



# REVIEW OF *EVALUATION OF TEACH FOR AMERICA IN TEXAS SCHOOLS*

*Reviewed By*

Edward J. Fuller, Penn State University  
Nathan Dadey, University of Colorado Boulder

April 2013

## Summary of Review

A new study funded by Teach for America (TFA) attempts to identify the effect of TFA teachers and alumni on student test scores. The report, by Edvance Research, matched schools and students within those schools on both demographic and achievement characteristics. It then used the matched student data in a multi-level regression analysis to estimate the effect of being taught by a TFA teacher on mathematics and reading test scores for two groups of students: those in grades 4 and 5 and those in grades 6 through 8. Of the eight comparisons performed, three were statistically significant: mathematics and reading scores of middle-level students taught by TFA alumni and mathematics scores for middle-level students taught by novice TFA teachers. While the findings were large enough to be relevant to policymakers, numerous issues with the sample construction, matching procedures, and statistical analyses lead us to conclude the outcomes cannot be attributed to TFA teachers. Furthermore, these issues make it likely that the actual size of the TFA teacher effects differ than what is found in the report. Thus, the report should not be accepted as solid evidence of any TFA teacher or TFA alumni effect.

**Kevin Welner**

*Project Director*

**William Mathis**

*Managing Director*

**Erik Gunn**

*Managing Editor*

---

**National Education Policy Center**

School of Education, University of Colorado

Boulder, CO 80309-0249

Telephone: (802) 383-0058

Email: [NEPC@colorado.edu](mailto:NEPC@colorado.edu)

<http://nepc.colorado.edu>

Publishing Director: Alex Molnar

---



This is one of a series of Think Twice think tank reviews made possible in part by funding from the Great Lakes Center for Education Research and Practice. It is also available at <http://greatlakescenter.org>.

*This material is provided free of cost to NEPC's readers, who may make non-commercial use of the material as long as NEPC and its author(s) are credited as the source. For inquiries about commercial use, please contact NEPC at [nepc@colorado.edu](mailto:nepc@colorado.edu).*

# REVIEW OF *EVALUATION OF TEACH FOR AMERICA IN TEXAS SCHOOLS*

*Edward J. Fuller, Penn State University  
Nathan Dadey, University of Colorado Boulder*

---

## I. Introduction

There is widespread agreement among both researchers and policymakers that teacher quality is the single most important school-based factor influencing student test scores. There has long been evidence that students in high-needs schools—those schools serving large percentages of poor and minority students—have had less access to well-qualified and stable groups of teachers<sup>1</sup>. Largely in response to this concern, Teach for America (TFA) was created to provide high-quality teachers to high-need schools as a strategy to improve student achievement and reduce the achievement gap. The underlying assumption of the TFA strategy is that academically accomplished individuals provided with even a short period of training will have a larger positive impact on student test scores than existing teachers in high-need schools. There has been contentious debate over the effectiveness of TFA teachers in improving student test scores and other student outcomes, with little high-quality and independent research to inform policymakers about the actual effects of TFA teachers on student test scores.

At the beginning of March, Edvance Research, Inc. and Teach for America released a report titled *Evaluation of Teach for America in Texas Schools*.<sup>2</sup> The report purportedly focused on estimating the impact of Teach for America (TFA) teachers on student test scores on the state-mandated Texas Assessment of Academic Skills (TAKS) in two sets of grade levels: <sup>3</sup> grades 4-5 and grades 6-8, which we refer to as the elementary grades and middle grades in this review.<sup>4</sup> More specifically, the report compares differences between average student TAKS scores for students taught by novice TFA and non-TFA teachers as well as by TFA alumni<sup>5</sup> and experienced non-TFA teachers<sup>6</sup> in mathematics and reading. The report adds to the growing literature base on TFA teachers in particular and the effects of alternatively certified teachers in general.

## II. Findings and Conclusions of the Report

In sum, the main finding was three of the eight comparisons of student TAKS scores between TFA and non-TFA teachers yielded statistically significant results. All three statistically significant findings were in the middle grades, with two of the results in mathematics and one in reading. In mathematics, students of both novice TFA teachers and TFA alumni had greater TAKS scores than students of novice and experienced non-TFA teachers, respectively, after controlling for prior scores and other factors. The report describes the differences in achievement in students between TFA and non-TFA teachers as “substantial,” with the largest impact for TFA alumni (p. 2). In reading, students of TFA alumni had slightly greater TAKS scores than experienced non-TFA teachers after controlling for prior scores and other factors. Table 1 includes a summary of the results. The three statistically significant results had moderate to large effect sizes. The effects, if accurate, would be large enough to have policy implications.

In contrast, the report found no statistically significant differences at the elementary school level between novice TFA and novice non-TFA teachers in either mathematics or reading. The results of the comparisons between TFA alumni and experienced non-TFA teachers at the elementary grade level was not presented with the results of the other regressions<sup>7</sup>; the findings, however, are the same as those for novice TFA teachers: no statistically significant differences. Similarly, between novice TFA and novice non-TFA teachers, there was no statistically significant result in reading in the middle grades.

**Table 1. Summary of Results by Grades Served, Subject Area, and Teacher Experience**

Grades	Subject	Novice Teachers		Experienced Teachers	
		TAKS Score Difference	Effect Size	TAKS Score Difference	Effect Size
4 - 5	Reading	4.1	0.04	-4.1	-0.05
	Math	10.0	0.11	3.5	0.04
6 - 8	Reading	3.8	0.04	10.4*	0.11
	Math	16.9**	0.19	23.3**	0.27

Note: Asterisks indicate statistical significance (\*\*p<.001, \*p<.05).

