COMMENTARY

education policy analysis archives

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

epaa aape

Arizona State University

Volume 23 Number 2 January 5th, 2015 ISSN 1068-2341

Issues of Teacher Performance Stability are Not New: Limitations and Possibilities

Thomas L. Good
The University of Arizona
&
Alyson L. Lavigne
Roosevelt University
United States

Citation: Good, T. L., & Lavigne, A. L. (2014). Issues of teacher performance stability are not new: Limitations and possibilities. *Education Policy Analysis Archives, 23*(2).
http://dx.doi.org/10.14507/epaa.v23.1916

Abstract: Morgan, Hodge, Trepinski, and Anderson (2014) have written an article that continues to confirm what we have known for some time—teacher effects on student achievement have limited stability. In this commentary, we address the *other* potential contributions this work can make to inform practice, policy, and research. While illustrating Morgan et al.'s inattention to history, we take the opportunity to reframe their findings. Considering the authors' work in the context of past and current research, we illustrate that this collective set of stable evidence should convince policymakers that it is *not* reasonable to assume that teachers and teaching is stable across time. Beyond this important opportunity to influence policy, we believe these findings underscore the need to build upon and expand the dependent measures we use to define and understand good teaching. After all, as we have noted (Lavigne & Good, 2014; in press) good teaching involves much more than increasing students' scores on standardized achievement tests.
Keywords: teacher evaluation policy; teacher stability; teacher effects; teacher performance

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

Manuscript received: 11/01/2014
Revisions received: 11/28/2014
Accepted: 11/28/2014

Los problemas de estabilidad del desempeño docente no son nuevos: Limitaciones y posibilidades.
Resumen: Morgan, Hodge, Trepinski, y Anderson (2014) han escrito un artículo que sigue confirmando lo que hemos sabido sobre la estabilidad limitada que los efectos docentes tienen en los logros de los estudiantes. En este comentario, queremos señala *otras* contribuciones potenciales de este trabajo que podrían informar la práctica, política e investigación educativa. Mientras que señalamos que Morgan y colegas no prestaron suficiente atención a la historia, tenemos la oportunidad de replantear sus hallazgos. Teniendo en cuenta el trabajo de los autores en el contexto de investigaciones pasadas y actuales, proponemos que este conjunto colectivo de pruebas estable debe convencer a los políticos de que no es razonable suponer que los docentes y la enseñanza es estable a través del tiempo. Más allá de esta importante oportunidad de influir en la política, creemos que estos resultados ponen de relieve la necesidad de aprovechar y ampliar las medidas dependientes que usamos para definir y entender la buena enseñanza. Después de todo, como hemos señalado (Lavigne & Good 2014; en prensa) buena enseñanza implica mucho más que el aumento de calificaciones de los estudiantes en pruebas estandarizadas
Palabras clave: política de evaluación docentes; estabilidad del profesorado; efectos docentes; desempeño docente

Problemas de estabilidade de desempenho dos professores não são novos: Limitações e possibilidades.
Resumo: Morgan, Hodge, Trepinski, e Anderson (2014) escreveram um artigo que confirma o que já sabemos sobre a estabilidade limitada dos efeitos que os professores têm sobre as avanços dos estudantes. Nese comentário, observamos *outras* contribuições potenciais do presente trabalho, que poderia informar a prática, política e pesquisa educacional. Enquanto notamos que Morgan e seus colegas não prestaram atenção suficiente para a história, temos a oportunidade de repensar as suas conclusões. Considerando-se o trabalho dos autores no contexto do passado e atual da pesquisa, propomos que este conjunto coletivo de provas estáveis deveria convencer os políticos de que não é razoável supor que os professores e o ensino é estável ao longo do tempo. Além deste importante oportunidade de influenciar a política, acreditamos que estes resultados destacam a necessidade de construir e expandir as medidas que usamos para definir e compreender o que é o bom ensino. Afinal, como já observamos (Lavigne & Good 2014; no prelo) bom ensino envolve muito mais do que o aumento na pontuação em testes padronizados
Palavras-chave: avaliação da política educacional; estabilidade dos professores; fins educativos; desempenho dos professores

