What Works Clearinghouse™

ies INSTITUTE OF EDUCATION SCIENCES

October 2012

WWC Review of the Report "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood"[1]

The findings from this review do not reflect the full body of research evidence on teacher value-added or teacher effectiveness.

What is this study about?

The study examined differences in student achievement when students were taught by a teacher with high or low value-added, a measure of teacher effectiveness. See the blue box for more details on teacher value-added.

The study sample included about 3,300 cohorts of math and reading students in grades 4–8 in a large, urban school district. A cohort consists of the students within a given school, grade, and subject area for each year from 1991 to 2009.

This What Works Clearinghouse (WWC) review focuses on the report's analyses of what happened to students' achievement as a result of the movements of teachers with relatively high or low value-added estimates. In particular, these analyses examined what happened to average student achievement in a cohort when one of four events took place:

1. a high value-added teacher started teaching students in the cohort;[2]
2. a high value-added teacher stopped teaching students in the cohort;
3. a low value-added teacher started teaching students in the cohort; or
4. a low value-added teacher stopped teaching students in the cohort.

Cohorts that did not experience one of these events served as the comparison group.

Features of Teacher Value-Added In This Study

Value-added is a measure of teachers' contributions to the achievement growth of their students.

In this study, it was estimated by first predicting each student's performance based on the performance of other students with similar prior achievement and characteristics. Students' actual performance was compared to their predicted performance, and these differences were averaged over all of a teacher's students. This average difference was then compared across all teachers in the district within a particular subject, resulting in a ranking or distribution of teachers.

For each subject, teachers with value-added estimates in the top 5% of this distribution were designated "high value-added" teachers, while teachers in the bottom 5% of this distribution were designated "low value-added" teachers.

A stylized example helps illustrate this concept. Suppose that, on average, students of teacher A scored 15 points higher than they were predicted to score, and students of teacher B scored 2 points lower than they were predicted to score. The value added estimate of teacher A (+15) exceeds that of teacher B (–2) by 17 points.

In this study, the movement of teachers into and out of schools and classrooms was not a result of any policy, program, or intervention; it occurred naturally.

What did the study find?

Study authors reported that when a high value-added teacher started to teach students in a cohort, or a low value-added teacher stopped teaching students in a cohort—two events that are hypothesized to increase student achievement—the event was associated with a statistically significant increase in reading and math test scores.

In addition, when a high value-added teacher stopped teaching students in a cohort, or a low value-added teacher started teaching students in a cohort—two events that are hypothesized to decrease student achievement—the event was associated with a statistically significant decrease in test scores.

WWC Rating

The research described in this report meets WWC evidence standards with reservations

Strengths: The cohorts of students being compared were well matched on test scores before an event occurred (e.g., a high value-added teacher started teaching).

Cautions: Most of the analyses in this study examined how different levels of teacher value-added correspond to students' long-term outcomes, such as college attendance and employment outcomes. However, these analyses are not eligible for WWC review because there are no distinct intervention and comparison groups.

Appendix A: Study details

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood* (NBER Working Paper 17699). Cambridge, MA: National Bureau of Economic Research.

Setting

The study was conducted in one large urban school district in the United States.

Study sample

The sample included 3,202 to 3,384 cohorts (depending on the analysis) of math and reading students in grades 4–8 from 1991 to 2009. A cohort consists of the students within a given school, grade, subject area, and year.

Intervention group

Teacher value-added in this study was estimated by first predicting each student's performance based on the performance of other students with similar prior achievement and characteristics. Students' actual performance was compared to their predicted performance, and these differences were averaged over all of a teacher's students. This average difference was then compared across all teachers in the district within a particular subject, resulting in a ranking or distribution of teachers.

For each subject, teachers with value-added estimates in the top 5% of this distribution were designated "high value-added" teachers, while teachers in the bottom 5% of this distribution were designated "low value-added" teachers.

The study authors analyzed what happened to average student achievement in a cohort when one of four events took place:

1. a high value-added teacher started teaching students in the cohort;
2. a high value-added teacher stopped teaching students in the cohort;
3. a low value-added teacher started teaching students in the cohort; or
4. a low value-added teacher stopped teaching students in the cohort.

Typically, the teacher entered one classroom; therefore, not all the students in the cohort would have been exposed directly to the teacher. For example, if a fourth-grade classroom teacher entered a school with three fourth-grade classrooms, she would teach her own students, but students in the other two classrooms would not be directly exposed to her teaching.

The movement of teachers into and out of schools and classrooms was not a result of any policy, program, or intervention; it occurred naturally.

Comparison group

The comparison group was the cohort of students in the same subject and grade at the school in the year before the change in teachers. In other words, the study compared the test scores of students in the year after getting a high value-added teacher to the test scores of students in the same school, subject, and grade in the year before the teacher's arrival.

Outcomes and measurement

The study analyzed standardized student math and reading assessment scores. For a more detailed description of this outcome measure, see Appendix B.

Support for implementation

Does not apply to this study.

Reason for review

This study was identified for review by receiving significant media attention.

Appendix B: Outcome measures for each domain

Student achievement	
School-grade-subject-year mean English language arts and math test scores	English language arts and math test scores were available for all students in grades 3–8 for every year from 1989 through 2009, with the exception of seventh-grade scores in 2002. The tests were specific to the district until the late 1990s, at which point the district began using statewide assessments in grades 4 and 8, with all grades transitioning to the state assessment by 2006. For the analysis, the study authors standardized individual students' test scores so that they had a mean of zero and a standard deviation of one within each grade, subject, and year in the entire district. The authors then took the average standardized test scores for each school, subject, grade, and year for students in grades 4–8 between 1991 and 2009 and analyzed them jointly.