## III. The Report’s Rationale for Its Findings and Conclusions

The conclusions in the report are based on the results of a multi-level regression analysis technique called hierarchical linear modeling (HLM). The HLM procedure used student TAKS scores in the 2010-11 school year for students taught by TFA teachers/alumni and

non-TFA teachers. Importantly, due to data limitations, the students were grouped at the school level, not the teacher level. This hierarchical linear model was used to estimate the effect of students having been taught by TFA teachers or TFA alumni in schools employing at least one TFA teacher or TFA alumnus. In addition, the HLM procedure accounted for the nested structure of the data (i.e., students within schools). To control for factors other than whether a student was taught by a TFA teacher or a TFA alumnus (henceforth referred to as TFA teachers/alumni), schools and students were matched according to demographic and achievement characteristics using propensity score matching (PSM)<sup>8</sup> at the school and student levels. First, schools employing at least one TFA teacher/alumnus (316 schools) were matched to a much larger set of schools not employing any TFA teachers/alumni that included almost all remaining public schools in Texas (7,882 schools). The initial match resulted in the 316 TFA schools being matched to 924 non-TFA schools. This set of 316 schools was divided across subjects and grades in various ways, with sample sizes of schools ranging from 25 to 55 for novice TFA teachers and 14 to 18 for TFA alumni, depending on the subject area and grades addressed. The initial and final analytic samples are displayed in Table 2 (mathematics) and Table 3 (reading).<sup>9</sup>

**Table 2. Comparison of the Number of Schools in the Initial and Final Samples for Mathematics**

School	Sample	Novice Teachers		Experienced Teachers	
		TFA	Not-TFA	TFA	Not-TFA
<i>Grades</i>					
<i>4th and 5th</i>	Initial	81	385	81	385
	Final	25	90	14	98
<i>6th, 7th, and 8th</i>	Initial	113	375	113	349
	Final	51	205	12	200

**Table 3. Comparison of the Number of Schools in the Initial and Final Samples for Reading**

School	Sample	Novice Teachers		Experienced Teachers	
		TFA	Not-TFA	TFA	Not-TFA
<i>Grades</i>					
<i>4th and 5th</i>	Initial	81	385	81	385
	Final	37	298	14	298
<i>6th, 7th, and 8th</i>	Initial	113	375	113	375
	Final	55	157	18	181

Appendix C in the report suggests an initial sample of 194 schools, while the text of the report lists the initial sample of schools as 316. The report does state that there was some

overlap of campuses, but the overlap is never described, nor is the apparent difference in sample sizes explained in the report. The answer likely lies somewhere between the two numbers, but there is no way to determine the actual number of schools in the overall initial sample or to assess how many schools were included in both the elementary and middle grades analyses. One would have hoped that the samples of elementary grades schools and middle grades schools would have remained separate to ensure grade span was not associated with student achievement.

Also note that the report never includes the number of teachers included in any of the various samples. This is because the state did not include any teacher information in the data set used for the study. Thus, we never know how many actual TFA teachers/alumni were included in the analysis.

Once the two sets of schools were identified, students of TFA teachers/alumni in a TFA school were matched with students of non-TFA teachers/alumni in a non-TFA school. To further account for pre-existing differences while estimating the effects on student test scores, the student-level HLM analysis controlled for prior student test scores and variables that the matching procedure did not fully account for, including student demographics, school demographics and school achievement levels.

The report assumes no serious methodological flaws in either the PSM or HLM procedures, and thus, concludes the statistically significant findings indicate students taught by TFA teachers have greater test scores than students not taught by TFA teachers. The findings are then used to support the conclusion that middle grade students of TFA alumni, as well of novice TFA teachers in mathematics, gain half a year to a full year of additional learning relative to students of non-TFA teachers. While the report mentions at the end of the paper some important limitations of the study and caution that the reported TFA effects may be indicative of more than *just* the impact of TFA teachers, it does not nuance the conclusions in a way that communicates the tentative nature of the findings. As shown in the remainder of this review, the serious methodological flaws in the matching procedure and in the hierarchical linear model raise substantial concerns about the accuracy of the results and, consequently, the report's conclusions about the effectiveness of TFA teachers.

#### IV. The Report's Use of Research Literature

The report vaguely mentions eight studies (seven of which were conducted with students outside of Texas) that examined the impact of TFA teachers on student test scores and provides a cursory review of only one of these studies.<sup>10</sup> Of the eight, four found TFA teachers had a positive effect on student test scores, although only one of the four studies was a peer-reviewed journal article. Three of the eight studies found greater effects for non-TFA teachers; all three of those studies were peer-reviewed journal articles. The final article mentioned was a peer-reviewed journal article that found no difference between TFA and non-TFA teachers. Given that three of the reports reviewed that found positive

effects for TFA were not peer-reviewed studies, one must be cautious in giving any weight to them, considering the potential methodological or interpretational issues that peer review often uncovers. Importantly, the report does not review or critique the methodological approaches used in the eight non-Texas studies.

The report also provides a slightly more detailed review of three previous Texas studies of TFA teachers, only one of which was a peer-reviewed journal article. Unlike the review of the non-Texas studies, the report adequately examined some of the methodological problems in the three Texas studies cited. Interestingly, however, the report unquestioningly accepts the positive results of the Ware *et al.* (2011) study from Texas that had glaring methodological issues that rendered the positive achievement results meaningless with respect to any TFA teacher effect.<sup>11</sup> The Ware *et al.* study, in fact, relied on changes in the *percentage* of students passing the TAKS mathematics and reading test as the measure of achievement. Changes in percentage passing, however, are not accurate indicators of changes in student performance.<sup>12</sup>

In sum, for a report that spans 109 pages, the review of the literature is certainly less than adequate, particularly with respect to the methodologies employed in other reports. Indeed, a more robust description of the methods employed in other studies—in particular, the study<sup>13</sup> highlighted as “the best evidence to date” (p. 10)—would have provided readers an opportunity to compare the methodology employed in the report at hand to those used in other reports.

