Washington State
Institute for
Public Policy

110 Fifth Avenue Southeast, Suite 214 • PO Box 40999 • Olympia, WA 98504-0999 • (360) 586-2677 • www.wsipp.wa.gov

May 2012

TEACHER COMPENSATION AND TRAINING POLICIES: IMPACTS ON STUDENT OUTCOMES

Washington State's Quality Education Council (QEC) was created by the 2009 Legislature to make recommendations regarding basic education policy, including financing of the school system.[1] The legislation stipulated that recommendations from the QEC "shall be based on evidence that the programs effectively support student learning."[2] To assist the group in compiling such evidence, the 2010 Legislature directed the Washington State Institute for Public Policy (Institute) to provide research support to the QEC.[3]

The enacting legislation also created a technical working group, with oversight by the QEC, to design a new state salary allocation model "to attract and retain the highest quality educators."[4] The workgroup requested that the Institute conduct research reviews on six topics related to teacher compensation and training. The reviews examine the impact on student outcomes from:

[1] The QEC is composed of eight members of the legislature and representatives from the Office of the Governor, Office of the Superintendent of Public Instruction, State Board of Education, Professional Educator Standards Board, and Department of Early Learning.
[2] HB 2261 § 114, Laws of 2009.
[3] SB 6444 § 608 (13), Laws of 2010.
[4] HB 2261 § 601, Laws of 2009.

Summary

Washington State's Quality Education Council (QEC) was created by the legislature to make recommendations regarding basic education policy and finance. The legislature also directed the Washington State Institute for Public Policy (Institute) to provide research support to the QEC. This report summarizes six research reviews of the impacts of teacher compensation and training policies on student outcomes (measured by test scores).

The research reveals that some of Washington State's existing policies regarding teacher compensation—such as paying for additional years of experience, NBPTS certification, and teacher induction—are, at least roughly, aligned with evidence regarding teacher effectiveness.

The research also suggests that creating financial incentives for teachers to obtain general graduate training and professional development is not associated with improvements in student test scores. We did find evidence that more focused training, such as in-subject master's degrees and content-specific professional development, can improve student outcomes.

For this report, we have not estimated the benefits and costs of these findings. We will calculate benefits and costs for these topics prior to the 2013 legislative session.

Suggested citation: Pennucci, A. (2012). *Teacher compensation and training policies: Impacts on student outcomes.* (Document No. 12-05-2201). Olympia: Washington State Institute for Public Policy.

Research Approach

For the six research reviews, we focus on a single type of educational outcome: student academic performance. Washington's public school system has many other goals, such as promoting economic well-being, critical thinking, and citizenship.[5] While these goals are important, this review focuses on a narrower question: what works to improve academic outcomes? In the research literature, these outcomes are primarily measured by changes in standardized test scores. Other relevant measures of student academic performance, such as grade repetition and high school graduation, have not yet been measured by a sufficient number of studies to allow for meta-analysis of those outcomes.

We also recognize the importance of non-cognitive abilities, such as "socioemotional skills, physical and mental health, perseverance, attention, motivation, and self confidence."[6] These outcomes have also not yet been measured in a consistent way across the K-12 research literature for meta-analysis.

For these reasons, the findings presented in this report focus on student test scores. Our research approach uses two basic steps.

1. We include <u>all</u> methodologically sound research in our review. To estimate whether a particular K-12 policy or program is likely to affect student academic performance, we systematically assess the findings of all methodologically sound research we can locate. We include studies in our review after screening for methodological rigor and relevance to Washington State. We include random assignment studies, although there are relatively few of these "gold-standard" examples. We also include rigorous quasi-experimental or observational evaluations when special care has been taken to isolate the causal effect of a K-12 policy or program on academic outcomes.

For each high-quality evaluation we find, we compute an "effect size"—a statistical summary measure that indicates the degree to which an evaluated policy or program changes an outcome. Effect sizes are in standard deviation units. The results can be interpreted using the fifth grade Measures of Student Progress math test in 2011 as an example: the average score was 409.2 and the standard deviation, 45.6; thus, a 0.5 effect size represents an increase in average scores by 22.8 points (= 45.6 * 0.5). An effect size of 0.02 (closer to the results presented in this report) increases average scores by about one point.

2. We then compute a policy option's expected impact on student test scores based on the group of methodologically sound studies. We combine the effect sizes from each study to determine whether, on average, outcomes can be expected to change with the policy or program under consideration.[7] While it may be tempting to examine only one or two studies on a topic, we think a restricted review of existing research may lead to unrealistic or biased expectations. By considering all methodologically sound studies on a topic, we seek to determine the average evidence-based effectiveness of each K-12 topic. While above-average performance is always desired, we base expectations on the average evidence-based results. If the empirical evidence is insufficient to draw conclusions about a policy's effectiveness, we say so.

The remaining sections of this report summarize the Institute's reviews of the six topics related to teacher compensation and training. The research shows that some of Washington State's existing policies regarding teacher compensation—such as paying for additional years of experience, induction, and NBPTS certification—are, at least roughly, aligned with the evidence regarding effectiveness. The research also suggests that creating financial

[5] RCW 28A.150.210

[6] Heckman, J. (2008). Schools, skills and synapses. *Economic Inquiry* 46(3), 289–324.

[7] As described in the technical appendix, we calculate mean-difference effect sizes for each methodologically sound study and then meta-analyze these individual effects to produce a weighted average effect size for a group of studies on a particular topic. Generally, we follow the procedures in Lipsey, M. and Wilson, D. (2001). *Practical Meta-analysis*. Thousand Oaks, CA: Sage Publications, with one important exception. Many studies of education topics are based on data that are organized hierarchically; students are nested in classes, which are nested in schools, which are nested in districts. To account for this, we adjust effect sizes and inverse variance weights using methods suggested in Hedges, L. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics* 32(4): 341-370.

incentives for teachers to obtain general graduate training and professional development is not associated with improvements in student test scores. We did find evidence that more focused training, such as in-subject master's degrees and content-specific professional development, can improve student outcomes.

A. Years of Teaching Experience

Washington's teacher salary allocation model starts with base pay for a teacher with a bachelor's degree and zero years of experience ($33,401 in the 2011-12 school year).[8] The salary allocation increases with each additional year of experience up to 16 years.[9] The amount of the increase varies according to the highest level of education attained by the teacher; on average, the salary allocation rises by 2.1 percent per year of teaching experience (7.5 percent in the first five years).

A1. *Effectiveness by Years of Experience*. We located and analyzed 38 high-quality studies from across the United States that examine the relationship between teachers' years of experience and growth in their students' test scores. The studies measure teacher effectiveness at different points in teachers' careers in comparison with novice teachers. From these 38 studies, we computed 146 separate effect sizes at different points in teachers' careers.[10]

Exhibit 1 (next page) plots the average effect size from these 38 studies of teachers at different points in their careers.[11] The error bands around the point estimates indicate the degree of uncertainty around each estimate. The red series is a regression-fitted line.

These results indicate that in the first few years on the job, a teacher progresses considerably in her or his ability to improve the academic performance of students. The effect increases rapidly in years one to five, and then begins to level off. The marginal gains in effectiveness become smaller after these initial years.[12]

B. Graduate Degrees

In addition to increasing salary by years of experience, Washington's teacher salary allocation model provides increases for educational credits earned beyond a bachelor's degree and for a master's degree or higher. In Washington State, the master's degree step on the allocation model is approximately 18 to 20 percent higher than the base bachelor-only salary (depending on the number of years of experience). The research literature examines the effectiveness of teachers with a master's degree or higher, in comparison to teachers with less than a master's degree.

B1. *Impact of Graduate Degrees in General*. We located and analyzed 26 high-quality studies from across the United States that examine the relationship between teachers having a master's degree and growth in their students' test scores. Exhibit 2 (next page) plots the average impacts from these studies.[13]

We conclude from this analysis that there is no consistent relationship between teachers with graduate degrees and increased student outcomes as measured by test scores. While a few studies show graduate degrees to be effective, and a few indicate no or a negative impact, our average estimate, as shown by the vertical line in Exhibit 2, is slightly negative and very close to zero.[14]

[8] LEAP Document 1, May 23, 2011.
[9] To view the salary allocation model, visit http://leap.leg.wa.gov/leap/budget/leapdocs/2011L1.pdf.
[10] Some studies include multiple samples, and many studies measure outcomes at multiple points in teachers' careers. Some studies include estimates for teachers by year bands (e.g., 1-2, 3-5, 6-10). For these studies, we estimate the effect size at the midpoint (e.g., year 4 for the 3-5 range). Other studies measure an average annual gain per year of experience, and include a squared term in the regression equation to model potential changes over time. For these studies, we calculated effect sizes at 2, 5, 10, and 20 years of experience (using the formula: coefficient * years + coefficient * years squared).
[11] For all effect sizes in this report, we present the weighted (for sample size) average impact on student test scores.
[12] These estimates are based on a shifting population of teachers, and estimates at later stages of teachers' careers may confound experience gains with effects of attrition. The average effectiveness of teachers who discontinue teaching in earlier stages of their careers may be different than for those who stay longer.
[13] Exhibit 2 includes 29 effect sizes because some studies included multiple samples.
[14] The studies displayed in Exhibit 2 all examine individual teacher data linked with individual students to test the impact. An additional seven studies examine aggregate data—the percentage of teachers in a school or district with a master's degree, and average student test scores. The results from these seven studies were similar, with a weighted mean average effect size of -0.001.

