge

Baylor University

kari_hodge@baylor.edu

Kari Hodge is a doctoral candidate in the Department of Educational Psychology at Baylor University. Her primary areas of research include technology in teacher education as well as psychometric investigations of assessments.

Tonya M. Trepinski

Texas A&M International University

tonya.trepinski@tamiu.edu

Dr. Tonya Trepinski completed her Ph.D. in the Department of Educational Psychology at Baylor University and is now an assistant professor at Texas A&M International University in the College of Education. Her primary areas of research include applied behavior analysis in teacher education programs and autism spectrum disorders.

Lorin W. Anderson

University of South Carolina

andregroup@sc.rr.com

Dr. Lorin Anderson is a Carolina Distinguished Professor Emeritus at the University of South Carolina, where he served on the faculty from August, 1973, until his retirement in August, 2006. His recent research interests are focused on the quality of education provided for children of poverty throughout the world. His most recognized work is *A Taxonomy of Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*, which was published in 2001.

education policy analysis archives

Volume 22 Number 95 October 6th, 2014 ISSN 1068-2341



Readers are free to copy, display, and distribute this article, as long as the work is attributed to the author(s) and **Education Policy Analysis Archives**, it is distributed for non-commercial purposes only, and no alteration or transformation is made in the work. More details of this Creative Commons license are available at

<http://creativecommons.org/licenses/by-nc-sa/3.0/>. All other uses must be approved by the author(s) or **EPAA**. **EPAA** is published by the Mary Lou Fulton Institute and Graduate School of Education at Arizona State University. Articles are indexed in CIRC (Clasificación Integrada de Revistas Científicas, Spain), DIALNET (Spain), [Directory of Open Access Journals](#), EBSCO Education Research Complete, ERIC, Education Full Text (H.W. Wilson), QUALIS A2 (Brazil), SCImago Journal Rank; SCOPUS, Socolar (China).

Please contribute commentaries at <http://epaa.info/wordpress/> and send errata notes to Gustavo E. Fischman fischman@asu.edu

Join EPAA's Facebook community at <https://www.facebook.com/EPAAAPE> and **Twitter feed** @epaa_aape.

education policy analysis archives
editorial board

Editor **Gustavo E. Fischman** (Arizona State University)

Associate Editors: **Audrey Amrein-Beardsley** (Arizona State University), **Rick Mintrop**, (University of California, Berkeley)
Jeanne M. Powers (Arizona State University)

Jessica Allen University of Colorado, Boulder

Gary Anderson New York University

Michael W. Apple University of Wisconsin, Madison

Angela Arzubiaga Arizona State University

David C. Berliner Arizona State University

Robert Bickel Marshall University

Henry Braun Boston College

Eric Camburn University of Wisconsin, Madison

Wendy C. Chi* University of Colorado, Boulder

Casey Cobb University of Connecticut

Arnold Danzig Arizona State University

Antonia Darder University of Illinois, Urbana-Champaign

Linda Darling-Hammond Stanford University

Chad d'Entremont Strategies for Children

John Diamond Harvard University

Tara Donahue Learning Point Associates

Sherman Dorn University of South Florida

Christopher Joseph Frey Bowling Green State University

Melissa Lynn Freeman* Adams State College

Amy Garrett Dikkers University of Minnesota

Gene V Glass Arizona State University

Ronald Glass University of California, Santa Cruz

Harvey Goldstein Bristol University

Jacob P. K. Gross Indiana University

Eric M. Haas WestEd

Kimberly Joy Howard* University of Southern California

Aimee Howley Ohio University

Craig Howley Ohio University

Steve Klees University of Maryland

Jaekyung Lee SUNY Buffalo

Christopher Lubienski University of Illinois, Urbana-Champaign

Sarah Lubienski University of Illinois, Urbana-Champaign

Samuel R. Lucas University of California, Berkeley

Maria Martinez-Coslo University of Texas, Arlington

William Mathis University of Colorado, Boulder

Tristan McCowan Institute of Education, London

Heinrich Mintrop University of California, Berkeley

Michele S. Moses University of Colorado, Boulder

Julianne Moss University of Melbourne

Sharon Nichols University of Texas, San Antonio

Noga O'Connor University of Iowa

João Paraskveva University of Massachusetts, Dartmouth

Laurence Parker University of Illinois, Urbana-Champaign

Susan L. Robertson Bristol University

John Rogers University of California, Los Angeles

A. G. Rud Purdue University

Felicia C. Sanders The Pennsylvania State University

Janelle Scott University of California, Berkeley

Kimberly Scott Arizona State University

Dorothy Shipps Baruch College/CUNY

Maria Teresa Tatto Michigan State University

Larisa Warhol University of Connecticut

Cally Waite Social Science Research Council

John Weathers University of Colorado, Colorado Springs

Kevin Welner University of Colorado, Boulder

Ed Wiley University of Colorado, Boulder

Terrence G. Wiley Arizona State University

John Willinsky Stanford University

Kyo Yamashiro University of California, Los Angeles

* Members of the New Scholars Board

archivos analíticos de políticas educativas
consejo editorial

Editor: **Gustavo E. Fischman** (Arizona State University)