## V. Review of the Report’s Methods

As noted above, the report in question used two methods: (1) propensity score matching (PSM) to match schools and students to create two groups of students similar on the observed variables selected by the researchers, except on the question of whether they were being taught by TFA teachers/alumni, and (2) hierarchical linear modeling (HLM) to estimate the effect of a student being taught by TFA teachers/alumni. Below, we first review and critique the PSM procedure and then follow with a review and critique of the HLM model. After reviewing the methods of the study, we then provide a four-part critique of the study: critique of overall approach, critique of the sample used, critique of the PSM procedure, and critique of the HLM analysis.

### Review of Sample Construction and Matching Procedure

The first step of the PSM procedure was to match schools that employed at least one TFA mathematics or reading teacher in grades 4 through 8 in the 2010-11 school year with schools that did not employ a TFA teacher in the same subjects and grade levels in the same school year. The schools were matched based on several demographic and achievement characteristics. Demographic characteristics included the percentages of students by ethnicity, economically disadvantaged status, special education status, limited

English proficiency, student mobility, total student enrollment, the percentage of teachers in their first year of teaching, and the number of teacher full-time equivalents at the school. The achievement characteristics included the percentage of students meeting the state passing standard in mathematics and reading across all grade levels in the schools for which test scores were reported. The report states 924 of the potential 7,882 non-TFA campuses were initially identified through the PSM procedure as being a set of matched non-TFA campuses. In addition, all schools in the districts with at least one TFA campuses were also included “as a backup of the comparison campuses in case the 924 non-TFA comparison campuses could not provide a sufficient number of comparable students in the student-level matching” (p. 86). Ultimately, 1,641 comparison campuses were identified, with 717—44%—of the non-TFA campuses coming from this backup list. The interpretations of the study rely heavily on the assumption that student and school characteristics were controlled for through the matching procedure. The failure to match all schools, though, undercuts the confidence in the findings as accurate indicators of any TFA effect.

After schools were identified and matched, the second step of the PSM was to match students based on both demographic and achievement characteristics. To be included in the matching procedure, students had to be in either schools employing a TFA teacher or one of the comparison schools identified through the first step of the PSM. A sample of students from within these schools was constructed that complied with the federal educational rights and privacy act (FERPA). Because students’ personal characteristics were included in the study and the students were grouped at the individual course level, a substantial number of students had to be excluded from the analysis to comply with FERPA. The flowcharts presented in Appendix C of the study document the number of students in the initial sample. Of students in the TFA schools, about 60% of the initial sample was excluded at the elementary grades and 70% at the middle grades.

Students were then matched on both demographic and achievement characteristics. The demographic characteristics included gender, ethnicity, economically disadvantaged status, special education status, limited English proficiency, and mobility. The achievement characteristics included 2009-10 student TAKS mathematics and reading achievement scores as well as achievement scores on all other available TAKS tested content areas (science, social studies, and writing) for each respective grade level.

### **Critique of Overall Methods**

There are two major critiques of the overall methods employed in the study. First, since the research questions posed in the report spoke about estimating the effect of teachers on student test scores, the unit of analysis should have been the teacher or course. Yet, because of the masking of student data to comply with FERPA restrictions, the study could not be conducted at the teacher/classroom level and was, instead, conducted at the school/campus level. Importantly, then, the findings of the report *do not* allow for conclusions to be reached about the impact of *teachers* on student test scores. The findings speak more to the impact of the schools than teachers on student test scores. Even if we

accepted the claim that the findings reflect the impact of teachers, the failure to match teacher and classroom characteristics between TFA and non-TFA teachers led to the omission of numerous factors that influence student test scores; thus, they are certainly inaccurate estimates of any TFA effect on student test scores.

Second, as noted in the study at hand, the ideal methodology would have been to randomly assign students to teachers and then compare student achievement of TFA and non-TFA teachers matched on various characteristics; this was the approach used in the 2004 study by Decker, Mayer, and Glazerman<sup>14</sup> of TFA teachers. The PSM method used in the Texas report has the potential to approximate the outcome of a randomized experiment by matching characteristics to control for factors affecting an outcome—in this case, student test scores. In this study, however, there were serious problems with the procedures, which are described below.

### Critique of Sample Construction

We identified a number of issues related to the creation of the final sample of students. First, the number of students excluded from the initial sample was substantial. Indeed, as shown in Table 4, only about 38% of the original elementary and 27% of the original middle grades samples of students in TFA schools remained in the sample after the successive application of the two major reasons for the exclusion of students. In comparison, the percentage of students remaining in the sample for students in non-TFA schools was 74% for the elementary grades and almost 87% for the middle school grades. Importantly, the report does not provide an accurate comparison of the characteristics of

**Table 4. Effect of Excluding Schools with Students not Enrolled in Mathematics Classes on the Number of Schools Included in the Sample**

	Grades 4-5				Grades 6-8			
	TFA		Not TFA		TFA		Not TFA	
	N	%	N	%	N	%	N	%
<i>Original Sample</i>	4,801	--	73,225	--	22,275	--	247,000	--
<i>Excluding Students</i>								
<i>Not Enrolled in Math</i>	1,967	41%	59,960	82%	6,774	30%	233,572	95%
<i>Absent too Many Days</i>	1,821	38%	54,215	74%	6,087	27%	213,888	87%

Notes: % refers to percent of original sample (the top row). The figures are for mathematics only. However, the patterns of exclusions in reading are similar enough to those in mathematics that we have excluded the reading statistics for the sake of simplicity.

students in the initial sample with the characteristics of the students in the much smaller FERPA-compliant sample used in the analysis. The analysis should have compared and

reported the demographics and achievement of all students taught by TFA teachers/alumni in TFA schools to the final sample of students taught by TFA teachers in TFA schools after the masking of data to comply with FERPA. Because the report did not furnish such an analysis, we do not know how the masking procedure affected the demographic and achievement profiles of students taught by TFA teachers. This matters because if the final sample of students taught by TFA teachers was substantially different from the initial sample, then the findings would not necessarily be generalizable to the entire population of students taught by TFA teachers.