Exhibit 1
Estimates of the Effect of Years of Teaching Experience on Student Outcomes
Meta-analysis of 38 studies

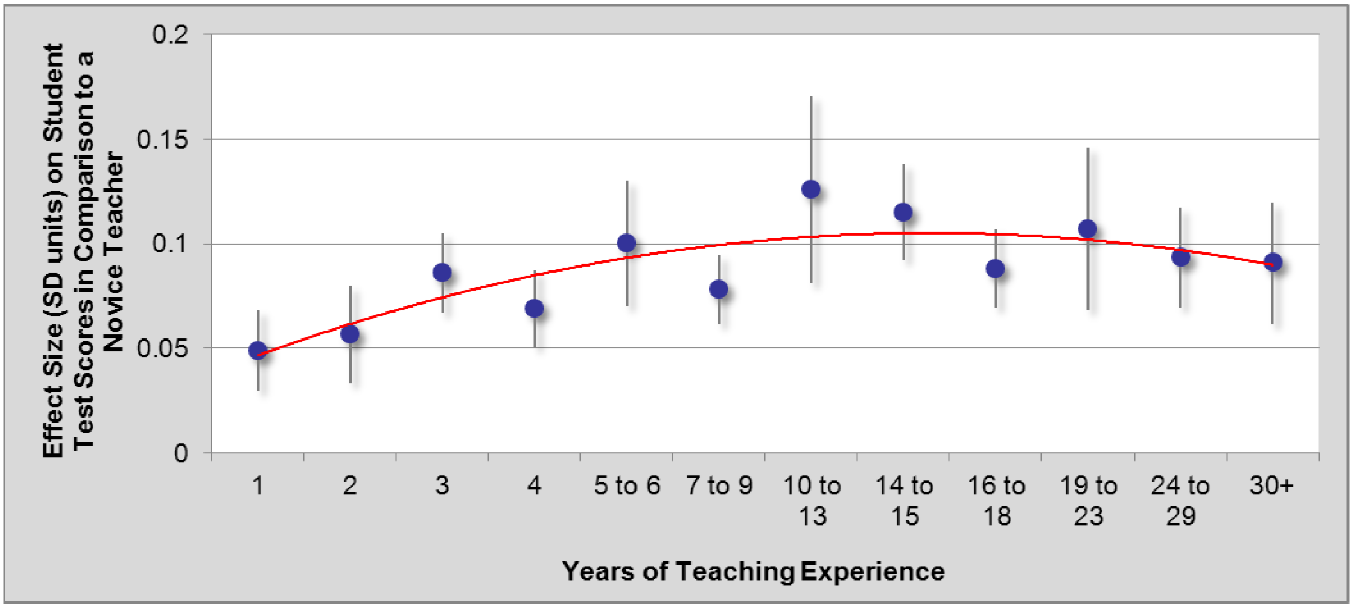
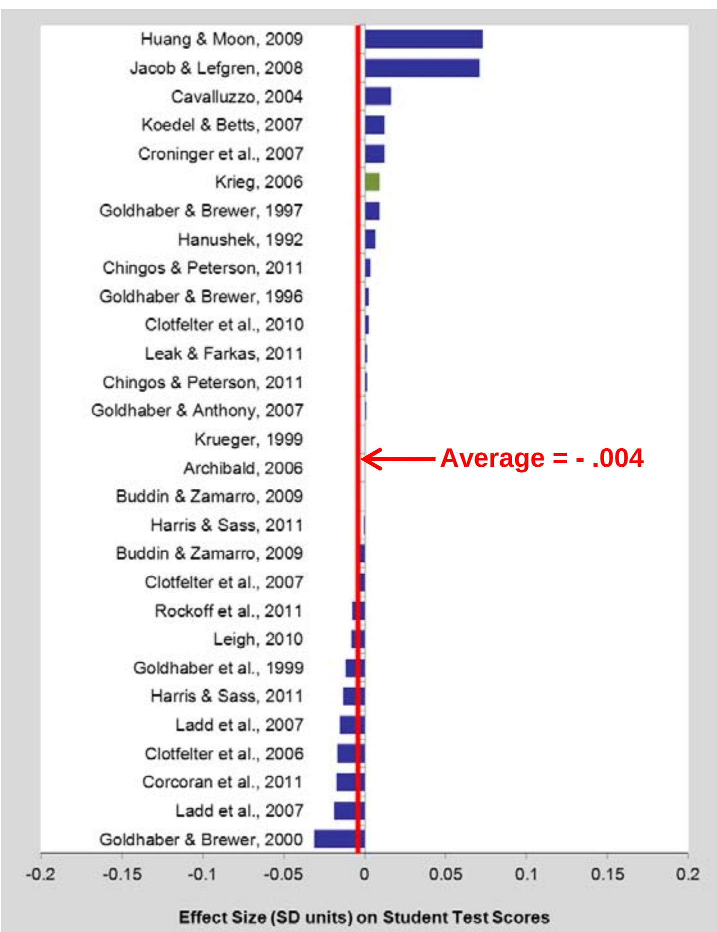


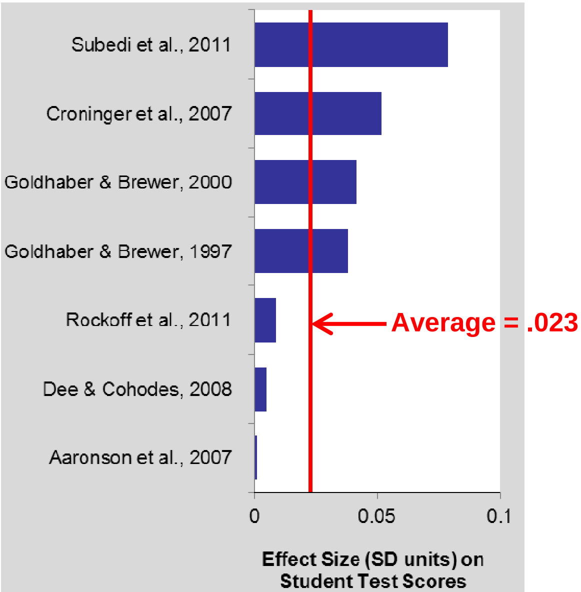
Exhibit 2
Estimates of the Effect of Teacher Graduate Degrees on Student Outcomes
Meta-analysis of 26 studies



B2. *In-Subject Graduate Degrees*. We also located seven rigorous studies that examine whether having an in-subject graduate degree—such as mathematics for a math teacher, science for a science teacher, or elementary education for an elementary school teacher—is associated with higher test scores. Exhibit 3 (next page) plots the results. The average impact is larger than for graduate degrees in general, and positive.

Therefore, while it appears that graduate degrees in general do not have a consistent impact on student test scores, in-subject degrees likely have a positive impact on student learning.

Exhibit 3
Estimates of the Effect Teacher In-Subject Graduate Degrees on Student Outcomes
Meta-analysis of 7 studies



Thus far, we have reviewed research related to Washington State's salary allocation model for teachers, which bases pay increases on years of experience and educational attainment beyond a bachelor's degree. For the remainder of this report, we turn to other ways of compensating teachers, including bonuses and support for professional development.

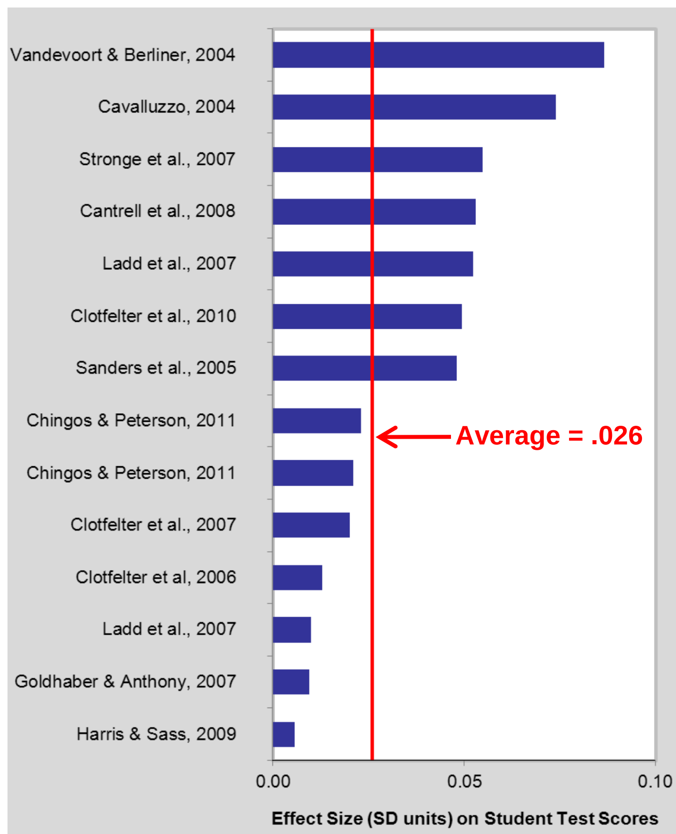
C. National Board for Professional Teaching Standards (NBPTS) Certification

Washington State provides a $5,090 annual bonus for teachers with NBPTS certification.[15] NBPTS is a voluntary national teacher certification system that is complementary to state certification. To obtain this certification, teachers participate in a series of assessments and develop a written and video portfolio to demonstrate their content knowledge and instructional skills. The certification is valid for ten years.