Introduction

Morgan, Hodge, Trepinski, and Anderson (2014) have written a useful article providing yet more data to demonstrate both the low stability of teacher performance and teacher effectiveness over time. Their work assumes additional *potential* value in that it studied the effects of 132 teachers over five years using measures of teacher performance (observation ratings) and teacher effectiveness (standardized achievement tests), with assessments in multiple curriculum subjects. The authors found a weak relationship between an observational measure of teaching performance (in this case the TAP observation system, see Jerald & Van Hook, 2011) and standardized measures of student achievement. This complements others' recent reports of low correlations between other observational measures and student achievement (Cohen, in press; Kane, Kerri, & Pianta, 2014) and

evidence that teacher actions are not very stable even from lesson to lesson (Patrick & Mantzicopoulos, 2014). This collective set of stable evidence should convince policymakers that the assumptions imbedded in Race to the Top about the easy use of observational and achievement data to evaluate teaching are faulty. Identifying effective teachers is more difficult than they believed.

Having acknowledged the authors' contribution to the literature, we express our disappointment with their inattention to history. We believe that their ahistorical framing of the issue is misleading and this presentation limits the value of their contribution. The authors start their abstract with the observation that, "The last five to ten years has seen a renewed interest in the stability of teacher behavior and effectiveness" (Morgan et al., 2014, p. 1). And they restate their position in the introductory part of the article. Yet, the reader is never told what previous research had found about the stability of teacher behavior and effectiveness. This is like starting in the middle of a book! A better and more logical starting point for the authors would have been the provision of a brief review of the research on teacher stability—what preceded this renewed interest in the stability of teacher effects? If the authors had reviewed earlier literature, they would have found that their major finding was known long before they conducted their research. And access to this information is not difficult to obtain. For example, Konstantopoulos (2014) noted that issues of teacher stability have been studied for *decades* and provided a brief and effective review of this literature. He noted that stability of teaching has been studied in different ways including the degree of stability of teacher effects: when a teacher teaches the same material to a different group of students (e.g., Rosenshine, 1970), across instructional periods in the same year (e.g. Emmer, Evertson, & Brophy, 1979; Rosenshine, 1970), and when a teacher teaches different classes of students over time (e.g., Brophy, 1979). Across these varied contexts, Konstantopoulos (2014) concluded that the stability of teacher effects was *low*. Other earlier researchers also had reported that teacher effects on students' achievement were not highly stable. For example, Berliner (1976) reported, "Our own research, just completed, involved about 200 elementary school teachers, each of which taught a 2-week, specially designed teaching unit in reading and mathematics. Residual gain scores for each subject matter were calculated. These measures of effectiveness using different content and the same students were correlated. From these data we found that measures of effectiveness in the two curriculum areas correlated about .30" (p. 379).

Brophy (1973) studied 165 elementary school teachers' stability of residual gain estimates of their impact on student achievement over 3 years and found that roughly 14% of teachers had high effects on students over 3 years 14% had low effects for three consecutive years. Further, some teachers showed linear increases or decreases over time and 49% of teachers' residual gain scores were inconsistent over time. Elsewhere, Good and Grouws (1977) studied over 100 third and fourth grade teachers' stability over time on the Iowa test of Basic Skills. They found even *lower* levels of stability than had Brophy when they considered teacher residual gains over time across all math subtests.

We acknowledge that Morgan et al. (2014) provided an important replication with a rich data set. However, we are uncertain whether or not they are aware of their replication. We have no basis for concluding why the authors did not mention the easily available historical research on teacher stability. Perhaps they were not aware of previous research or thought it unimportant? Perhaps they felt that they had insufficient space to acknowledge it? Or perhaps they thought that framing the problem without any history made their argument "fresher" or more unique? Or they may have felt that their replication would be viewed as less important than a "new" contribution. This is possible as, historically, replications have been perceived as less valuable (for a review of bias towards replications, see Makel & Pluker, 2014). Yet, we and others contend otherwise. The argument for replication has been made exceedingly well by Makel and Plucker (2014). They write, "If education

research is to be relied upon to develop sound policy and practice, then conducting replications on important findings is essential to moving toward a more reliable and trustworthy understanding of educational environments" (p. 313). Makel and Plucker (2014) have strongly recommended that factual evidence should be valued over putative novelty. We agree, and this is especially the case when new and replicating factual information can be linked to important policy decisions/actions. If we consider the increasing changes in the curriculum, diversity of student populations, changes in the teaching force, and arguably better statistical measures of teachers' impact on student learning (i.e., value-added estimates rather than residual gain scores), it would seem important to determine if earlier research describing the low stability of teacher effects was still accurate. In other words, given these changes, is teacher performance and effectiveness now more stable than in another era? This is a timely and important question as states are using teacher evaluation data to make hire, fire, and grant tenure to teachers.