Second, the report failed to consistently provide accurate and comprehensive information about which students were included in the final sample and which students were not included. While the flowcharts in Appendix C provide some detail about the exclusion of

*The failure to match teacher and classroom characteristics between TFA and non-TFA teachers led to the omission of numerous factors that influence student test scores.*

students, other areas in the report describe additional exclusions not included in the flowcharts. For example, footnote 50 on page 23 notes, “Among student demographic variables, non-economically disadvantaged students were excluded from the analysis because TFA primarily focuses on economically disadvantaged communities.” In addition, students taking a modified, alternate, or Spanish-language version of the 2010-11 TAKS were excluded from the analysis. Our own estimates using student-level data for the TAKS in 2010-11 suggests employing these three exclusions would result in excluding between 30% and 40% of all students depending on the grade level and subject area.<sup>15</sup> These exclusions would be particularly problematic if applied after the PSM, as the groups would no longer be comparable on the variable upon which they were matched. Unfortunately, the report provides very little information about these exclusions, and we could not verify the characteristics of the students in the various samples used in the report. Ultimately, readers simply do not have enough information to fully understand to which conclusions and implications the exclusions applied, and thus should not blindly accept as appropriate the conclusions made in the study.

Third, the inclusion of the average of prior TAKS scores over the three previous years in the hierarchical linear model suggests that students with missing TAKS scores were excluded from the analysis.<sup>16</sup> As noted above, the study provides no details about the number and percentage of students excluded from the sample for these reasons and whether the exclusions affected TFA and non-TFA students equally. Our own analysis of student-level Texas TAKS data from the same year and of schools in the same regions as the report suggests using the above three exclusion criteria would exclude more than 50% of all students from the original population of students.

Fourth, as was the case with the creation of the sample of students, many details of the construction of the final sample of schools were missing from the report. For example, as

noted above, 717 (44%) of the non-TFA campuses were from the district “back up” set of comparison schools and, as the report admitted, were not subjected to the PSM procedure. As a consequence, the report could *not* fully control for school characteristics; the HLM analysis and the report’s conclusions, however, assumed the PSM procedure *did* adequately control for school characteristics.

Fifth, the initial sample of TFA schools differed markedly from both the non-TFA schools and the final sample of TFA schools. Specifically, as shown in Table 5, only about 54% of TFA schools had students enrolled in mathematics courses, compared with about 99% of non-TFA schools—making the schools employing TFA teachers substantially different in this regard from schools not employing TFA teachers. Our analysis of Texas school data reveals that schools with sixth-grade mathematics classes are, in fact, substantially different in important ways from schools with no mathematics classes.<sup>17</sup> For example, schools with mathematics classes in the sixth grade are more likely to participate in the state’s school accountability rating system. They also have a greater number of enrolled students and lower student mobility rates; in addition, they are more likely to be a traditional middle school serving grades 6 through 8. These issues should have been explored in the report to ensure they had no effect on student test scores.

**Table 5. Effect of Excluding Schools with Students not Enrolled in Mathematics Classes on the Number of Schools Included in the Sample in Math**

	Grades 4-5				Grades 6-8			
	TFA		Not TFA		TFA		Not TFA	
	N	%	N	%	N	%	N	%
<b>Original Sample</b>	81	--	385	--	113	--	375	--
<b>After Excluding Students</b>								
<b>Not Enrolled in Math</b>	43	53%	329	85%	61	54%	370	99%
<b>Absent too Many Days</b>	43	53%	329	85%	61	54%	369	98%

Notes: % refers to percent of original sample (the top row). As with table 2, the patterns in math are similar to those in reading, the reading statistics have been excluded for the sake of simplicity.

Sixth, the report never details specific information about the number or characteristics of teachers included in the final samples used in the PSM procedure and HLM analysis. Clearly, a number of schools were removed from the analysis, yet we do not know how this affected the number of TFA teachers/alumni. Moreover, we do not know how many students were taught by each teacher. Because the data were masked to comply with FERPA, some teachers had fewer students included in the analysis while other teachers had greater numbers of students in the analysis. The study does not address or explore this issue in any manner. Due to data restrictions, the study could not report on any teacher characteristics, thus we also do not know if TFA teachers/alumni were significantly

different in important ways from non-TFA teachers/alumni. These differences could have influenced the study such that the findings would have been very different.

Finally, the total sample of schools in the TFA alumni analysis was extremely small. For grades 4 and 5, only 14 TFA alumni schools were included; for grades 6 through 8, only 12 TFA alumni schools were included. With such a small sample, the inclusion of a few outlier schools with very high achievement could skew the results for TFA alumni schools. The study does not explore the possible effects of such a small sample, but it does state readers should interpret the results for the TFA alumni with caution. Importantly, two of the three statistically significant results were from the middle grades alumni analysis—the very analysis with the smallest sample of schools and thus the most susceptible to the effects of outliers.