C1. *Impact of NBPTS Teachers*. For this review, we located and analyzed 12 studies that examine growth in student test scores for teachers with and without NBPTS certification. As Exhibit 4 shows, having NBPTS certification is consistently associated with improvements in student test scores.

The research does not, however, clearly indicate whether the NBPTS process itself imrpoves teaching, or whether it simply recognizes above-average teachers. Nonetheless, NBPTS-certified teachers, on average, outperform their non-NBPTS certified peers in terms of improving student test scores.

Exhibit 4
Estimates of the Effect of Having a NBPTS-Certified Teacher on Student Outcomes
Meta-analysis of 12 studies



[15] The bonus amount is adjusted annually for inflation and is given for the duration of the certification (10 years). An additional $5,000 bonus is awarded to NBPTS-certified teachers who work in high-poverty schools.

D. Pay for Performance

Pay for performance policies link part of teachers' salaries—usually as a bonus in addition to base pay—to measures of their effectiveness. "Effectiveness" is typically measured by gains in student test scores, observations of teaching practices, or both. Teacher performance pay programs tend to face opposition and few have continued beyond a pilot phase, in part due to the complexity of implementation.[16] For example, one challenge is figuring out how to reward teachers who do not teach tested subjects (usually reading and math), such as the arts.

Some programs in the United States have been rigorously evaluated in terms of how teacher bonuses impact student test scores in the short-run. Those programs are summarized in Exhibit 5; most provide a bonus based on multiple measures, including increases in test scores from individual teachers or school-wide. The bonus amount ranges from $1,000 to $15,000 per teacher.

D1. *Impact of Pay for Performance programs*. We located 12 studies, most of them very recent, which examined pay for performance for teachers. The 12 studies are summarized in Exhibit 5.

Exhibit 5

Summary of Teacher Pay for Performance Programs Included in the Research Review

Study	Program name	Location	Bonus only?	Individual or school-wide award basis?	Test scores or other criteria?	Approx. bonus amount
Dee & Keys, 2004	Career Ladder Evaluation System	Tennessee	Bonus + career ladder*	Individual	Principal evaluations	Up to $7000
Figlio & Kenny, 2007	NELS:88 (national survey)	United States	Mixed	Mixed	Varies	Varies
Fryer, 2011	School-wide Performance Bonus Program	New York City	Bonus only	School	Test scores + other	Up to $3,000
Glazerman & Seifullah, 2010	Teacher Advancement Program (TAP)	Chicago	Bonus + career ladder*	Individual	Test scores + other	Up to $6,320
Goodman & Turner, 2010	School-wide Performance Bonus Program	New York City	Bonus only	School	Test scores + other	Up to $3,000
Hudson, 2010	Teacher Advancement Program (TAP)	United States	Bonus + career ladder*	Both	Test scores + other	Up to $3,000
Ladd, 1999	Dallas incentive program	Dallas	Bonus only	School	Test scores + other	$1,000
Marsh et al., 2011	School-wide Performance Bonus Program	New York City	Bonus only	School	Test scores + other	Up to $3,000
Springer & Winters, 2009	School-wide Performance Bonus Program	New York City	Bonus only	School	Test scores + other	Up to $3,000
Springer et al., 2009	Texas Educator Excellence Grant (TEEG) Program	Texas	Bonus only	Usually Individual	Test scores + other	Up to $10,000
Springer et al., 2010	Project on Incentives in Teaching (POINT)	Nashville	Bonus only	Individual	Test scores only	Up to $15,000
Vigdor, 2010	ABCs of Public Education	North Carolina	Bonus only	School	Test scores only	Up to $1,500

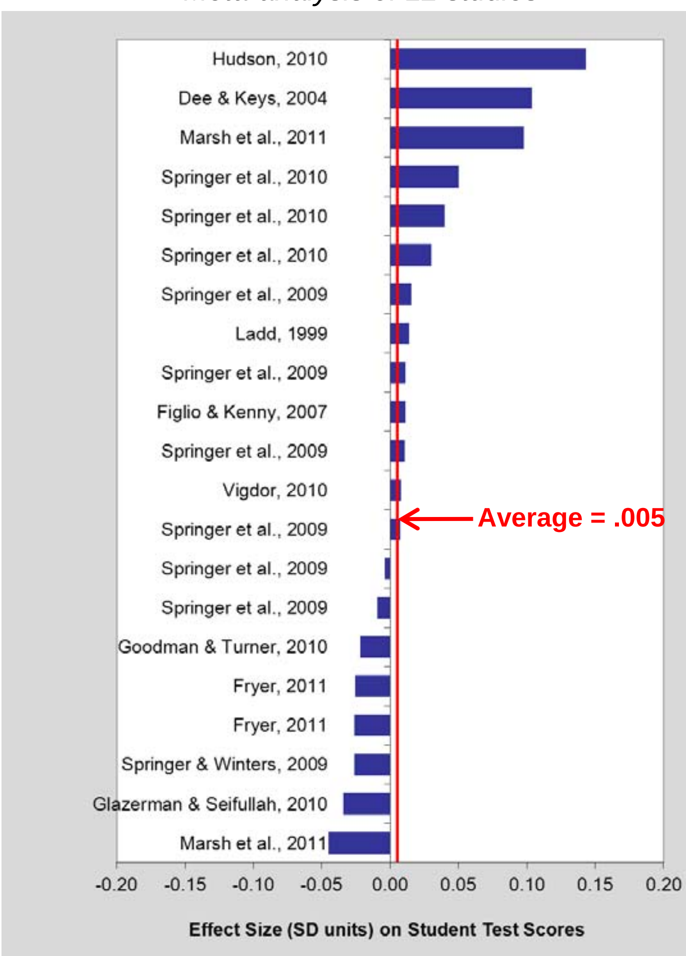
*A career ladder provides salary increases for additional responsibilities and authority (such as mentors or "master" teachers)

[16] Ballou, D. (2001). Pay for performance in public and private schools. *Economics of Education Review 20*, 51-61.

As we saw with our review of the effectiveness of graduate degrees, a few studies of teacher performance pay found positive effects while a few found negative effects. Exhibit 6 reveals a small positive impact on student test scores, on average.

The existing research does not address, unfortunately, the long-term impacts on student achievement from paying teachers for performance. Because the available evidence is limited to pilot programs, we do not yet know the potential impact of a well-established program.

Exhibit 6
Estimates of the Effect of Teacher Pay for Performance Programs on Student Outcomes
Meta-analysis of 12 studies



E. Teacher Induction

Across the United States, schools frequently provide "induction" for new teachers who have no prior classroom experience. School administrators assign a veteran teacher to mentor a novice teacher, offering guidance and support in their first and second years at the school. Some induction programs provide additional support such as professional development, structured peer group interaction, and observation of veteran teachers.

Washington State's Beginning Educator Support Team (BEST) program provides grants to districts and schools to help them implement teacher induction programs. BEST grants were awarded to 28 districts (some as part of a consortium) in the 2011-12 school year.[17]

Exhibit 7 summarizes the programs included in this review.

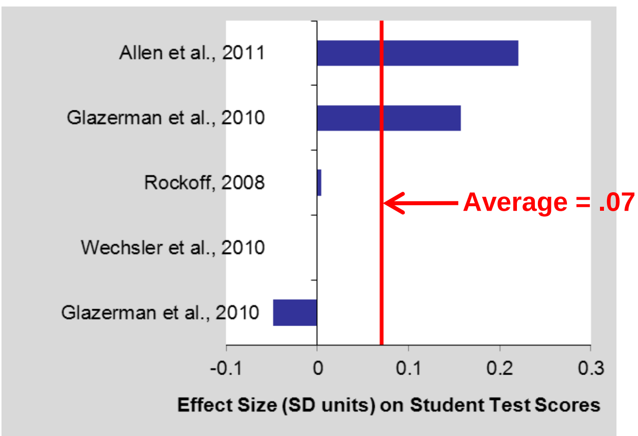
E1. *Impact of Induction Programs*. The four studies summarized in Exhibit 7 (next page) include five results, plotted in Exhibit 8 (next page). Three of these studies compare more intensive programs to "induction-as-usual," because some form of mentoring (often informal) was typically already occurring in the schools studied. The results suggest an overall positive impact on student achievement as measured by test scores, although this finding should be considered preliminary given so few studies.[18]

[17] For more information, visit: http://www.k12.wa.us/BEST/default.aspx.
[18] Many studies of teacher induction programs measure impacts on teacher retention, which is often the primary focus of these programs. We have not meta-analyzed those impacts.