At the same time, had the authors acknowledged that the field has known for decades that teaching effects on student achievement have limited stability, they would have framed their research focus more accurately and appropriately. And, had they examined stability in other professions (e.g., coaching), they would have found that expectations for stable performance over time cannot easily be assumed. For example, an examination of the top football teams illustrates that fluctuation in performance is considerable even week to week. Table 1 illustrates how much change occurred in the top ranked football teams over time and in one week.

Table 1
A Comparison of the 25 Top Football Teams as Ranked by Associated Press at Preseason, Week 12, and Week 13, 2014: How Stable are Top Teams Over Time?

	Ranking*		
Team	Preseason	Week 12	Week 13
Florida State	1	2	1
Alabama	2	4	2
Oregon	3	3	3
Oklahoma	4	—	23
Ohio State	5	8	7
Auburn	6	9	16
UCLA	7	14	11
Michigan State	8	12	10
South Carolina	9	—	—
Baylor	10	6	6
Stanford	11	—	—
Georgia	12	16	9
LSU	13	20	—
Wisconsin	14	22	14
USC	15	—	24
Clemson	16	18	—
Notre Dame	17	15	—
Mississippi	18	10	8
Arizona State	19	7	13
Kansas State	20	13	12
Texas A&M	21	—	—
Nebraska	22	11	21
North Carolina	23	—	—
Missouri	24	—	19
Washington	25	—	—
Mississippi State	—	1	4
TCU	—	5	5
Arizona	—	17	15
Duke	—	19	25
Marshall	—	21	18
Colorado State	—	23	22
Georgia Tech	—	24	17
Utah	—	25	20

Note. * The symbol, "—" signifies not ranked.

Clearly, these data indicate that stability of football performance shows noticeable change from the preseason to week 12. Notably 8 teams ranked in the preseason top 25 were no longer

ranked at week 12, and 8 teams ranked in week 12 were not ranked in the preseason. Some of these fluctuations were dramatic as teams ranked 4, 9, 11, and 15 in the preseason were not ranked at week 12. Further, in week 12 the teams ranked 1st and 5th were not ranked in the preseason. Even in one week, week 12 to 13, there is variation both within the week and over time. For example, Oklahoma who was ranked 4 in the preseason fell out of the ratings in the 12th week, and returned as the 23rd team in week 13. And as we submit our paper on November 21 we suspect other evidence of unstable performance will appear in week 14 and beyond.

Applying these data to high stakes testing of teachers we suggest that one could consider the preseason ranking as a "valued added" score. The expert raters take into consideration the quality of returning and new player talent/achievement potential, the schedule, and other factors and then rank the teams in terms of expected performance ranks. Should coaches at schools like South Carolina, Stanford, Texas and Washington be fired because their talented teams did not live up to expectations? Should coaches at schools like TCU, Arizona, Georgia Tech, Utah, Colorado State and Duke be awarded bonuses and tenure because they achieved more with teams who were judged to have less talent?

Konstantopoulos (2014) has noted that stability of performance varies markedly both in athletics and business. Sobolevska (2014) has examined the performance of the top 200 professional golfers in America and their stability over six consecutive years. She concluded that the stability of golfers is very low and many top golfers do not appear on the list from one year to the next. Generally speaking, it would seem that the conditions that surround professional golfing are quite standard in comparison to the conditions that teachers face. Professional golfers use comparable equipment, and the rules for determining successful performance are vastly clearer than for teachers. Although conditions change from course to course on a given day, top golfers playing in the same tournament face the same course and the same weather conditions. Teachers on the other hand, face conditions that are inherently complex and unequal. Some teachers are teaching primarily talented students or teaching students who are from low-income homes and who often have not received appropriate preschool opportunities and or educational challenges. Some students come to school rested and well prepared for instruction, whereas other students arrive at school tired, hungry, and unprepared. Morgan et al.'s framing of the research question implicitly suggests that it is reasonable to assume that teacher performance and effects are stable. We find this assumption misleading if not erroneous.