### **Critique of Matching Procedure**

As with the sampling procedure, there are a number of serious concerns about the matching procedures employed in the study. Our critiques are contained in three different broad areas: critique of the variables included in the PSM procedure, critique of the variables omitted from the PSM procedure, and critique of the results of the PSM procedure

#### ***Critique of the Variables Included in the PSM Procedure***

There were several important issues related to the variables included in the PSM analysis. First, the demographic and achievement characteristics employed in the matching of schools were problematic. With respect to demographic characteristics, the use of the percentage of students participating in the federal free- and reduced-price meal program (FARM) masks the variation in the percentage of students participating in the free-lunch side of the program. Other research has found the percentage of *free-lunch* students has a more significant impact on student scores than the percentage of students participating in the *reduced-price* lunch program<sup>18</sup>. Since schools with similar FARM percentages could have dramatically different percentages of students participating in the free meal portion of the program, the results of the study are likely inaccurate because of the failure to use the more important measure, which was publicly available on the Texas Education Agency (TEA) website.

The school achievement characteristic used to match schools was similarly problematic. Specifically, the report relies on the percentage of students passing the TAKS as the matching criteria. Because of the binary nature of the passing metric, however, the percentage passing masks real differences in achievement levels as measured by average scale scores<sup>19</sup>; two schools with identical percentages of students passing the TAKS mathematics test could have dramatically different average scale scores. Again, given the impact of peer effects on achievement, the failure to accurately account for the true level of achievement in a school likely resulted in inaccurate results.

A final issue related to the variables included in the PSM was the matching of novice TFA teachers to novice non-TFA teachers and TFA alumni to experienced non-TFA alumni teachers. The report relied on a variable of educator experience to classify teachers as novice (two or fewer years of teaching) or experienced (greater than two years of teaching). This classification, provided by TEA, is problematic, however. We investigated the accuracy of TEA's identification of novice teachers in the four regions of the state included in the study. Of the 16,107 teachers identified as being novice in 2010-11, we found 5% had more than two years of teaching experience in Texas public schools based on teacher employment records from the 1987-88 through 2009-10 school years<sup>20</sup>—meaning they were not, in fact, novice teachers. The error rate was even greater for those identified by TEA as “beginning” teachers—more than 10% actually had documented teaching experience in Texas public schools prior to 2010-11. In addition, we also found that another 5% of the “novice” teachers were either designated as beginning teachers or had four or more years of experience in 2011-12, thus raising questions about the accuracy of the identification of novice teachers. With a large sample of teachers, this would likely not be an important issue. Remember, however, that the alumni comparisons relied on very, very small numbers of teachers. Thus, the mis-identification of only a few teachers could influence the results. We do not know how the error rate was distributed across TFA and non-TFA teachers, and the report did not examine this possibility. Lacking that information, we are unable to hypothesize how these types of errors might affect the estimates of the effectiveness of TFA teachers and alumni. Given the importance of controlling for years of experience when estimating teacher effects on student test scores, the report's failure to ensure teacher experience was accurately identified is a serious cause for concern about the accuracy of its results.

### *Critique of the Variables Excluded in the PSM Procedure*

Using PSM as a strategy to control for important factors related to the achievement of students requires including *all* factors in the matching procedure that influence both student assignment to TFA teachers/alumni *and* student test scores. Then the PSM can approximate a random experiment, and any difference in test scores between the two groups could then be attributed to some TFA effect. Below, we make the case below that not all relevant variables were included in the PSM procedure for three reasons: data unavailability, decisions by the researchers involved in the study, and data that is simply not measured.

**Unavailability of Data.** Even though some important data was collected by the state and theoretically available, the data was not made available for the study, primarily due to FERPA considerations by the state. For example, the researchers did not have access to accurate information on the actual years of experience of the teacher or teacher characteristics such as certification test scores, route to certification, or selectiveness of undergraduate institution. Ideally, the study would have matched teachers on these characteristics to control for the influence of such factors on student test scores. In fact, the TEA did not make the teacher data available for the study, so there was no matching of any teacher characteristics. Note that the researchers could have at least examined the

characteristics across TFA and non-TFA schools using publicly available data, but chose not to do so.

**Decisions by Researchers.** A number of different variables potentially associated with student test scores were available to use in the study, but were not employed in the matching procedure or in the HLM analysis. First, there is no evidence the study's matching procedure included a school's grade span, charter status, participation in the regular state accountability system, percentage of mobile students, percentage of students

*While the report is interested in teacher effects, the study had access to only individual-, student-, and school-level data.*

at risk, average scale score on TAKS, or percentage of students not taking the standard TAKS.<sup>21</sup> All of these factors were publicly available on the TEA website and would be important to employ in constructing an appropriate set of comparison schools. The exclusion of readily available variables calls into question the accuracy of the matching procedure and, hence, the conclusions of the report.

Relatedly, the PSM procedure failed to control for length of school day, week, or year. Normally, one could assume instructional time is equal across schools. However, a number of high-profile charter schools such as KIPP and YES Preparatory utilize longer days, weeks, or years. Further, five middle schools and 12 elementary schools in Houston Independent School District are part of the Apollo 20 effort that requires schools to use extended time strategies. As noted in the appendix of the report (p. 106), the greatest proportion of students included in the analysis was in the Houston region, and this region had at least 20 elementary schools and 28 middle schools using extended time strategies. Since Apollo 20 schools, KIPP charter schools, and YES Preparatory schools seek and often do employ TFA teachers, the report should have either excluded extended time schools or included this factor in the PSM procedure. Such data is publicly available, and the report could have easily identified such schools. Failure to include it could have very likely led to the TFA effects being confounded with extended time effects. In fact, the evaluation of the impact of Apollo 20 schools<sup>22</sup> on TAKS scores found very similar effect sizes for middle school mathematics as the Edvance report; that study concluded extended time was one reason for the effect of Apollo schools on student achievement. Thus, the TFA "teacher" effect found in this report may, in fact, be explained by the relationship between extended time schools and higher student TAKS scores.