Exhibit 7
Summary of Teacher Induction Programs included in the Research Review

Study	Program name	Location	Mentoring only, or other program components?	Compared with?
Allen et al., 2011	My Teaching Partner-Secondary	Virginia	On-line mentoring + workshops and videos	Induction-as-usual
Glazerman et al., 2010	Two programs (analyzed together): Educational Testing Service (ETC) & New Teacher Center (NTC)	National (U.S.)	Mentoring + professional development, peer group, and observation of veteran teachers	Induction-as-usual
Rockoff, 2008	New Teacher Center (NTC) model	New York City	Mentoring	No mentoring
Wechsler et al., 2010	State-Funded Mentoring and Induction Program	Illinois	Mentoring + professional development, formative assessments	Induction-as-usual

Exhibit 8
Estimates of the Effect of Teacher Induction Programs on Student Outcomes
Meta-analysis of 4 studies



F. Professional Development

In Washington, as in other states, teachers must complete certain professional development (PD) requirements in order to maintain certification and add endorsements.[19] We analyzed research that examines impacts on student test scores from various approaches to teacher PD. The analysis addressed a basic question: what are the potential impacts from putting more resources into teacher training?

The specific approaches studied were diverse. We organized the research literature into two categories: “general” and “content-specific,” broadly defined. Studies of general PD measured training in terms of time (variation in total in-service hours among teachers) or additional PD resources given to struggling schools to be used at the schools’ discretion. Content-specific PD focused on instructional strategies specific to a grade level and subject area. Exhibit 9 (next page) provides a brief summary of the studies included in the analysis. All of the studies examined elementary and/or middle schools.

For both types of PD, we standardized all measured impacts in terms of the effect of an additional day (eight hours) of training. Because teachers typically participate in more than one day of PD per year, the actual impacts were larger than shown in Exhibits 10 and 11 (page 10). Many of the programs studied, particularly for content-specific PD, involved two-week summer institutes with follow-up sessions during the school year.

[19] For more information, visit: http://www.pesb.wa.gov and http://www.k12.wa.us/certification/TeacherMain.aspx.

Exhibit 9

Summary of Teacher Professional Development Programs Included in the Research Review

Study	Location	When provided	Type of PD studied	Content area	Grade level
General PD					
Duffy et al., 1986	Mid-western district	School year	Explicit verbal explanations	General + reading	Elementary
Harris & Sass, 2011	Florida	School year	General in-service hours	General	Elementary & middle
Jacob & Lefgren, 2004	Chicago	School year	At the school's discretion	General	Elementary & middle
McGill-Franzen et al., 1999	Eastern U.S. district	School year	General classroom management & reading instruction	General + reading	Elementary
Sloan, 1993	South Bend, IN	School year	At the school's discretion	General	Elementary
Content-specific PD					
Angrist & Lavy, 2001	Jerusalem, Israel	School year	Mathematics instructional techniques	Math	Elementary
Borman et al., 2008	Los Angeles, CA	Summer + school year	"System-wide Change"	Science	Elementary
Carpenter et al., 1989	Madison, WI	Summer	"Cognitively Guided Instruction"	Math	Elementary
Garet et al., 2008	United States	Summer + school year	"Language Instruction for Teachers of Reading and Spelling"	Reading	Elementary
Garet et al., 2010	United States	Summer + school year	Content knowledge & technique	Math	Middle school
Harris & Sass, 2011	Florida	School year	Content-specific in-service hours	Reading & math	Elementary and middle school
Johnson et al., 2007	Ohio	Summer + school year	"Discovery Model Schools Initiative"	Science	Middle school
Santagata et al., 2011	Urban U.S. district	School year	Researcher-developed training based on high-TIMSS* countries	Math	Middle school
Saxe et al., 2001	Los Angeles, CA	Summer + school year	"Integrated Mathematics Assessment"	Math	Elementary

* TIMSS = Third International Mathematics and Science Study (1999)

The estimates can be scaled up using a simple multiplier (e.g., multiply by ten for a ten-day program).[20] For both categories, the increased time or new approach is compared with professional development as-usual.

F1. *Impact of General Professional Development*. We located and analyzed five studies, yielding eight results that examine the impact of increasing the time or overall resources for teacher PD. Overall, there is no impact on student test scores from providing "more of the same" PD.

F2. *Impact of Content-Specific Professional Development*. We located and analyzed eight results from eight studies of focused, content-specific PD for teachers. These results are positive overall, suggesting that providing more focused PD can improve student learning.

[20] We did not have sufficient information to model any possible diminishing returns from larger increases in PD resources.

Exhibit 10
Estimates of the Effect of an Additional Day of General Professional Development on Student Outcomes *Meta-analysis of 5 studies*

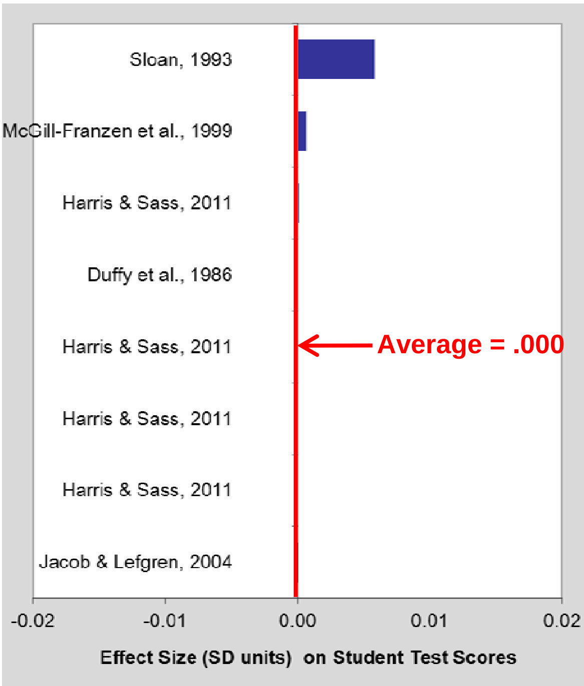
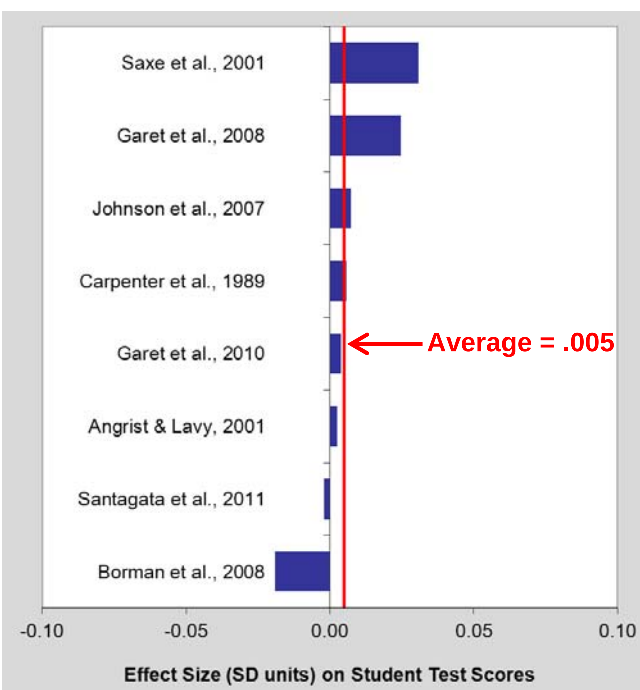


Exhibit 11
Estimates of the Effect of an Additional Day of Content-Specific Professional Development on Student Outcomes *Meta-analysis of 8 studies*