Failure to Report History: A Lost Opportunity to Influence Policy

Policymakers hold extraordinary and often exaggerated beliefs about what teachers can and cannot accomplish. We (Lavigne & Good, in press) and others (Berliner, 2014; Biddle, 2014) have noted, teachers certainly influence achievement but that the effects of poverty on student achievement are huge as do data from the Program for International Student Assessment (PISA) show. Data from the PISA assessment illustrates that countries that distribute resources to schools equally are those that have higher student achievement. Countries like the US that distribute resources unequally across schools have lower student achievement. The PISA data show that low SES accounts for about 20% of the variance in student achievement scores. Although the authors cannot provide an exhaustive historical context for everything they discuss, we do think it would have been reasonable for them to have briefly explored the question of how much can teachers be expected to overcome in a given year and to have questioned the extent to which teachers can be expected to have stable effects given the fluctuations and their own lives (Spencer, 1986) and in the lives of students they teach.

Yet, generally, the research community, like these authors, was silent and did not question initially the *easy* assumptions when policymakers and Race to the Top (RttT) advocates advanced simplistic strategies for conducting high stake teacher evaluations by using classroom observation and student achievement data. Previous research provided many caveats about the difficulty of linking classroom process and student achievement (Brophy & Good, 1986; Everston & Green 1986; Good & Brophy, 2008). Although talking "truth to power" may not have slowed the powerful forces (including RttT and substantial funding from the Bill and Melinda Gates Foundation) clamoring for evaluating teachers on performance and outcome measures, but, it would have been worth the effort. What if educational researchers early in the formation of RttT had pointed out that previous research provided clear knowledge that teacher performance and effectiveness was not generally highly stable over consecutive years? Could this stance have led to reasoned debate on issues such as – are today's observational and statistical techniques sufficiently better (contain less measurement error) and more capable of linking teaching and learning now than they were previously?

The history of educational reform has been one of consistent failure as the field moves from fad to fad (Cuban, 2013; Good & Braden, 2000; McCaslin, 1996; Payne, 2010; Ravitch, 2010; Tyack & Cuban, 1977). Over time policy makers have repeatedly implemented simple solutions for reform based upon little if any research. These reforms are costly and waste enormous amounts of time and money. These solutions are quickly abandoned because they do not provide immediate answers to complex problems. But, after these failures, yet again another simple but costly reform appears. We have reviewed these issues and the interested reader can obtain our detailed arguments elsewhere (Lavigne & Good, 2014, in press). But briefly, Lavigne and Good (2014) characterized the failure as a series of steps. Each crisis is an acute concern about students' poor achievement on standardized tests. Each crisis is based upon the performance of American students in comparison to their international peers. The new general solution is based on the rejection of the status quo and a call for something new and notably different from current practice. In time data suggest that the new reform has not solved the problem and new reforms are eventually sought. Typically the professional research community has not actively spoken against problematic reforms in a timely fashion. Ultimately, again American policymakers embark on costly new endeavors even without carefully defining what the new movement involves. In reviewing these characteristics of failed reform, we believe that had Morgan et al. (2014) missed an important opportunity to make a modest attempt at challenging these chronic patterns of failed reform. We contend that it is more reasonable for research to precede rather than to follow reform efforts. Further, we also believe that when research is available it should be recognized and used.

Measuring Good Teaching: Another History Lesson

Much of the current reform rhetoric is based on the assumption that we can measure good teaching. What is good teaching and can it be captured with any observational instrument? Within the context of this article, it seems the authors assume that appropriate teaching involves increasing student performance on standardize achievement tests. And, the authors persuasively note the importance of validating observational instruments (and even discuss how to do it), yet, they have little to say about the validity of the TAP program/instrument that they used in their study. The authors only address this issue (on page 6) when they quote Jerald and Van Hook (2011, p. 4), that the instrument's 'indicators provide sufficient breadth to ensure that evaluation ratings reflect the kind of effective instructional practices that predict positive learning outcomes. Unfortunately, the authors do not provide a reference for Jerald and Van Hook (2011); however, an examination of

that source does not provide a clear demonstration that the instrument is predictive of value-added achievement.

Readers are still left wondering, how does the TAP measure good teaching and how/if/why the TAP can be considered an appropriate measure of good teaching? These appear to be important issues to address, particularly because Morgan et al. (prior to describing the observational procedures they used in the study) note, "…there is remarkably little research to guide such critical decisions as which teachers to hire, retain, remunerate, and promote" (Rice, 2003).

