The University of Texas at Austin

Education Research Center

www.texaserc.utexas.edu

POLICY BRIEF

The Impact of Teacher Professional Development on Student Achievement in Math and Science

Jayce R. Warner, Carol L. Fletcher, & Lisa S. Garbrecht

December 2019

What We Studied

This study examined the impact of a multi-year, large-scale professional development program on student achievement in mathematics and science. Math and science teachers at the elementary and secondary levels who participated in the professional development program were matched to a control group. Results revealed no differences in STAAR test scores between treatment and control groups for all grades and subjects except elementary students in rural areas. In rural areas, fifth grade students of teachers in the treatment group outperformed students of teachers in the control group on the STAAR Math and STAAR Science assessments. These results suggest that the large-scale professional development program is particularly beneficial for rural elementary school teachers, perhaps due to rural teachers having a greater need for the connections and resources gained by participating in the program and to elementary teachers having more flexibility in the amount of time they spend teaching math and science in their classrooms.

Effect sizes for the statistically significant differences were very small. Estimates of the variance in test scores at the three hierarchical levels (students, teachers, and schools) revealed that teachers account for relatively little of the variation in students' scores. This finding suggests that expectations that professional development programs themselves will significantly impact student achievement on standardized tests may be unrealistic. It also points to the potential value of systemic professional development programs that involve schools and school leaders in addition to training individual teachers.

Although there has been strong evidence that professional development increases teachers' knowledge and skills, which in turn leads to more effective instruction (Darling-Hammond et al., 2009), few studies have examined the direct link between professional development and student achievement, and fewer still have done so with the rigor needed to facilitate causal inferences regarding the effectiveness of professional development on student outcomes (Yoon, Duncan, Lee, Scarloss, & Shapley, 2007). Moreover, studies have provided mixed evidence for the overall effectiveness of professional development on student achievement. Some studies have found positive effects (Heller et al., 2012; McMeeking et al.,2012), null effects (Garet et al., 2008; Garet et al., 2010), or even negative effects of professional development on student outcomes (Borman, Gamoran, & Bowdon, 2008). As a result, education leaders are left without the clear guidance needed to form accurate expectations of the potential impact that different professional development programs can have on student achievement.

How We Analyzed the Data

The mixed results from efficacy and effectiveness studies of professional development demand that more attention be given to the way such studies are conducted. Such was the purpose of the current study. Using hierarchical linear modeling (HLM), we investigated the impact of a large-scale math and science teacher professional development program on students' standardized test scores. Multilevel models like HLM allow for the outcome variance to be partitioned across levels, which can lead to better understanding of the strength of program effects since it allows one to examine the proportion of variance explained at each level. The following research questions guided our analysis:

1. What is the effect of the professional development program on student test scores?
2. What proportion of variance in student test scores can be explained by differences at each level of analysis (students, teachers, and schools) and by the treatment (the professional development program)?

The teachers that composed the treatment group participated as *Math Teacher Mentors* or *Science Teacher Mentors* within the Texas Regional Collaboratives (TRC) professional development program during the 2014-15 and 2015-16 school years. The TRC funds and facilitates sustained, intensive professional development for math, science, and computer science teachers across the state. This is done by providing funds, training, and resources to regional *collaboratives* led by local teacher development specialists. Leaders in each collaborative determine their own professional development agenda that meets the requirements of the subgrants they receive. Each collaborative is required to recruit a number of teacher mentors, who participate in a minimum of 100 hours of high-quality, content-focused professional development training each school year. In the 2015-16 school year, there were a total of 694 Math Teacher Mentors and 976 Science Teacher Mentors who participated in the TRC program. This study tested whether students of these TRC teachers outperformed non-TRC students on the math and science STAAR tests.

A control group was created by matching TRC teachers to non-TRC teachers using propensity scores. Propensity score matching is a regression technique that is used to identify control group participants that most closely match treatment group participants on a variety of characteristics. For every TRC teacher, one non-TRC teacher was identified through this matching process. Students for all TRC and non-TRC teachers were included in the analysis as treatment group or control group participants, respectively.

Eight separate analyses were conducted, one for each grade level and subject for which state standardized test scores were available and in which sample sizes were adequate. These included STAAR Math assessments at 5th, 6th, 7th, and 8th grade, STAAR Science assessments at 5th and 8th grade, the Algebra I STAAR End-of-Course exam at 9th grade, and the Biology STAAR End-of-Course exam at 9th grade. Each analysis proceeded in two phases. The first phase tested for treatment effects using the entire sample of matched participants. Next, because TRC collaboratives tend to reach teachers of smaller, rural school districts more than the larger, urban districts, a second phase tested for effects on a subsample that only included participants from the smaller, more rural types of school districts.