**Unmeasured Data.** If factors associated with student test scores are not measured, then clearly they would be unavailable to use in the PSM procedure. If these unmeasured characteristics—such as school climate, strong school leadership, teacher qualifications, or teacher turnover—are not randomly distributed across both groups of students (those taught by TFA teachers/alumni and those *not* taught by TFA teachers/alumni), then the PSM will yield inaccurate results unless appropriate statistical approaches in the HLM are employed. Given that only a handful of characteristics were included in the PSM match

and that the HLM analysis did not use student- or school-fixed effects, the reported effects of TFA are very likely to be confounded with important, but unmeasured, characteristics of students, teachers, and schools.

### ***Critique of the Results of the PSM Procedure***

One important overall critique is that the PSM failed to achieve comparability in the middle grades between the TFA and non-TFA schools on multiple factors: the proportion of students who were economically disadvantaged, percentages of African American and Hispanic Students, and percentage of Limited English Proficiency students. The report does note the lack of comparability in the PSM procedure and appropriately includes the factors that were not matched well in the PSM procedure in the subsequent HLM analysis. The extent to which this inclusion controlled for the poorly matched factors is unclear, as the authors do not describe any additional analysis that would assure the reader of the effectiveness of the approach. If the approach failed, then the estimates of the effectiveness of TFA teachers and alumni would be confounded. While the authors do acknowledge these problems, they do not adequately explain these limitations in relation to their main policy findings.

### **Statistical Analysis**

As with the sampling and PSM procedures, there were also multiple issues with the HLM analysis employed in the study.

One major problem with the HLM analysis is that it did not adjust for the measurement error in TAKS scores. With any test score, there is an amount of uncertainty, or measurement error, present. Failure to adjust such models for measurement error often leads to inaccurate results. This study, then, likely has inaccurate estimates of any TFA effect on student test scores. Even if not included in the model, the study should have at least mentioned the issue of measurement error.

Second, the report takes a problematic approach to derive the covariate it uses to control for prior student scores. In an appendix footnote on page 88, the report states that the covariate was the average of all prior scores in a particular subject area. The average, though, masks differences in changes in scale scores over time. For example, suppose we have two students whose eighth-grade math score was 900 and whose average math score from the three prior grades was 700. Let us also suppose student A had prior scores of 700, 700, and 700 in grades 5, 6, and 7, while student B had prior scores of 600, 700, and 800. Both students would have the same average score, but student B would arguably be much more likely to be expected to obtain a score of 900 in the eighth grade than student A. By not actually controlling for specific prior scores for each year or for the *rate* of student progress, the model likely masks differences among students and among schools, producing inaccurate results.

Third, as noted above, the PSM procedure could not control for unmeasured factors. This issue could have been rectified in the HLM analysis by employing school-fixed effects as a means of controlling for the unobserved characteristics of schools. The report, however, did not employ such a strategy and did not even mention the issue of potential unobserved characteristics that could influence the results. Nor did the study conduct additional sensitivity analyses to allay concerns about unobserved characteristics influencing the outcomes. The failure to even mention or investigate the possibility of unmeasured characteristics influencing the outcomes of the analysis calls into question the accuracy of the results of the HLM analysis.

Fourth, the way in which teaching experience was included in the analysis may be problematic. The report defined novice teachers as teachers within their first two years of teaching and experienced teachers as those with more than two years of experience. Given that most research suggests a fairly steep learning curve over at least the first three years of teaching experience,<sup>23</sup> differences in the number and percentage of students taught by teachers in their first and second years of teaching between TFA and non-TFA teachers, could cause the results of the HLM estimates to be inaccurate. The same issue would apply to TFA alumni and experienced non-TFA teachers, since there was no matching or control variable employed for years of experience. While the effects of additional years of experience are greatest in the initial years of teaching, some research has found that improvements in effectiveness in improving test scores increases over a longer time period.<sup>24</sup> Experience levels, then, could have varied dramatically between the two groups of teachers in ways not captured by the simple division between teachers in their first two years and those more than two years of experience. The extent to which actual years of experience are comparable between TFA and non-TFA teachers is unclear because the study did not report any information about the distribution of teachers or students by the years of experience of the teachers.

Finally, the report did not contextualize the findings within the broader set of issues surrounding TFA. In particular, the study did not examine or note the potential impact of the attrition of TFA teachers. This is particularly salient, as teacher turnover at the school level depresses student achievement.<sup>25</sup> Despite repeated claims that TFA teachers as a group do not have a high rate of attrition from teaching, various reports repeatedly find TFA teachers *do* have a very high turnover rates; one such report was conducted in Texas.<sup>26</sup> Even if we accept the findings of the report, the failure to assess the negative impact of high rates of TFA teacher attrition on student test scores means the findings may be overstated.

## **VI. Review of the Validity of the Findings and Conclusions**

The validity of the report's conclusions is questionable for numerous reasons. The overall goals of the report were to isolate the effect of TFA teachers/alumni and determine the size of these TFA teacher/alumni effects on student test scores. To accomplish this, the study employed PSM to isolate the TFA effect, but the estimated effect is mostly likely

confounded with other factors influencing test scores because of (1) key omitted variables, (2) problems with the included variables, and (3) failure of the PSM procedure to include all schools in the matching process and subsequent failure to achieve comparability at the school level. To estimate the size of the TFA teacher/alumni, effects the authors employed HLM. Issues with measurement error and aggregation of prior years of student test scores, as well as problems related to the omission of key variables akin to those omitted from the PSM, suggest the size of the TFA effects are likely different than what is stated in the report.