Editores. Asociados **Alejandro Canales** (UNAM) y **Jesús Romero Morante** (Universidad de Cantabria)

Armando Alcántara Santuario Instituto de Investigaciones sobre la Universidad y la Educación, UNAM México

Claudio Almonacid Universidad Metropolitana de Ciencias de la Educación, Chile

Pilar Arnaiz Sánchez Universidad de Murcia, España

Xavier Besalú Costa Universitat de Girona, España

Jose Joaquin Brunner Universidad Diego Portales, Chile

Damián Canales Sánchez Instituto Nacional para la Evaluación de la Educación, México

María Caridad García Universidad Católica del Norte, Chile

Raimundo Cuesta Fernández IES Fray Luis de León, España

Marco Antonio Delgado Fuentes Universidad Iberoamericana, México

Inés Dussel FLACSO, Argentina

Rafael Feito Alonso Universidad Complutense de Madrid, España

Pedro Flores Crespo Universidad Iberoamericana, México

Verónica García Martínez Universidad Juárez Autónoma de Tabasco, México

Francisco F. García Pérez Universidad de Sevilla, España

Edna Luna Serrano Universidad Autónoma de Baja California, México

Alma Maldonado Departamento de Investigaciones Educativas, Centro de Investigación y de Estudios Avanzados, México

Alejandro Márquez Jiménez Instituto de Investigaciones sobre la Universidad y la Educación, UNAM México

José Felipe Martínez Fernández University of California Los Angeles, USA

Fanni Muñoz Pontificia Universidad Católica de Perú

Imanol Ordorika Instituto de Investigaciones Economicas – UNAM, México

Maria Cristina Parra Sandoval Universidad de Zulia, Venezuela

Miguel A. Pereyra Universidad de Granada, España

Monica Pini Universidad Nacional de San Martín, Argentina

Paula Razquin UNESCO, Francia

Ignacio Rivas Flores Universidad de Málaga, España

Daniel Schugurensky Universidad de Toronto-Ontario Institute of Studies in Education, Canadá

Orlando Pulido Chaves Universidad Pedagógica Nacional, Colombia

José Gregorio Rodríguez Universidad Nacional de Colombia

Miriam Rodríguez Vargas Universidad Autónoma de Tamaulipas, México

Mario Rueda Beltrán Instituto de Investigaciones sobre la Universidad y la Educación, UNAM México

José Luis San Fabián Maroto Universidad de Oviedo, España

Yengny Marisol Silva Laya Universidad Iberoamericana, México

Aida Terrón Bañuelos Universidad de Oviedo, España

Jurjo Torres Santomé Universidad de la Coruña, España

Antoni Verger Planells University of Amsterdam, Holanda

Mario Yapu Universidad Para la Investigación Estratégica, Bolivia

arquivos analíticos de políticas educativas
conselho editorial

Editor: **Gustavo E. Fischman** (Arizona State University)
Editores Associados: **Rosa Maria Bueno Fisher** e **Luis A. Gandin**
(Universidade Federal do Rio Grande do Sul)

Dalila Andrade de Oliveira Universidade Federal de Minas Gerais, Brasil
Paulo Carrano Universidade Federal Fluminense, Brasil
Alicia Maria Catalano de Bonamino Pontifícia Universidade Católica-Rio, Brasil
Fabiana de Amorim Marcello Universidade Luterana do Brasil, Canoas, Brasil
Alexandre Fernandez Vaz Universidade Federal de Santa Catarina, Brasil
Gaudêncio Frigotto Universidade do Estado do Rio de Janeiro, Brasil
Alfredo M Gomes Universidade Federal de Pernambuco, Brasil
Petronilha Beatriz Gonçalves e Silva Universidade Federal de São Carlos, Brasil
Nadja Herman Pontifícia Universidade Católica –Rio Grande do Sul, Brasil
José Machado Pais Instituto de Ciências Sociais da Universidade de Lisboa, Portugal
Wenceslao Machado de Oliveira Jr. Universidade Estadual de Campinas, Brasil

Jefferson Mainardes Universidade Estadual de Ponta Grossa, Brasil
Luciano Mendes de Faria Filho Universidade Federal de Minas Gerais, Brasil
Lia Raquel Moreira Oliveira Universidade do Minho, Portugal
Belmira Oliveira Bueno Universidade de São Paulo, Brasil
Antônio Teodoro Universidade Lusófona, Portugal
Pia L. Wong California State University Sacramento, U.S.A
Sandra Regina Sales Universidade Federal Rural do Rio de Janeiro, Brasil
Elba Siqueira Sá Barreto Fundação Carlos Chagas, Brasil
Manuela Terrasêca Universidade do Porto, Portugal
Robert Verhine Universidade Federal da Bahia, Brasil
Antônio A. S. Zuin Universidade Federal de São Carlos, Brasil