Three-level hierarchical linear modeling, with students nested within teachers nested within schools, was used to test for differences between treatment and control groups. Among many other variables entered as covariates in the analyses, students' STAAR scores from the prior school year were used as a means of controlling for students' aptitude and prior knowledge. Because there was no STAAR Science assessment given at the 4th and 7th grade levels, students' 2015 STAAR Math scores from 4th and 7th grade were used as the pretest covariate for the 2016 STAAR Science analyses in 5th and 8th grade, respectively. Correlations between students' prior-year math scores and current-year science scores were similar to the correlations between prior and current math scores (Pearson's r = .71 to .83), suggesting that the prior-year math score may be an adequate pretest covariate for 5th and 8th grade science outcomes.

What We Discovered

Treatment Effects

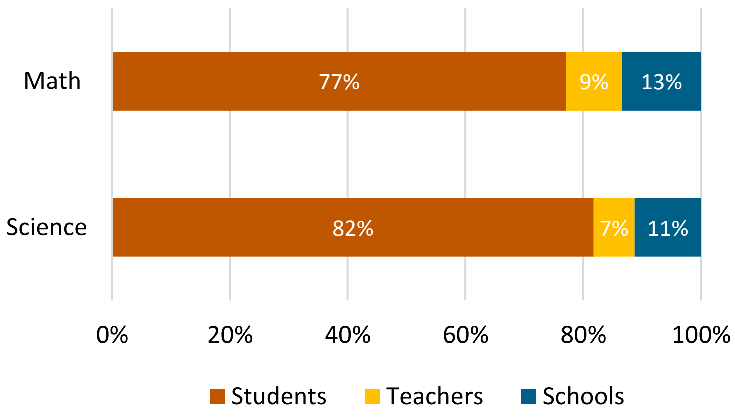
No effects of the treatment were detected for any of the dependent variables in the first phase of analyses. In the second phase of analyses, which focused on smaller, more rural districts, two effects of the treatment were found. Students of TRC teachers outperformed control group students on the 5th grade STAAR Math assessment and the 5th grade STAAR Science assessment. The sample for the STAAR Math analysis consisted of 3,993 students across 80 teachers, and the sample for STAAR Science consisted of 9,327 students across 164 teachers. The number of participants was approximately even between treatment and control groups for both analyses. Controlling for all other variables in the model, 5th grade students of TRC teachers scored .19 standard deviations higher than control group students on the STAAR Math assessment and .10 standard deviations higher on the STAAR Science assessment than their peers in the control group.

Variance Estimates

Overall Variance. The magnitudes of these treatment effects were estimated by calculating the proportions of variance in student test scores that could be explained by the treatment (Raudenbush & Bryk, 2002). These estimates showed the differences between treatment and control groups to be very small, as the treatment explained less than 1% of the total variance in students' 5th grade math scores and less than 1% of the total variance in students' 5th grade science scores.

Variance at Each Level. Using HLM to analyze the data has the added benefit that the variance in student test scores can be partitioned across the levels (students, teachers, and schools) accounted for in the analysis. This provides an estimate of the proportion of variance that is accounted for at each level, or, in other words, the proportion of variance in test scores that can be attributed to differences between units within each level. For 5th grade math, 77% of the variance in test scores was attributed to differences between students, 9% was attributed to differences between teachers, and 13% was attributed to differences between schools. For 5th grade science, 82% of the variances was found to be between students, 7% between teachers, and 11% between schools (see Figure 1).

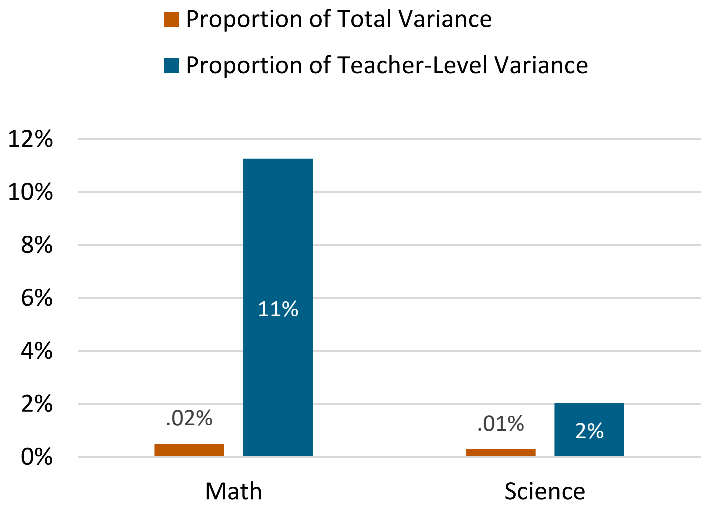
Figure 1. Proportion of Variance* in 5th Grade STAAR Math and Science Test Scores Accounted for at the Student, Teacher, and School Level



**Intraclass correlations from the unconditional model*

When estimating effect sizes in multilevel models such as this one, it is important to note the level at which the treatment (the intervention) takes place. As the intervention in this study was teacher professional development, the treatment occurred at the teacher level. In other words, the treatment was designed and implemented to impact factors related to teachers and their instruction. No intervention was implemented directly with students, so no part of the treatment can be said to have occurred at the student level. Similarly, the professional development was not designed to influence schoolwide policy or practice, nor was it implemented across all teachers at each school. As a result, it is not likely that the treatment had direct impacts on school-level factors. Because the treatment only occurred at the teacher level, it can be useful to calculate the proportion of the *teacher-level* variance that can be explained by the treatment (in addition to the proportion of the *total* variance in student scores explained by the treatment as reported above, which, as noted above, was less than 1%). For 5th grade math, 11% of the teacher-level variance in student scores was explained by the treatment. For 5th grade science, 2% of the teacher-level variance in student scores was explained by the treatment (see Figure 2).