It is useful to *note* that a conclusion remarkably similar to Rice's was expressed fifty years earlier by a committee appointed by the American Educational Research Association.

> The simple fact of the matter is that, after forty years of research on teacher effectiveness during which a vast number of studies have been carried out, one can point to few outcomes that a superintendent of schools can safely employ in hiring a teacher or granting him tenure, that an agency can employ in certifying teachers or that a teacher-education faculty can employ in planning or improving teacher-education programs. (AERA, 1953, p. 657).

So we and readers are left with a least two questions. First, did the field know anything more about good teaching when Rice wrote in 2003 than was known in 1953? Similarly, do we know any more now in 2014 than we did in 1953? Simply put, if we are to design an observational system believed to be predictive of student achievement, we need to have some knowledge based on theory and research that links teaching to higher or lower levels of student achievement and some reason to believe that the observational system we use includes those key teacher actions. Given that the authors used Rice's conclusions that the field has limited capacity for making critical decisions such as teacher retention, it seems important, to ask why it is plausible to assume now that observational systems generally and the TAP specifically can be used for high-stakes decisions. Clearly, the authors were willing to at least implicitly accept the belief that TAP successfully captures the teacher actions that lead to higher value added scores.

Other Opportunities to Inform Practice, Policy, and Research

We now turn to consider the value of Morgan et al.'s work (2014) in the context of the brief history that we have provided. We comment on additional contributions this work can make today to inform policy, practice, and research as they further explore their data.

Study Participants

It is not clear why researchers did not include teachers who changed grades or schools. Obviously, they could not be included in the overall quantitative analyses, but an examination of teachers who moved would seemingly provide useful descriptive information. For example, were teachers who changed grades/schools more or less stable than those who stayed in their same context? Should we "anticipate" or account for instability (of teacher actions or teacher effects on students) as a natural function of adjusting to a new setting? Did teachers who moved from an elementary school setting to a middle school setting (or vice versa) have higher or lower effects? That is, there is considerable evidence to suggest that teachers in elementary schools are rated higher on observational measures than are teachers in middle schools (Mihaly & McCaffrey, 2014). Were these findings also obtained in this research? Further, it is not clear whether subject matter was an important mediator of teacher ratings (were teachers rated higher in math than in reading)?

Measures of Effective Teaching

The above points raise an additional question: How were 5 year of data over multiple years and subjects combined? It is not clear how the authors combined their data to show the effects of teachers on students. We are told that each year grade 3-8 students were administered the Palmetto Assessment of State Standards. We also know that students in grades 4-7 were administered science and social studies tests. So when the authors discuss the performance of a teacher for whom we have multiple subjects, is the performance an aggregation of multiple tests over years? We are not suggesting that the data analysis is inappropriate; however we do suggest that in our reading of the article it was not clear to us how the data analysis was conducted for teachers in grades 4-8.

Observers

More information about coders and their training and deployment would be helpful in understanding the research methods. For example, the authors described the observations as "expert" observations. Given the current focus on high stakes testing, expert observers often suggest the use of highly trained external observers who have passed rigorous training. However, we are informed that the observers were school administrators, mentor teachers, and master teachers. It is not clear what criteria were used for defining master teachers or who made those designations. How were they trained on the observation instrument and were they familiar with using the TAP prior to this research? How was coder drift accounted for (seemingly in an article dealing with stability of teacher performance, we might expect some discussion of coder stability)? Were teachers observed by different observers? And, is it possible that who is doing the observation may be more important to reliability than how many observations were conducted? The authors do address this issue, but only briefly and we think more information would help readers to understand better their observational procedures. Further, it is not clear why observational results were rounded. There may be good reason for doing this, but without more information it seems that this decision would limit variation and potentially restrict the correlation between observation data and value-added data.

The Importance of Context

The authors provide commentary from other researchers that likely mediate and explain the lack of stability of performance and effectiveness and the low relationship between teacher actions and achievement. They also add to this discussion by noting, "Finally, there are contextual differences, such as grade level, subject matter, and classroom size and composition" (p. 14). We agree with the authors that there is substantial evidence that these are important considerations. And, for an especially thoughtful discussion of these context issues, see Berliner (2014). However, since the authors have grade level and subject matter data, why not discuss these possibilities in their own data set?