The issues of the attribution and size of the effects are further complicated by (a) a mismatch between the report's intended and actual-level analyses and (b) unclear sample characteristics.

The report's research questions and subsequent interpretations of the findings do not reflect the nature of the authors' data. While the report is interested in *teacher* effects, the study had access to only *individual-*, *student-*, and *school-*level data. Situations such as this are common and understandable; however, the limitations of the data should have been explicitly acknowledged, and the authors would have been well served in reformulating their research questions and interpretations at the school level. It is unclear the degree to which the report's findings at the campus level translate to the effectiveness of teachers. The second issue that cuts across both the attribution and estimation of the size of the effects is that of sample size. The exact characteristics of the samples used are unclear due to numerous exclusions, which were often referenced only in footnotes. The degree to which the samples are representative of the population of Texas students taught by, and schools employing, TFA teachers or alumni is not included in the report. Thus, the attribution of the reported effects to TFA teachers and TFA alumni stands on extremely weak conceptual and empirical grounds.

There are multiple methodological issues with the report's PSM and HLM procedures. These issues are not trivial—they are likely to have caused the final results of the HLM analysis to be inaccurate and to have led to improper conclusions. While the report finds some effects large enough to be relevant to policy, the errors in its analysis make attributing the effects to TFA teachers impossible. And by ignoring repeated reports of high TFA turnover rates, the report fails to assess how high TFA attrition rates might affect student test scores, increasing the chances the findings are overstated.

Perhaps most importantly, the report's hierarchical linear model estimates *campus- or school-* level effects rather than *teacher-*level effects. This issue is similar to the issues surrounding the estimation of the PSM at the campus level, and produces differences in student test scores between campuses with at least one TFA teacher and campuses without. This stands at odds with the report's express research questions, which are about the effect of TFA *teachers* on student test score performance. It is unclear the degree to which the report's findings at the campus level translate to the effectiveness of teachers.

## VII. Usefulness of the Report for Guidance of Policy and Practice

As documented in detail above, the report has serious flaws that jeopardize all of the conclusions related to the effects of TFA teachers on student test scores. While the report concludes TFA alumni cause higher performance on the TAKS math and reading assessments in middle grades, and novice TFA teachers cause higher performance on the TAKS mathematics assessments in the middle grades, the study only indirectly investigates the impact of TFA teachers on student test scores. Thus, conclusions about actual teacher effectiveness are misplaced. Compounding this issue, numerous methodological issues render its conclusions unusable for policymakers.

What most parents and members of the public expect from teachers and schools far exceeds test scores on multiple-choice tests. Tests such as TAKS assess only a limited amount of information about what a student knows and can do. A more complete assessment of the effect of TFA teachers would examine how TFA teachers influence other important outcomes, such as a student's ability to communicate, collaborate, and reflect on her or his own learning—the skills needed in the 21<sup>st</sup> century.

The report's main benefit is to inform future investigations of TFA teachers and student test scores. It suggests future investigations of TFA would be well served by understanding the characteristics of TFA alumni, whose campuses show large effects, as well as by focusing on in-depth evaluations of TFA in grades 6 to 8, in which the report did find effects. Such evaluations must rely on complete sets of students in classrooms taught by TFA teachers matched to similar classrooms taught by non-TFA teachers within the same school. The conclusion that TFA teachers have a positive impact on student TAKS test scores in mathematics and reading is simply not adequately supported by this report, and it should not be considered solid evidence of any such effect.

## Notes and References

---

1 For example, see the following as well as many other studies:

Lankford, H., Loeb, S. & Wyckoff, J. (2002). “Teacher Sorting the Plight of Urban Schools: A Descriptive Analysis.” *Education Evaluation and Policy Analysis*, 24(1), 37-62.

Loeb, S. & Reigninger, M. (2004). *Public policy and teacher labor markets: What we know and why it matters*. The Education Policy Center at Michigan State University, April 2004.

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

Darling-Hammond, L., & Green, J. (1990). Teacher quality and equality. In J. Goodlad & P. Keatings (Eds.), *Access to knowledge: An agenda for our nation's schools* (pp. 237-259). New York, NY: College Entrance Examination Board.

2 Turner, H., Goodman, D., Adachi, E., Brite, J., & Decker, L. (2012). *Evaluation of Teach For America in Texas schools*. San Antonio, TX: Edvance Research, Inc. Retrieved April 5, 2013, from [http://www.edvanceresearch.com/Teach\\_for\\_America\\_Texas\\_Schools.htm](http://www.edvanceresearch.com/Teach_for_America_Texas_Schools.htm).

3 The report also explores the characteristics of campuses that employed TFA teachers and students taught by TFA teachers. In addition the report also examines how the effect of TFA teachers varies by Texas TFA region and teacher experience. While we do not address these findings in the paper, we found these issues to be problematic as well.

4 The report never clearly describes whether only elementary schools with the same grad span were used in the analysis for grades 4 and 5 nor does the report detail which schools were used to examine grades 6-8.

5 TFA alumni are defined in the report as former TFA members who are still employed in a teaching position after completion of their two-year TFA commitment to teach. TFA novices are defined as teachers who are still completing their two-year TFA assignment.

6 Experienced non-TFA teachers are defined as those teachers with more than two years of teaching experience and who were not members of TFA at any point in time.

7 The results of the elementary grade level regressions for TFA-Alumni were presented with a set of exploratory analysis (see p. 64 as well as Appendix F) while the remaining six regressions were presented separately in the confirmatory analysis section.

8 Propensity score matching (PSM) is a statistical procedure that allows a research to match students, teachers, or schools based on a set of observable characteristics chosen by the researcher. PSM procedures results in accurate matches if and only if all relevant factors were measured and included in the matching procedure. If important factors were mis-measured or simply not measured at all, then the PSM procedure will not produce accurate matches. If PSM produces accurate matches, then differences between the matched students, teachers or schools represent the effect of the intervention in question—in this case the effect of TFA teachers.