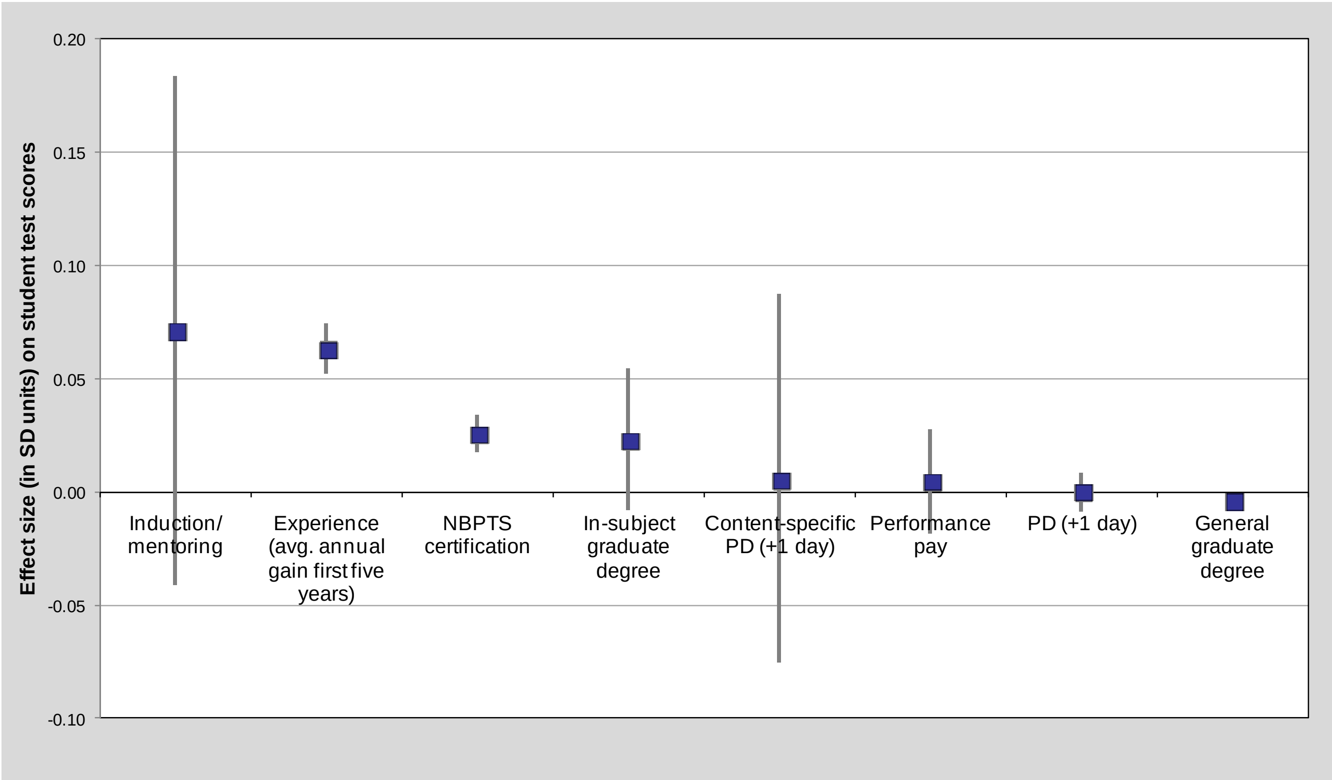
Summary

The research summarized in this report reveals that some of Washington State's existing policies regarding teacher compensation and training—such as paying for additional years of experience, NBPTS certification, and teacher induction—are, at least roughly, aligned with the evidence regarding effectiveness. The research also suggests that creating financial incentives for teachers to obtain general graduate training and professional development is not associated with improvements in student test scores.

We did find evidence that more focused training, such as in-subject master's degrees and content-specific professional development, can improve student outcomes. Exhibit 12 (next page) summarizes the results across the six topics. The error bands around the estimates indicate the precision of each estimate; statistically significant impacts have error bands that do not cross the zero line. The findings regarding teaching experience and NBPTS certification are the strongest.

For this report, we have not estimated the benefits and costs of these findings. We will calculate benefits and costs for these topics prior to the 2013 legislative session.

Exhibit 12
Summary of Meta-Analytic Findings Regarding Impacts on Student Test Scores from Different Policies Related to Teacher Compensation and Training



Studies Used in the Meta-analyses

Years of Teaching Experience

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics, 25*(1), 95-135.

Akerhielm, K. (1995). Does class size matter? *Economics of Education Review, 14*(3), 229-241.

Archibald, S. (2006). Narrowing in on educational resources that do affect student achievement. *Peabody Journal of Education, 81*(4), 23-42.

Borland, M. V., Howsen, R. M., & Trawick, M. W. (2005). An investigation of the effect of class size on student academic achievement. *Education Economics, 13*(1), 73-83.

Boyd, D., Lankford, H., Loeb, S., Rockoff, J., & Wyckoff, J. (2008). The narrowing gap in New York City teacher qualifications and its implications for student achievement in high-poverty schools. *Journal of Policy Analysis & Management*, 27(4), 793-818.

Brown, B. W., & Saks, D. H. (1975). The production and distribution of cognitive skills within schools. *Journal of Political Economy, 83*(3), 571-593.

Buddin, R., & Zamarro, G. (2009). Teacher qualifications and student achievement in urban elementary schools. *Journal of Urban Economics 66*(2), 103-115.

Chingos, M. M., & Peterson, P. E. (2011). It's easier to pick a good teacher than to train one: Familiar and new results on the correlates of teacher effectiveness. *Economics of Education Review, 30*(3), 449-465.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *The Journal of Human Resources, 41*(4), 778-820.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review, 26*(6), 673-682.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2010). Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects. *Journal of Human Resources, 45*(3), 655-681.

Corcoran, S. P., Jennings, J. L., & Beveridge, A. A. (2011, April). *Teacher effectiveness on high- and low-stakes tests*. Unpublished manuscript, New York University.

Croninger, R. G., Rice, J. K., Rathbun, A., & Nishio, M. (2007). Teacher qualifications and early learning: Effects of certification, degree, and experience on first-grade student achievement. *Economics of Education Review, 26*(3), 312-324.

Goldhaber, D. D., & Brewer, D. J. (1996). Evaluating the effect of teacher degree level on educational performance. In W. J. Fowler, Jr. (Ed.), *Developments in school finance, 1996: Fiscal proceedings from the Annual NCES State Data Conference* (pp. 197-210). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Goldhaber, D. D., & Brewer, D. J. (1997). Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *The Journal of Human Resources, 32*(3), 505-523.

Goldhaber, D. D., & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis, 22*(2), 129-145.

Goldhaber, D. D., Brewer, D. J., & Anderson, D. J. (1999). A three-way components analysis of educational productivity. *Education Economics, 7*(3), 199-208.

Goldhaber, D., & Anthony, E. (2007). Can teacher quality be effectively assessed? National board certification as a signal of effective teaching. *The Review of Economics and Statistics, 89*(1), 134-150.

Goldhaber, D., Liddle, S., Theobald, R., & Walch, J. (2010). *Teacher effectiveness and the achievement of Washington's Students in Mathematics* (CEDR Working Paper 2010-06). Bothell: University of Washington Bothell, Center for Education Data & Research.

Hanushek, E. A. (1992). The trade-off between child quantity and quality. *Journal of Political Economy, 100*(1), 84-117.

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics, 95*(7-8), 798-812.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371-406.

Huang, F. L., & Moon, T. R. (2009). Is experience the best teacher? A multilevel analysis of teacher characteristics and student achievement in low performing schools. *Educational Assessment, Evaluation and Accountability, 21*(3), 209-234.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*(1), 101-136.

Jepsen, C., & Rivkin, S. (2002). *Class size reduction, teacher quality, and academic achievement in California public elementary schools*. San Francisco: Public Policy Institute of California.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review, 27*(6), 615-631.

Koedel, C., & Betts, J. R. (2007). *Re-examining the role of teacher quality in the educational production function*. Unpublished manuscript, University of Missouri-Columbia, Department of Economics.

Krieg, J. M. (2006). Teacher quality and attrition. *Economics of Education Review, 25*(1), 13-27.

Krueger, A. B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics, 114*(2), 497-532.

Kukla-Acevedo, S. (2009). Do teacher characteristics matter? New results on the effects of teacher preparation on student achievement. *Economics of Education Review, 28*(1), 49-57.
Ladd, H. F., Sass, T. R., & Harris, D. N. (2007, February). *The impact of national board certified teachers on student achievement in Florida and North Carolina: A summary of the evidence prepared for the National Academies Committee on the evaluation of the impact of teacher certification by NBPTS*. Unpublished manuscript.
Leak, J. A., & Farkas, G. (2011). *Effects of teacher credentials, coursework, and certification on student achievement in math and reading in kindergarten: An ECLS-K study*. Evanston, IL: Society for Research on Educational Effectiveness.
Leigh, A. K. (2010). Estimating teacher effectiveness from two-year changes in students' test scores. *Economics of Education Review, 29*(3), 480-488.
Ost, B. (2009, November). *How do teachers improve? The relative importance of specific and general human capital*. Unpublished manuscript, Cornell University, Ithaca, NY.
Pil, F. K., & Leana, C. (2009). Applying organizational research to public school reform: The effects of teacher human and social capital on student performance. *Academy of Management Journal, 52*(6), 1101-1124.
Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review, 94*(2), 247-252.
Subedi, B. R., Swan, B., & Hynes, M. C. (2011). Are school factors important for measuring teacher effectiveness? A multilevel technique to predict student gains through a value-added approach. *Education Research International, 2011*. doi: 10.1155/2011/532737
Xu, Z., Hannaway, J., & Taylor, C. (2009, March). *Making a difference? The effects of Teach for America in high school* (Working Paper 17. Revised). Washington, DC: The Urban Institute, National Center for Analysis of Longitudinal Data in Education Research.