The authors also place considerable emphasis upon the fact that observational ratings for teachers became higher over time, but it is not clear that this "instability" was not actually a function of teachers. For example, there is some evidence that teachers increase in their performance during the first few years of teaching, but this evidence is not reviewed in the paper nor is any consideration given to the fact that related research suggests that new teachers become more effective at least for the first few years. Furthermore, the authors suggest that in general observational ratings were higher than value added performance. However, they note a context finding suggesting that in average or above average performing schools, observation ratings and value added ratings were similar. But that in poor performing schools, teachers had higher mean observational ratings than value added ratings. The authors conclude, "This suggests that on average the observational ratings may over estimate teacher effectiveness in lower performing schools." (p. 12). However, this

conclusion appears to be based on a general average and it is not clear whether this was true equally across all subjects. Further, there are alternate explanations for this finding which are not explored including the possibility that teachers in low performing schools were actually scoring higher on the TAP because they were actually performing the behaviors better over time that this instrument measures.

Concluding Remarks

We end our remarks with a question the authors eventually raise. Is it reasonable to assume that teachers and teaching should be stable? This is a central issue and one that policy makers have spent little time addressing. Today's simple policy orientation is that we can identify good teachers and reward them and identify poor teachers and remediate them or terminate them. Unfortunately, policymakers have not considered these assumptions carefully including the lack of teacher stability over consecutive years (as discussed here). When we apply this knowledge to simulations of high-stakes decision-making, a significant number of teachers are misclassified (Guarino, Reckase, & Wooldridge, 2012; Schochet & Chiang, 2010). Effective teachers are fired and ineffective ones and rewarded. The costs of these misclassifications to teachers, schools, students are insurmountable. If researchers want to understand those teacher actions that relate to student achievement, they need to be very sure that they are studying teachers who have stable performance and effects. However, from research presented by Brophy in 1973 to research presented by Berliner in 2014, we know that such teachers are not common (recall the Brophy, 1973, reported that 14% of highly effective teacher and 14% of low effective teachers held their rating over 3 consecutive years).

> Berliner (2014) aptly summarizes the issue of teacher stability,
> Although hard to ferret out in their "pure" form as an independent main effect, teacher effects on student achievement exist, and they are likely to be strong enough for us all to worry about who teaches our children and what their training has been. There does seem to be a small percentage of teachers who show consistency no matter what classroom and school compositions they deal with. Those few teachers who have strong and consistent positive effects on student outcomes, we should learn from and reward. And, those few teachers who have strong negative effects on student outcomes need to be helped or removed from classrooms. But the fundamental message from the research is that the percentage of such year-to-year, class-to-class, and school-to-school effective and ineffective teachers appears to be much smaller than is thought to be the case. (p. 27)

We agree. Perhaps one direction for future research is an examination of the patterns that Berliner (2014) addresses, with a focus on when and why stability should be expected. Given that the majority of teachers do not fall into stable patterns, as traditionally defined, the field might benefit from further considering our expectations for reasonable stability of professional practice. Further, future research should build upon and expand the dependent measures we use to define/understand good teaching. After all, good teaching involves much more than increasing students' scores on standardized achievement tests. Good teaching includes helping students to become better problem finders and problem solvers, as well as encouraging student civility, social responsibility, and much more (Lavigne & Good, in press). It is prudent to recall that as recently as the late 60's teachers were not considered to have much impact on students' achievement and that students' success in schools was primarily determined by student and family variables. If we look using good research procedures, we may well find evidence that some aspects of teaching and their consequences on students are more enduring than teacher effects on standardized achievement scores.

References

American Educational Research Association, Committee on the Criteria on Teacher Effectiveness. (1953). Second Report of the *Journal of Educational Research*, *46*, 641–658.

Berliner, D. (1976). A status report on the study of teacher effectiveness. *Journal of Research in Science Teaching*, *13,* 369–382. http://dx.doi.org/10.1002/tea.3660130415

Berliner, D. (2014). Exogenous variables and value-added assumptions: A fatal flaw. *Teachers College Record, 116,* 1–31.