9 See Appendix C-1 to C-4 in the report for greater detail.

10 Decker, P.T., Mayer, D.P., & Glazerman, S. (2004). *The effects of Teach For America on students: Findings from a national evaluation*. Mathematica Policy Research, Inc. Retrieved April 8, 2013, from <http://www.mathematica-mpr.com/publications/pdfs/teach.pdf>.

11 Downing, M. (2011, February 1). Are Teach For America teachers better? \$39k later, we still don't know (blog entry). *Hairballs* blog, Houstonpress.com . Retrieved March 24, 2013, from [http://blogs.houstonpress.com/hairballs/2011/02/\\_anyone\\_whos\\_been\\_holding.php](http://blogs.houstonpress.com/hairballs/2011/02/_anyone_whos_been_holding.php).

12 Koretz, D. (2008). *Measuring Up: What Educational Testing Really Tells Us*. Cambridge, Massachusetts: Harvard University Press.

13 Decker, P.T., Mayer, D.P., & Glazerman, S. (2004). *The effects of Teach For America on students: Findings from a national evaluation*. Washington, DC: Mathematica Policy Research, Inc. Retrieved April 8, 2013, from <http://www.mathematica-mpr.com/publications/pdfs/teach.pdf>.

14 Decker, P.T., Mayer, D.P., & Glazerman, S. (2004). *The effects of Teach For America on students: Findings from a national evaluation*. Washington, DC: Mathematica Policy Research, Inc. Retrieved April 8, 2013, from <http://www.mathematica-mpr.com/publications/pdfs/teach.pdf>.

15 The report does not make clear whether only students with all four sets of test scores were included in the analysis or some statistical procedure was used to impute scores, or whether some statistical technique was used to adjust the results for missing data.

16 The report does not make clear whether only students with all four sets of test scores were included in the analysis or some statistical procedure was used to impute scores, or whether some statistical technique was used to adjust the results for missing data.

17 Schools with no mathematics classes at a grade level still offer mathematics. However, mathematics and all other subject areas are taught by one teacher to a classroom of students. In such schools, students remain in the same classroom with the same teacher for core subject areas.

18 For a discussion and links to peer-reviewed research on the topic, see

Baker, B. D. (2011, April 27). Research warning label: Analysis contains inadequate measurement of student poverty (blog post). *School Finance 101*. Retrieved April 8, 2013, from <http://schoolfinance101.wordpress.com/2011/04/27/research-warning-label-analysis-contains-inadequate-measurement-of-student-poverty/>.

19 Koretz, D. (2008). *Measuring Up: What Educational Testing Really Tells Us*. Cambridge, MA: Harvard University Press.

20 This example is based on an analysis by Ed Fuller of publicly available data on teachers, their years of experience, and employing district. The dataset used in the analysis was created by matching teacher employment data from 1988 through 2010 to the 2011 teacher employment file. Results are available upon request.

21 This is a different measure than the percentage of students in special education. Many students take a non-standard TAKS, but are not identified as special education, For example, student with 504 status can take a non-standard TAKS and such student are not identified as special education.

22 Fryer, R.G., Jr. (2011). *Creating “no excuses” (traditional) public schools: Preliminary evidence from an experiment in Houston* (Working Paper No. 17494). Cambridge, MA: National Bureau of Economic Research.

23 For a review of many recent papers on this topic, see

Kukla-Acevedo, S. (2012). Do teacher characteristics matter? New results on the effects of teacher preparation on student achievement. *Economics of Education Review*, 28(1):49-57.

24 Kukla-Acevedo, S. (2012). Do teacher characteristics matter? New results on the effects of teacher preparation on student achievement. *Economics of Education Review*, 28(1):49-57.

25 Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2006). How changes in entry requirements alter the teacher workforce and affect student achievement. *Education Finance and Policy*, 1(2), 176–216.

Ronfeldt, M., Loeb, S., & Wyckoff, J. (2013). How teacher turnover harms student achievement. *American Educational Research Journal*, 50(1), 4-36.

26 For example, see the following reports:

Fuller, E.J. & Nolen, A. (2010, December). *An emerging understanding of the Arkansas teacher pipeline: An exploratory analysis using state data bases and teacher survey*. Little Rock, AR: Arkansas Partnership for Teacher Quality.

Ware, A., LaTurner, J. A., Parsons, J., Okulicz-Kozaryn, A., Garland, M., & Klopfenstein, K. (2011). *Teacher preparation programs and Teach For America research study*. Dallas, TX: The University of Texas at Dallas, Education Research Center.

DOCUMENT REVIEWED:

**Evaluation of Teach for America in Texas Schools**

AUTHORS:

Herbert M. Turner, David Goodman, Eishi Adachi, Jessica Brite, and Lauren E. Decker

PUBLISHER/THINK TANK:

Edvance Research, Inc.

DOCUMENT RELEASE DATE:

March 4, 2013

REVIEW DATE:

April 9, 2013

REVIEWER(S):

Edward J. Fuller, Penn State University  
Nathan Dadey, University of Colorado Boulder

E-MAIL ADDRESS:

[ejf20@psu.edu](mailto:ejf20@psu.edu)

PHONE NUMBER:

(814) 865-2233

SUGGESTED CITATION:

Fuller, E.J & Dadey, N.D. (2013). *Review of "Evaluation of Teach for America in Texas Schools."* Boulder, CO: National Education Policy Center. Retrieved [date] from <http://nepc.colorado.edu/thinktank/review-evaluation-TFA-texas>.