Graduate Degrees

Archibald, S. (2006). Narrowing in on educational resources that do affect student achievement. *Peabody Journal of Education, 81*(4), 23-42.
Buddin, R., & Zamarro, G. (2009, June). *Teacher qualifications and middle school student achievement* (Working Paper WR-671-IES). Santa Monica, CA: RAND.
Buddin, R., & Zamarro, G. (2009). Teacher qualifications and student achievement in urban elementary schools. *Journal of Urban Economics 66*(2), 103-115.
Cavalluzzo, L. C. (2004, November). *Is national board certification an effective signal of teacher quality?* Alexandria, VA: The CNA Corporation.
Chingos, M. M., & Peterson, P. E. (2011). It's easier to pick a good teacher than to train one: Familiar and new results on the correlates of teacher effectiveness. *Economics of Education Review, 30*(3), 449-465.
Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *The Journal of Human Resources, 41*(4), 778-820.
Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review, 26*(6), 673-682.
Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2010). Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects. *Journal of Human Resources, 45*(3), 655-681.
Corcoran, S. P., Jennings, J. L., & Beveridge, A. A. (2011, April). *Teacher effectiveness on high- and low-stakes tests*. Unpublished manuscript, New York University.
Croninger, R. G., Rice, J. K., Rathbun, A., & Nishio, M. (2007). Teacher qualifications and early learning: Effects of certification, degree, and experience on first-grade student achievement. *Economics of Education Review, 26*(3), 312-324.
Goldhaber, D., & Anthony, E. (2007). Can teacher quality be effectively assessed? National board certification as a signal of effective teaching. *The Review of Economics and Statistics, 89*(1), 134-150.
Goldhaber, D. D., & Brewer, D. J. (1996). Evaluating the effect of teacher degree level on educational performance. In W. J. Fowler, Jr. (Ed.), *Developments in school finance, 1996: Fiscal proceedings from the Annual NCES State Data Conference* (pp. 197-210). Washington, DC: U.S. Department of Education, National Center for Education Statistics.
Goldhaber, D. D., & Brewer, D. J. (1997). Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *The Journal of Human Resources, 32*(3), 505-523.
Goldhaber, D. D., & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis, 22*(2), 129-145.
Goldhaber, D. D., Brewer, D. J., & Anderson, D. J. (1999). A three-way components analysis of educational productivity. *Education Economics, 7*(3), 199-208.
Hanushek, E. A. (1992). The trade-off between child quantity and quality. *Journal of Political Economy, 100*(1), 84-117.
Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics, 95*(7-8), 798-812.
Huang, F. L., & Moon, T. R. (2009). Is experience the best teacher? A multilevel analysis of teacher characteristics and student achievement in low performing schools. *Educational Assessment, Evaluation and Accountability, 21*(3), 209-234.
Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*(1), 101-136.
Koedel, C., & Betts, J. R. (2007, April). *Re-examining the role of teacher quality in the educational production function*. Unpublished manuscript, University of Missouri-Columbia, Department of Economics.

Krieg, J. M. (2006). Teacher quality and attrition. *Economics of Education Review, 25*(1), 13-27.
Krueger, A. B. (1999). Experimental estimates of education production functions. *The Quarterly Journal of Economics, 114*(2), 497-532.
Ladd, H. F., Sass, T. R., & Harris, D. N. (2007, February). *The impact of national board certified teachers on student achievement in Florida and North Carolina: A summary of the evidence prepared for the National Academies Committee on the evaluation of the impact of teacher certification by NBPTS*. Unpublished manuscript.
Leak, J. A., & Farkas, G. (2011). *Effects of teacher credentials, coursework, and certification on student achievement in math and reading in kindergarten: An ECLS-K study*. Evanston, IL: Society for Research on Educational Effectiveness.
Leigh, A. K. (2010). Estimating teacher effectiveness from two-year changes in students' test scores. *Economics of Education Review, 29*(3), 480-488.
Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy, 6*(1), 43-74.

In-Subject Graduate Degrees

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics, 25*(1), 95-135.
Croninger, R. G., Rice, J. K., Rathbun, A., & Nishio, M. (2007). Teacher qualifications and early learning: Effects of certification, degree, and experience on first-grade student achievement. *Economics of Education Review, 26*(3), 312-324.
Goldhaber, D. D., & Brewer, D. J. (1997). Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *The Journal of Human Resources, 32*(3), 505-523.
Goldhaber, D. D., & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis, 22*(2), 129-145.
Dee, T. S., & Cohodes, S. R. (2008). Out-of-field teachers and student achievement: Evidence from matched-pairs comparisons. *Public Finance Review, 36*(1), 7-32.
Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy, 6*(1), 43-74.
Subedi, B. R., Swan, B., & Hynes, M. C. (2011). Are school factors important for measuring teacher effectiveness? A multilevel technique to predict student gains through a value-added approach. *Education Research International.*

NBPTS Certification

Cantrell, S., Fullerton, J., Kane, T. J., & Staiger, D. O. (2008). *National board certification and teacher effectiveness: Evidence from a random assignment experiment* (Working Paper No. 14608). Cambridge: NBER.
Cavalluzzo, L. C. (2004). *Is national board certification an effective signal of teacher quality*? Alexandria, VA: The CNA Corporation.
Chingos, M. M., & Peterson, P. E. (2011). It's easier to pick a good teacher than to train one: Familiar and new results on the correlates of teacher effectiveness. *Economics of Education Review*, 30, 3, 449-465
Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *The Journal of Human Resources*, 41(4), 778-820.
Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, 26, 6, 673-682.
Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2010). Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects. *Journal of Human Resources*, 45(3): 655-681.
Goldhaber, D., & Anthony, E. (2007). Can teacher quality be effectively assessed? National board certification as a signal of effective teaching. *The Review of Economics and Statistics*, 89(1), 134-150.
Harris, D. N., & Sass, T. R. (2007*). The effects of NBPTS-certified teachers on student achievement* (Working Paper 4). Washington, DC: The Urban Institute, National Center for Analysis of Longitudinal Data in Education Rsch..
Ladd, H. F., Sass, T. R., & Harris, D. N. (2007). The impact of national board certified teachers on student achievement in Florida and North Carolina: A summary of the evidence. *National Academies Committee on the evaluation of the impact of teacher certification by NBPTS.* Unpublished manuscript.
Sanders, W. L., Ashton, J. J., Wright, S. P. (2005). *Comparison of the effects of NBPTS certified teachers with other teachers on the rate of student academic progress* (Final report). Retrieved from ERIC database.
Stronge, J. H., Ward, T. J., Tucker, P. D., Hindman, J. L., McColsky, W., & Howard, B. (2007). National Board certified teachers and non-national board certified teachers: Is there a difference in teacher effectiveness and student achievement? *Journal of Personnel Evaluation in Education*, 20(3-4), 185-210.
Vandevoort, L. G., Amrein-Beardsley, A., & Berliner, D. C. (2004). National Board Certified teachers and their students' achievement. *Education Policy Analysis Archives,* 12(46).

Pay for Performance

Dee, T. S., & Keys, B. J. (2004). Does merit pay reward good teachers? Evidence from a randomized experiment. *Journal of Policy Analysis and Management,* 23(3), 471-488.
Figlio, D. N., & Kenny, L. W. (2007). Individual teacher incentives and student performance. *Journal of Public Economics*, 91(5-6), 901-914.
Fryer, R. G. (2011). *Teacher incentives and student achievement: Evidence from New York City public schools* (Working Paper No. 16850). Cambridge: National Bureau of Economic Research.
Glazerman, S., Seifullah, A. (2010). *An evaluation of the Teacher Advancement Program (TAP) in Chicago: Year two impact report*. Washington, DC: Mathematica Policy Research.
Goodman, S., & Turner, L. (2010). *Teacher incentive pay and educational outcomes: Evidence from the NYC Bonus Program*. Unpublished manuscript, Columbia University, New York.
Hudson, S. (2010). *The effects of performance-based teacher pay on student achievement.* Discussion Paper for the Stanford Institute for Economic Policy Research, Stanford University. Retrieved from: http://www.stanford.edu/group/siepr/cgi-bin/siepr/?q=system/files/shared/pubs/papers/09-023_Paper_Hudson.pdf
Ladd, H. F. (1999). The Dallas school accountability and incentive program: An evaluation of its impacts on student outcomes. *Economics of Education Review,* 18(1), 1-16.
Marsh, J. A., Springer, M. G., & McCaffrey, D. F. (2011). *A Big Apple for Educators: Final Evaluation Report*. Santa Monica: RAND Corp.
Springer, M. and Winters, M. (2009). *New York City's school-wide bonus pay program: Early evidence from a randomized trial* (Working Paper 2009-02). Nashville, TN: Vanderbilt Peabody College, National Center on Performance Incentives.
Springer, M. G., Ballou, D., Hamilton, L., Le, V.-N., Lockwood, J. R., McCaffrey, D. F., . . . Stecher, B. M. (2010). *Teacher pay for performance: Experimental evidence from the Project on Incentives in Teaching.* Nashville, TN: Vanderbilt University Peabody College, National Center on Performance Incentives.
Springer, M., et al. (2009). *Texas Educator Excellence Grant (TEEG) Program: Year three evaluation report.* National Center on Performance Incentives, Vanderbilt University, Nashville.
Vigdor, J. L. (2008). *Teacher salary bonuses in North Carolina* (Working Paper 15). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research.