Figure 2. Proportion of Variance* in 5th Grade STAAR Math and Science Test Scores Explained by the Treatment



**Incremental variance explained by adding treatment to the final model*

Policy Recommendations

Examining the variance explained at each level (i.e., students, teachers, and schools) offers direction for researchers, practitioners, and policymakers concerned with teacher professional development.

When the outcome variance (in this case, the variance in student test scores) is partitioned across levels, it represents the proportion of variance that is "up for grabs" at each level. In this study, the treatment was given only to teachers, who comprised the second level of the three-level hierarchy. For example, because only 9% of the variance in student math test scores and 7% of the variance in student science scores can be attributed to teacher-level factors, and because the intervention was only given at the teacher level, the treatment (the professional development program) has the potential to directly impact only 9% and 7%, respectively, of the variance in student math and science scores. This means that because test scores are largely dependent on student-level factors, it can be extremely difficult to move the needle on student achievement through teacher professional development alone.

There are two important implications of this. First, if effect sizes of the impact of professional development programs on student achievement are calculated only as proportions of the total variance (i.e., without partitioning the variance into each level), evaluators run the risk of attenuating the interpretation of program effects. Alternatively, reporting both types of effect sizes (i.e., the proportion of total variance explained by the treatment and the proportion of teacher-level variance explained by the treatment) would aid in understanding results. The results of the analysis for math in this study illustrates this as the treatment accounted for only 1% of the total variance in student math scores but 11% of the variance for which it had any potential to impact. Second, these results point to the value of systemic-minded professional development programs, that is, professional development programs that push for change at the school and district levels rather than only targeting teachers and instructional practice. In this study, 13% of the variance in student math scores and 11% of the variance in student science scores was between schools (the third level). The implication here is that interventions that target school and district policies (in addition to teacher practice) and involve school and district leaders (in addition to teachers) may increase the amount of variance in student test scores that can potentially be impacted by the treatment, thus increasing the potential to impact student achievement.

It is important to note the limitations of this study. One limitation is that the professional development experiences of the teachers in the control group were unknown. While teachers in the treatment group all participated in at least 100 hours of professional development, control group teachers could have participated in any number of professional development hours. Because it is likely that the control group teachers did participate in some other professional development programs (as teacher professional development is a common practice in Texas), this study did not compare teachers with professional development to teachers without, and thus does not speak to the question of whether teacher professional development in general is a worthwhile endeavor. Instead, this study compared one professional development program to the status quo, or what might be considered the typical professional development experiences that teachers have (though, again, that status quo is assumed rather than explicitly defined in this study).

Another limitation has to do with the potential for the presence of unknown confounding factors that affected the results. This study did not randomly assign teachers to treatment and control conditions. Although the matching procedure used to create the control group is designed to control for other factors that may affect the results, the lack of random assignment leaves open the possibility that there may be unknown confounding factors that influenced the results. Thus, caution should be taken when drawing conclusions regarding the results of the treatment effects testing in the study.

Nevertheless, the variance estimates reported here can be viewed with confidence as a clear indication of the infeasibility of achieving substantial impacts on student achievement through interventions that only target teachers. Similarly, the results of this study suggests that measuring student achievement gains may not be a valid means of evaluating the impact of teacher professional development.

References

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.

Darling-Hammond, L., Wei, R. C., Andree, A., Richardson, N., & Orphanos, S. (2009). *Professional Learning in the Learning Profession: A Status Report on Teacher Development in the US and Abroad.* Dallas, TX: National Staff Development Council.

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., ... & Zhu, P. (2008). *The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement*. NCEE 2008-4030. Washington, DC:National Center for Education Evaluation and Regional Assistance.

Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Walters, K., Song, M., ... & Doolittle, F. (2010). *Middle School Mathematics Professional Development Impact Study: Findings After the First Year of Implementation.* NCEE 2010-4009. Washington, DC:National Center for Education Evaluation and Regional Assistance.

Heller, J. I., Daehler, K. R., Wong, N., Shinohara, M., & Miratrix, L. W. (2012). Differential effects of three professional development models on teacher knowledge and student achievement in elementary science. *Journal of Research in Science Teaching*, *49*(3), 333-362.

McMeeking, S., Orsi, R., & Cobb, R. B. (2012). Effects of a teacher professional development program on the mathematics achievement of middle school students. *Journal for Research in Mathematics Education*, *43*(2), 159-181.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods, 2nd Ed*. Thousand Oaks, CA: Sage.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. L. (2007). Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement. Issues & Answers. REL 2007-No. 033. *Regional Educational Laboratory Southwest (NJ1)*.