Appendix C: Study findings for each domain

Domain and outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Student achievement: Events hypothesized to increase student achievement								
Average ELA and math test scores	Cohorts after a high value-added teacher joined	3,384 student school-grade-subject-year cohorts	0.23 (1.00)	0.20 (1.00)	0.03	0.03	1	< 0.001
Average ELA and math test scores	Cohorts after a low value-added teacher left	3,202 student school-grade-subject-year cohorts	0.16 (1.00)	0.14 (1.00)	0.02	0.02	1	< 0.001
Domain average for events hypothesized to increase student achievement						**0.03**	**1**	**Statistically significant**
Student achievement: Events hypothesized to decrease student achievement								
Average ELA and math test scores	Cohorts after a high value-added teacher left	3,304 student school-grade-subject-year cohorts	0.19 (1.00)	0.24 (1.00)	–0.05	–0.05	–2	< 0.001
Average ELA and math test scores	Cohorts after a low value-added teacher joined	3,286 student school-grade-subject-year cohorts	0.19 (1.00)	0.22 (1.00)	–0.03	–0.03	–1	< 0.001
Domain average for events hypothesized to decrease student achievement						**–0.04**	**–2**	**Statistically significant**

Table Notes: Table reports two sets of findings: one from the events hypothesized to increase student achievement, and one for those hypothesized to decrease student achievement. For mean difference, effect size, and improvement index values reported in the table, a positive number favors the intervention group and a negative number favors the comparison group. The effect size is a standardized measure of the effect of an event on student outcomes, representing the change (measured in standard deviations) in an average student's outcome that can be expected if the student is subject to the event; for this study, the effect size is the same as the difference in means because the authors analyzed standardized test scores as the dependent variable. The improvement index is an alternate presentation of the effect size, reflecting the change in an average student's percentile rank that can be expected if the student is subject to the event. The WWC-computed average effect size is a simple average rounded to two decimal places; the average improvement index is calculated from the average effect size. The statistical significance of the study's domain average was determined by the WWC. ELA = English language arts.

Study Notes: No corrections for clustering or multiple comparisons were needed. The mean values presented here are estimated from Figure 3 of the study. Standard deviations are assumed to be 1.00 because outcomes were expressed as z-scores. The *p*-values were reported in the study and are from F-tests of the hypothesis that the change in test score gains from the year before the event (as appropriate) to the year after equals zero.

Endnotes

[1] Single study reviews examine evidence published in a study (supplemented, if necessary, by information obtained directly from the author[s]) to assess whether the study design meets WWC evidence standards. The review reports the WWC's assessment of whether the study meets WWC evidence standards and summarizes the study findings following WWC conventions for reporting evidence on effectiveness. This study was reviewed using the single study review protocol, version 2.0. A quick review of this study was released on March 1, 2012, and this report is the follow-up review that replaces that initial assessment. The WWC rating applies only to the results that were eligible under this topic area and met WWC standards without reservations or met WWC standards with reservations, and not necessarily to all results presented in the study.

[2] Typically, the teacher entered one classroom in the school and grade. Therefore, not all the students in that school and grade would have been exposed directly to the teacher. For example, if a fourth-grade classroom teacher entered a school with three fourth-grade classrooms, she would teach her own students, but students in the other two classrooms would not be directly exposed to her teaching.

Recommended Citation

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2012, October). *WWC review of the report: The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood.* Retrieved from http://whatworks.ed.gov.

Glossary of Terms

Attrition Attrition occurs when an outcome variable is not available for all participants initially assigned to the intervention and comparison groups. The WWC considers the total attrition rate and the difference in attrition rates across groups within a study.

Clustering adjustment If intervention assignment is made at a cluster level and the analysis is conducted at the student level, the WWC will adjust the statistical significance to account for this mismatch, if necessary.

Confounding factor A confounding factor is a component of a study that is completely aligned with one of the study conditions, making it impossible to separate how much of the observed effect was due to the intervention and how much was due to the factor.

Design The design of a study is the method by which intervention and comparison groups were assigned.

Domain A domain is a group of closely related outcomes.

Effect size The effect size is a measure of the magnitude of an effect. The WWC uses a standardized measure to facilitate comparisons across studies and outcomes.

Eligibility A study is eligible for review if it falls within the scope of the review protocol and uses either an experimental or matched comparison group design.

Equivalence A demonstration that the analysis sample groups are similar on observed characteristics defined in the review area protocol.

Improvement index Along a percentile distribution of students, the improvement index represents the gain or loss of the average student due to the intervention. As the average student starts at the 50th percentile, the measure ranges from –50 to +50.

Multiple comparison adjustment When a study includes multiple outcomes or comparison groups, the WWC will adjust the statistical significance to account for the multiple comparisons, if necessary.

Quasi-experimental design (QED) A quasi-experimental design (QED) is a research design in which subjects are assigned to intervention and comparison groups through a process that is not random.

Randomized controlled trial (RCT) A randomized controlled trial (RCT) is an experiment in which investigators randomly assign eligible participants into intervention and comparison groups.

Single-case design (SCD) A research approach in which an outcome variable is measured repeatedly within and across different conditions that are defined by the presence or absence of an intervention.

Standard deviation The standard deviation of a measure shows how much variation exists across observations in the sample. A low standard deviation indicates that the observations in the sample tend to be very close to the mean; a high standard deviation indicates that the observations in the sample are spread out over a large range of values.

Statistical significance Statistical significance is the probability that the difference between groups is a result of chance rather than a real difference between the groups. The WWC labels a finding statistically significant if the likelihood that the difference is due to chance is less than 5% ($p < 0.05$).

Substantively important A substantively important finding is one that has an effect size of 0.25 or greater, regardless of statistical significance.

Please see the WWC Procedures and Standards Handbook (version 2.1) for additional details.