Teacher Induction

Allen, J. P., Mikami, A. Y., Pianta, R. C., Gregory, A., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science*, 333(6045), 1034-1037.
Glazerman, S., Isenberg, E., Dolfin, S., Bleeker, M., Johnson, A., Grider, M., . . . Ali, M. (2010). *Impacts of comprehensive teacher induction: Final results from a randomized controlled study.* Washington, DC: National Center for Education Evaluation and Regional Assistance.
Rockoff, J. E. (2008). *Does mentoring reduce turnover and improve skills of new employees?: Evidence from teachers in New York City (*Working Paper No. 13868). Cambridge: National Bureau of Economic Research.
Wechsler, M. E., Caspary, K., Humphrey, D. C., & Matsko, K. K. (2010). *Examining the effects of new teacher induction*. Menlo Park, CA: SRI International.

General Professional Development

Angrist, J. D., & Lavy, V. (2001). Does teacher training affect pupil learning? Evidence from matched comparisons in Jerusalem public schools. *Journal of Labor Economics*, 19(2), 343-369.
Duffy, G. G., Roehler, L. R., Meloth, M. S., Vavrus, L. G., Book, C., Putnam, J., & Wesselman, R. (1986). The relationship between explicit verbal explanations during reading skill instruction and student awareness and achievement: A study of reading teacher effects*. Reading Research Quarterly*, 21(3), 237-252.
Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, 95(7-8), 798-812.
Jacob, B. A., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *The Review of Economics and Statistics*, 86(1), 226-244.
McGill-Franzen, A., Allington, R. L., Yokoi, L., & Brooks, G. (1999). Putting books in the classroom seems necessary but not sufficient*. The Journal of Educational Research*, 93(2), 67-74.
Sloan, H. A. (1993). Direct instruction in fourth and fifth grade classrooms. Dissertation Abstracts International, 54(08), 2837A.

Content-specific Professional Development

Borman, G. D., Gamoran, A., & Bowdon, J. (2008). A randomized trial of teacher development in elementary science: First-year achievement effects. *Journal of Research on Educational Effectiveness*, 1(4), 237-264.
Carpenter, T. P., Fennema, E., Peterson, P. L., Chiang, C. P., & Loef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: An experimental study. *American Educational Research Journal*, 26(4), 499-531.
Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., . . . Silverberg, M. (2008). The impact of two professional development interventions on early reading instruction and achievement. Washington, DC: National Center for Educational Evaluation and Regional Assistance.

Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Walters, K., Song, M., . . . Warner, E. (2010). Middle school mathematics professional development impact study: Findings after the first year of implementation. Washington, DC: National Center for Education Evaluation and Regional Assistance.

Johnson, C. C., Kahle, J. B., & Fargo, J. D. (2007). A study of the effect of sustained, whole-school professional development on student achievement in science. *Journal of Research in Science Teaching*, 44(6), 775-786.

Santagata, R., Kersting, N., Givvin, K. B., & Stigler, J. W. (2011). Problem implementation as a lever for change: An experimental study of the effects of a professional development program on students' mathematics learning*. Journal of Research on Educational Effectiveness*, 4(1): 1-24.

Saxe, G., Gearhart, M., & Nasir, N. (2001). Enhancing students' understanding of mathematics: A study of three contrasting approaches to professional support. *Journal of Mathematics Teacher Educat*ion, 4(1), 55-79.

Technical Appendix: Meta-Analytic Procedures

To estimate the effects of programs and policies on outcomes, we use statistical procedures researchers have been developing to facilitate systematic reviews of evaluation evidence. This set of procedures is called “meta-analysis” and we employ that methodology in this study.[21]

STUDY SELECTION AND CODING CRITERIA

A meta-analysis is only as good as the selection and coding criteria used to conduct the study.[22] Following are the key choices we made and implemented.

Study Selection. We used four primary means to locate studies for meta-analysis of programs: (1) we consulted the bibliographies of systematic and narrative reviews of the research literature in the various topic areas; (2) we examined the citations in the individual studies themselves; (3) we conducted independent literature searches of research databases using search engines such as Google, Proquest, Ebsco, ERIC, PubMed, and SAGE; and (4) we contacted authors of primary research to learn about ongoing or unpublished evaluation work. After first identifying all possible studies via these search methods, we attempted to determine whether the study was an outcome evaluation that had a valid comparison group. If a study met this criterion, we then secured a paper copy of the study for our review.

Peer-Reviewed and Other Studies. We examined all evaluation studies we could locate with these search procedures. Many studies were published in peer-reviewed academic journals while others were from reports obtained from the agencies themselves. It is important to include non-peer reviewed studies, because it has been suggested that peer-reviewed publications may be biased to show positive program effects. Therefore, our meta-analysis includes all available studies that meet our other criteria, regardless of publication source.

Control and Comparison Group Studies. Our analysis only includes studies that had a control or comparison group or used a quasi-experimental design such as regression discontinuity with multiple, sophisticated controls. We do not include studies with a single-group, pre-post research design. This choice was made because it is only through rigorous studies that causal relationships can be reliably estimated.

Random Assignment and Quasi-Experiments. Random assignment studies were preferred for inclusion in our review, but we also included non-randomly assigned comparison groups. We only included quasi-experimental studies if sufficient information was provided to demonstrate comparability between the treatment and comparison groups on important pre-existing conditions such as age, gender, and pre-treatment characteristics such as test scores.

Enough Information to Calculate an Effect Size. Following the statistical procedures in Lipsey and Wilson,[23] a study had to provide the necessary information to calculate an effect size. If the necessary information was not provided, and we were unable to obtain the necessary information directly from the study author(s), the study was not included in our review.

Mean-Difference Effect Sizes. For this study, we coded mean-difference effect sizes for continuous measures following the procedures in Lipsey and Wilson.[24] For dichotomous measures, we used the D-cox transformation to approximate the mean difference effect size, as described in Sánchez-Meca, Marín-Martínez, and Chacón-Moscoso.[25] We chose to use the mean-difference effect size rather than the odds ratio effect size because we frequently coded both dichotomous and continuous outcomes (odds ratio effect sizes could also have been used with appropriate transformations).

Outcome Measures of Interest. Our primary outcomes of interest include standardized, validated assessments of student learning. Most of the studies control for students’ prior test scores using a value-added model. Most studies report reading and/or math outcomes.

Averaging Effect Sizes for Similar Outcomes. If both reading and math, or other subjects, were measured, we meta-analyzed the similar measures and used the combined effect size in the meta-analysis for that program. As a result, each study sample coded in this analysis is associated with a single effect size for a given outcome.

---

[21] In general, we follow the meta-analytic methods described in: Lipsey, M. W., & Wilson, D. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications.
[22] All studies used in the meta-analysis are identified in the references to this paper. Many other studies were reviewed, but did not meet the criteria set for this analysis.
[23] Lipsey & Wilson, 2001.
[24] Ibid.
[25] Sánchez-Meca, J., Marín-Martínez, F., & Chacón-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods*, *8*(4), 448-467.

PROCEDURES FOR CALCULATING EFFECT SIZES

Effect sizes summarize the degree to which a program or policy affects an outcome. In experimental settings this involves comparing the outcomes of treated participants relative to untreated participants. There are several methods used by analysts to calculate effect sizes, as described in Lipsey and Wilson.[26] The most common effect size statistic is the standardized mean difference effect size, and that is the measure we used in this analysis.