Brophy, J. (1973). Stability of teacher effectiveness. *American Educational Research Journal*, *10,* 245–252. http://dx.doi.org/10.3102/00028312010003245

Brophy, J. & Good, T. (1986). Teacher Behavior and Student Achievement. In M. Wittrock (Ed.), Third Handbook on Research on Teaching (pp. 328-275). Chicago: Rand McNally.

Cohen, J. (In press). The challenge of identifying high-leverage practices. *Teachers College Record*,*117*(8).

Cuban, L. (2013). *Inside the black box of classroom practice. Change without reform in American education.* Cambridge, MA: Harvard Education Press.

Emmer, E., Evertson, C., & Brophy, J. (1979). Stability of teacher effects in junior high classrooms. *American Educational Research Journal***,** 16, 71–75. http://dx.doi.org/10.3102/00028312016001071

Evertson, C., & Green, J. (1986). Observation as inquiry and method. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 162–213). New York, NY: Macmillan.

Good, T. & Brophy, J. (2008). *Looking in Classrooms*. New York: Longman.

Good, T. & Grouws, D. (1979). The Missouri Mathematics Effectiveness Project: An experimental study in 4th grade classrooms. *Journal of Educational Psychology*, 71, 355–362. http://dx.doi.org/10.1037/0022-0663.71.3.355

Good, T. & Grouws, D. (1977). Teaching effects: A process-product study in 4th-grade mathematics classrooms. *Journal of Teacher Education*, *28,* 49–54. http://dx.doi.org/10.1177/002248717702800310

Good, T. L. & Braden, J. S. (2000). *The great school debate: Choice, vouchers, and charters*. Mahwah, NJ: Erlbaum.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2012). *Can value-added measures of teacher education performance be trusted?* (Working paper # 18). Michigan State University Education Policy Center.

Jerald, C. & Van Hook, K. (2011). More than measurement: The TAP System's Lessons Learned for Designing Better Teacher Evaluation Systems. National Institute for Excellence in Teaching.

Kane, T., Kerr, K., & Pianta, R. (2014). Designing teacher evaluation systems: New guidance from the measures of effecting project. San Francisco, CA: John Wiley & Sons Inc.

Konstantopoulos, S. (2014). Teacher effects, value-added models, and accountability. *Teachers College Record***,** *116*(1), 1–21.

Lavigne, A. & Good, T. (2014). *Teacher and student evaluation: Moving beyond the failure of school reform.* New York, NY: Routledge.

Lavigne, A. & Good, T. (in press). *Improving teaching through observation and feedback: Going beyond state and federal mandates*. New York, NY: Routledge.

McCaslin, M. (1996). The problem of problem representation: The Summits' conception of student. *Educational Researcher, 25*(8), 13–15. http://dx.doi.org/10.3102/0013189X025008013

Makel, M. C. & Plucker, J. A. (2014). Facts are more important than novelty: Replication is the education science. *Educational Researcher*, *43*, 304–316. http://dx.doi.org/10.3102/0013189X14545513

Mihaly, K. & McCaffrey, D. (2014). Grade-level variation in observational measures of teacher effectiveness, in T. Kane, K Kerr and R. Pianta (Eds) *Designing Teacher Evaluation Systems; New guidance from the measures of effective teaching project*, pp. 9–49. San Francisco, CA: Jossey-Bass.

Morgan, G. B., Hodge, K. J., Trepinksi, 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). Retrieved from http://dx.doi.org/10.14507/epaa.v22n95.2014

Patrick, H. & Mantzicopoulos, P. (2014). Is effective teaching stable? *The Journal of Experimental Education, 0*(0), 1–25. http://dx.doi.org/10.1080/00220973.2014.952398

Payne, C. (2010). *So much reform, so little change: The persistence of failure in urban schools*. Cambridge, MA: Harvard Education Press.

Ravitch, D. (2010). *The death and life of the great American school system: How testing and choice are undermining education.* New York, NY: Basic Books.

Rice, J. (2003). *Teacher quality: Understanding the effectiveness of teacher attributes.* Washington, DC: Economic Policy Institute.

Rosenshine, B. (1970). The stability of teacher effects upon student achievement. *Review of Educational Research, 40*, 647–662. http://dx.doi.org/10.3102/00346543040005647

Schochet, P. Z. & Chiang, H. S. (2010). *Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains.* Report for National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education (Washington, D.C., 2010).

Sobolevska, G. (2014). Why do we expect teachers to be stable? An unpublished manuscript. University of Arizona, Tucson, AZ.