Weighted Mean Different Effect Size. The mean difference effect size was designed to accommodate continuous outcome data, such as student test scores, where the differences are in the means of the outcome.[27] The standardized mean difference effect size is computed with:

$$(1)\ ES = \frac{M_t - M_c}{\sqrt{\frac{(N_t - 1)SD_t^2 + (N_c - 1)SD_c^2}{N_t + N_c - 2}}}$$

In this formula, ES is the estimated effect size for a particular program; M_t is the mean value of an outcome for the treatment or experimental group; M_c is the mean value of an outcome for the control group; SD_t is the standard deviation of the treatment group; and SD_c is the standard deviation of the control group; N_t is the number of subjects in the treatment group; and N_c is the number of subjects in the control group. The variance of the mean difference effect size statistic in (1) is computed with:[28]

$$(2)\ ESVar = \frac{N_t + N_c}{N_t N_c} + \frac{ES^2}{2(N_t + N_c)}$$

In some random assignment studies or studies where treatment and comparison groups are well-matched, authors provide only statistical results from a t-test. In those cases, we calculate the mean difference effect size using:[29]

$$(3)\ ES = t\sqrt{\frac{N_t + N_c}{N_t N_c}}$$

In many research studies, the numerator in (1), $M_t - M_c$, is obtained from a coefficient in a regression equation, not from experimental studies of separate treatment and control groups. For such studies, the denominator in (1) is the standard deviation for the entire sample. In these types of regression studies, unless information is presented that allows the number of subjects in the treatment condition to be separated from the total number in a regression analysis, the total N from the regression is used for the sum of N_t and N_c, and the product term $N_t N_c$ is set to equal $(N/2)^2$.

Pre/Post Measures. Where authors report pre- and post-treatment measures without other statistical adjustments, first we calculate two between-groups effect sizes: (1) at pre-treatment and, (2) at post-treatment. Finally, we calculate the overall effect size by subtracting the post-treatment effect size from the pre-treatment effect size.

ADJUSTING EFFECT SIZES FOR SMALL SAMPLE SIZES

Since some studies have very small sample sizes, we follow the recommendation of many meta-analysts and adjust for this. Small sample sizes have been shown to upwardly bias effect sizes, especially when samples are less than 20. Following Hedges,[30] Lipsey and Wilson[31] report the “Hedges correction factor,” which we use to adjust all mean-difference effect sizes, (where N is the total sample size of the combined treatment and comparison groups):

$$(4)\ ES'_m = \left[1 - \frac{3}{4N - 9}\right] * \ ES_m$$

Adjusting Effect Sizes and Variances for Multi-Level Data Structures. Most studies in the education field use data that are hierarchical in nature. That is, students are clustered in classrooms, classrooms are clustered within schools, schools are clustered within districts, and districts are clustered within states. Analyses that do not account for clustering will underestimate the variance in outcomes at the student level (the denominator in equation 1 and, thus, may over-estimate the precision of

[26] Lipsey & Wilson, 2001.
[27] Ibid, Table B10, equation 1, p. 198.
[28] Ibid, Table 3.2, p. 72.
[29] Ibid, Table B10, equation 2, p. 198
[30] Hedges, L. V. (1981). Distribution theory for Glass’s estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107-128.
[31] Lipsey & Wilson, 2001, equation 3.22, p. 49.

magnitude on effect sizes. In studies that do not account for clustering, effect sizes and their variance require additional adjustments.[32] There are two types of studies, each requiring a different set of adjustments.[33] First, for student-level studies that ignore the variance due to clustering, we make adjustments to the mean effect size and its variance,

$$(5)\ ES_T = ES_m * \sqrt{1 - \frac{2(n-1)\rho}{N-2}}$$

$$(6)\ V\{ES_T\} = \left(\frac{N_t + N_c}{N_t N_c}\right)[1 + (n-1)\rho] + ES_T^{\,2}\left(\frac{(N-2)(1-\rho)^2 + n(N-2n)\rho^2 + 2(N-2n)\rho(1-\rho)}{2(N-2)[(N-2) - 2(n-1)\rho]}\right)$$

where ρ is the intraclass correlation, the ratio of the variance between clusters to the total variance; *N* is the total number of individuals in the treatment group, N_t, and the comparison group, N_c; and *n* is the average number of persons in a cluster, K. In the educational field, clusters can be classes, schools, or districts. For this study, we used 2006 Washington Assessment of Student Learning (WASL) data to calculate values of ρ for the school-level (ρ = 0.114) and the district level (ρ = 0.052). Class-level data were not available, so we use a value of ρ = 0.200 for class-level studies.

Second, for studies that report means and standard deviations at a cluster level, we make adjustments to the mean effect size and its variance:

$$(7)\ ES_T = ES_m * \sqrt{\frac{1 + (n-1)\rho}{n\rho}} * \sqrt{\rho}$$

$$(8)\ v\{ES_T\} = \left\{\left(\frac{N_t - N_c}{N_t N_c}\right) * \left(\frac{1 + (n-1)\rho}{n\rho}\right) + \frac{[1 + (n-1)\rho]^2 * ES_T^{\,2}}{2n\rho(K-2)}\right\} * \rho$$

We did not adjust effect sizes in studies reporting dichotomous outcomes. This is because the Cox transformation assumes the entire normal di**s**tribution at the student level.[34] However, when outcomes are dichotomous, or an effect size is calculated from studies where authors control for clustering with robust standard errors or hierarchical linear modeling, we use the "design effect" to calculate the "effective sample size".[35] The design effect is given by:

$$(9)\ D = 1 + (n-1)\rho$$

And the effective sample size is the actual sample size divided by the design effect. For example the effective sample size for the treatment group is:

$$(10)\ N_{t(eff)} = \frac{N_t}{D}$$

Computing Weighted Average Effect Sizes, Confidence Intervals, and Homogeneity Tests. Once effect sizes are calculated for each program effect, and any necessary adjustments for clustering are made, the individual measures are summed to produce a weighted average effect size for a program area. We calculate the inverse variance weight for each program effect and these weights are used to compute the average. These calculations involve three steps. First, the standard error, SE_T of each mean effect size is computed with:[36]

$$(11)\ SE_T = \sqrt{\frac{N_t + N_c}{N_t N_c} + \frac{ES^2}{2(N_t + N_c)}}$$

Next, the inverse variance weight w is computed for each mean effect size with:[37]

[32] Studies that employ hierarchical linear modeling, or fixed effects with robust standard errors, or random effects models account for variance and need no further adjustment for computing the effect size, but adjustments are made to the inverse variance weights for meta-analysis using these methods.
[33] These formulas are taken from: Hedges, L. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics, 32*(4), 341-370.
[34] Mark Lipsey (personal communication, November 11, 2007).
[35] Formulas for design effect and effective sample size were obtained from the Cochrane Reviewers Handbook, section 16.3.4, Approximate analyses of cluster-randomized trials for a meta-analysis: effective sample sizes. http://www.cochrane-handbook.org/
[36] Lipsey & Wilson, 2001, equation 3.23, p. 49.
[37] Ibid., equation 3.24, p. 49.

$$(12)\ w = \frac{1}{SE_T^2}$$

The weighted mean effect size for a group with i studies is computed with:[38]

$$(13)\ \overline{ES} = \frac{\sum(w_i ES_{T_i})}{\sum w_i}$$

Confidence intervals around this mean are then computed by first calculating the standard error of the mean with:[39]

$$(14)\ SE_{\overline{ES}} = \sqrt{\frac{1}{\sum w_i}}$$

Next, the lower, ES_L, and upper limits, ES_U, of the confidence interval are computed with:[40]

$$(15)\ \overline{ES_L} = \overline{ES} - z_{(1-\alpha)}\left(SE_{\overline{ES}}\right)$$

$$(16)\ \overline{ES_U} = \overline{ES} + z_{(1-\alpha)}\left(SE_{\overline{ES}}\right)$$

In equations (B18) and (B19), $z_{(1-\alpha)}$ is the critical value for the z-distribution (1.96 for α = .05). The test for homogeneity, which provides a measure of the dispersion of the effect sizes around their mean, is given by:[41]

$$(17)\ Q_i = \left(\sum w_i\, ES_i^2\right) - \frac{\left(\sum w_i\, ES_i^2\right)}{\sum w_i}$$

The Q-test is distributed as a chi-square with k-1 degrees of freedom (where k is the number of effect sizes).

Computing Random Effects Weighted Average Effect Sizes and Confidence Intervals. Next, a random effects model is used to calculate the weighted average effect size. Random effects models allow us to account for between-study variance in addition to within-study variance.[42] This is accomplished by first calculating the random effects variance component, v[43]

$$(18)\ v = \frac{Q_i - (k-1)}{\sum w_i - \left(\sum wsq_i / \sum w_i\right)}$$

where wsq_i is the square of the weight of ES_i (B15). This random variance factor is then added to the variance of each effect size and finally all inverse variance weights are recomputed, as are the other meta-analytic test statistics. If the value of Q is less than the degrees of freedom (*k-1*), there is no excess variation between studies and the initial variance estimate is used.

[38] Ibid., p. 114
[39] Ibid.
[40] Ibid.
[41] Ibid., p. 116
[42] Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*(2), 97-111.
[43] Ibid., p. 134

For further information, contact Annie Pennucci at
pennuccia@wsipp.wa.gov

Document No. 12-05-2201