Spencer, D. (1986). *Contemporary women teachers: Balancing school and home.* New York: Longman.

Tyack, D. & Cuban, L. (1997). *Tinkering toward utopia: A century of public school reform.* Cambridge, MA: Harvard University Press.

About the Authors

Thomas L. Good
University of Arizona
goodt@email.arizona.edu
Dr. Thomas L. Good is a professor emeritus of the Department of Educational Psychology at the University of Arizona. He was the long time editor of the *Elementary School Journal*. He formerly taught at the University of Missouri-Columbia and The University of Texas -Austin. He is a member of the National Academy of Education. His most recent publications include two books co-authored with Alyson L. Lavigne, entitled *Teacher and Student Evaluation: Moving Beyond the Failure of School Reform* (published in 2014) and *Improving Teaching through Observation and Feedback: Going Beyond State and Federal Mandates* (to be released in early 2015).

Alyson L. Lavigne
Roosevelt University
allavigne@roosevelt.edu
Dr. Alyson L. Lavigne, (pka Alyson Lavigne Dolan), is an assistant professor in the College of Education at Roosevelt University. She received her Ph.D. in Educational Psychology from the University of Arizona. Her work in the area teacher evaluation and supervision includes a co-edited *Teachers College Record* special issue on high-stakes teacher evaluation and two co-authored books with Thomas L. Good.

education policy analysis archives

Volume 23 Number 2 January 5th, 2015 ISSN 1068-2341

SOME RIGHTS RESERVED 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/EPAAAAPE 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), **Kevin Kinser** (University of Albany)
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 Jefferson County Public Schools in Golden, Colorado
Casey Cobb University of Connecticut
Arnold Danzig California State University, San Jose
Antonia Darder Loyola Marymount University

Linda Darling-Hammond Stanford University
Chad d'Entremont Rennie Center for Education Research and Policy
John Diamond Harvard University
Tara Donahue McREL International
Sherman Dorn Arizona State University
Christopher Joseph Frey Bowling Green State University
Melissa Lynn Freeman Adams State College
Amy Garrett Dikkers University of North Carolina Wilmington
Gene V Glass Arizona State University
Ronald Glass University of California, Santa Cruz
Harvey Goldstein University of Bristol
Jacob P. K. Gross University of Louisville

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
Michele S. Moses University of Colorado, Boulder
Julianne Moss Deakin University

Sharon Nichols University of Texas, San Antonio
Noga O'Connor University of Iowa

João Paraskveva University of Massachusetts, Dartmouth
Laurence Parker University of Utah
Susan L. Robertson Bristol University

John Rogers University of California, Los Angeles
A. G. Rud Washington State University
Felicia C. Sanders Institute of Education Sciences
Janelle Scott University of California, Berkeley

Kimberly Scott Arizona State University
Dorothy Shipps Baruch College/CUNY

Maria Teresa Tatto Michigan State University
Larisa Warhol Arizona State University
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 Center for Applied Linguistics
John Willinsky Stanford University
Kyo Yamashiro Los Angeles Education Research Institute

archivos analíticos de políticas educativas
consejo editorial

Editores: **Gustavo E. Fischman** (Arizona State University), **Jason Beech** (Universidad de San Andrés), **Alejandro Canales** (UNAM) y **Jesús Romero Morante** (Universidad de Cantabria)

Armando Alcántara Santuario IISUE, UNAM México

Claudio Almonacid University of Santiago, 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 DIE-CINVESTAV, Mexico
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 DIE-CINVESTAV México
Alejandro Márquez Jiménez IISUE, UNAM México
Jaume Martínez Bonafé, Universitat de València, España
José Felipe Martínez Fernández University of California Los Angeles, Estados Unidos

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 Universidad de San Andrés, Argentina
Ignacio Rivas Flores Universidad de Málaga, España
Daniel Schugurensky Arizona State University, Estados Unidos
Orlando Pulido Chaves Instituto para la Investigacion Educativa y el Desarrollo Pedagogico IDEP
José Gregorio Rodríguez Universidad Nacional de Colombia
Miriam Rodríguez Vargas Universidad Autónoma de Tamaulipas, México
Mario Rueda Beltrán IISUE, 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 Barcelona, España
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 Pontificia 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 Pontificia 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
António 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 University of York