WWC Intervention Report

A summary of findings from a systematic review of the evidence

ies INSTITUTE OF EDUCATION SCIENCES

Teacher Training, Evaluation, and Compensation August 2016

Teach For America

Program Description[1]

Teach For America (TFA) is a highly selective route to teacher certification that aims to place non-traditionally trained teachers in high-need public schools. Many *TFA* teachers hold bachelors' degrees from selective colleges and universities, in fields outside of education.[2] *TFA* teachers commit to teach for at least 2 years. *TFA* teachers receive 5–7 weeks of in-person training over the summer before they begin teaching, then continue to receive professional development and one-on-one coaching from *TFA* while teaching, in addition to support provided by their schools and districts.[3] As full-time employees of the public schools where they work, *TFA* teachers receive the same salary and benefits as other first- or second-year teachers in their school or district.

Research[4,5]

The What Works Clearinghouse (WWC) identified seven studies of teachers trained through *TFA* that both fall within the scope of the Teacher Training, Evaluation, and Compensation topic area and meet WWC group design standards.[6] Three studies meet WWC group design standards without reservations, and four studies meet WWC group design standards with reservations. Together, these studies included more than 65,324 students in grades pre-K–12, in geographically diverse states and districts.[7]

The WWC considers the extent of evidence for teachers trained through *TFA* on the academic achievement of students in grades pre-K–12 to be medium to large for two student outcome domains—mathematics achievement and English language arts achievement—and small for two student outcome domains—science achievement and social studies achievement. There were no studies that meet WWC group design standards in the two other student outcome domains and 11 teacher outcome domains.[8] (See the Effectiveness Summary on p. 6 for more details of effectiveness by domain.)

Effectiveness

TFA teachers were found to have positive effects on mathematics achievement, potentially positive effects on science achievement, and no discernible effects on social studies achievement and English language arts achievement for students in grades pre-K–12.

Report Contents

This intervention report presents findings from a systematic review of *TFA* conducted using the WWC Procedures and Standards Handbook, version 3.0, and the Teacher Training, Evaluation, and Compensation review protocol, version 3.2.

Table 1. Summary of findings[9]

Outcome domain	Rating of effectiveness	Improvement index *(percentile points)*		Number of studies	Number of students[a]	Extent of evidence
		Average	Range			
Mathematics achievement	Positive effects	+4	–1 to +8	6	65,324	Medium to large
Science achievement	Potentially positive effects	+7	na	1	36,104	Small
Social studies achievement	No discernible effects	+3	na	1	6,051	Small
English language arts achievement	No discernible effects	+1	–2 to +2	5	53,595	Medium to large

na = not applicable

[a] The reported sample sizes may count some individual students more than once because some studies examined data from multiple school years.

Program Information

Background

TFA was established in 1990 by Wendy Kopp, who currently serves on the organization's board of directors. The program is administered by the nonprofit organization Teach For America, Inc. Address: 25 Broadway, 12th Floor, New York, NY 10004. Web: www.teachforamerica.org. Telephone: (212) 279-2080.

Program details

TFA places recent college graduates and professionals in schools in low-income communities and requires its teachers to commit to at least 2 years of teaching. The program refers to its teachers as "corps members" during their 2-year commitment. *TFA* provides corps members with training and support.

TFA's highly selective admission process typically involves an online application that includes short-answer questions, followed by a telephone interview, and then a full-day, final interview. For the 2015 cohort of new corps members, *TFA* admitted 15% of applicants. To be eligible for admission, an applicant must be a college graduate with either an undergraduate grade point average (GPA) of at least 2.50 or a graduate school GPA of at least 3.50 on a 4.00 scale. Most *TFA* recruits are recent graduates from selective colleges and universities who majored in a field other than education. In 2015, about 30% of participants worked full-time before joining *TFA*.

Training is provided during the summer prior to beginning teaching. The training varies by region but usually includes an induction to the region in which they will teach, a residential training institute, and a regional orientation. During a regional induction that typically lasts about 5 days, corps members attend required sessions and have the opportunity to familiarize themselves with the location in which they will teach in the fall. The 5- to 7-week training institute includes: (a) teaching summer school in a real classroom under the supervision of a regular classroom teacher, typically for at least 2 hours per day; (b) being observed by and receiving feedback from *TFA* instructional coaches and experienced district teachers; (c) participating in small-group sessions to practice teaching, reflect on experiences and feedback, and analyze student progress; (d) receiving instruction from *TFA* instructional coaches in lesson planning clinics; and (e) completing coursework in instructional planning and delivery, classroom management, diversity, literacy development, and a *TFA* philosophical framework that emphasizes key leadership principles. The institute is followed by 1–2 weeks of regional orientation that includes sessions aimed at helping corps members establish student achievement goals, develop short- and long-term lessons plans, use data, understand their community, and build relationships with students.

TFA corps members receive ongoing support and professional development during their 2-year teaching commitment. A *TFA* staff member known as a "manager of teacher leadership and development" conducts classroom observations and provides coaching intended to help corps members improve their instructional practice and the academic achievement of their students. In addition to one-on-one coaching, corps members may meet in regional learning teams led by highly effective teachers to share best practices with other teachers in their subject area or grade. *TFA* also provides its teachers with toolkits that include sample tests and teaching resources tailored to the teacher's grade, subject area, and district. Regional *TFA* staff work with corps members to assist them in completing state certification requirements during their 2-year teaching commitment.

Cost

As of its 2015 fiscal year, *TFA* spent approximately $65,000 per corps member over the recruitment year and 2 years of teaching, with most costs being related to training and support. During summer training, corps members receive room and board. School districts pay a per-corps-member fee each school year. The average fee paid per first-year corps member by districts was $3,283 in *TFA*'s 2015 fiscal year. *TFA* teachers are regular full-time employees of their school districts; they apply for open teaching positions and receive the starting salary and benefits of similarly qualified teachers in the school district and, where applicable, are part of collective bargaining agreements.

Research Summary

The WWC identified 24 eligible studies that investigated the effects of *TFA* teachers on the academic achievement of students in elementary, middle, and high school.[10] An additional 21 studies were identified but do not meet WWC eligibility criteria for review in this topic area. Citations for all 45 studies are in the References section, which begins on p. 10.

Table 2. Scope of reviewed research

Grades	PK–12
Delivery method	Whole class
Program type	Teacher level

The WWC reviewed 24 eligible studies against group design standards. Three studies are randomized controlled trials that meet WWC group design standards without reservations, and four studies use quasi-experimental designs that meet WWC group design standards with reservations. Those seven studies are summarized in this report. The remaining 17 studies do not meet WWC group design standards.

Summary of studies meeting WWC group design standards without reservations

Clark et al. (2013) examined the effectiveness of *TFA* teachers compared to other teachers in their schools using a randomized controlled trial conducted in 45 secondary schools in 10 *TFA* regions in eight states. In each participating school, students were randomly assigned to either a math class taught by a *TFA* teacher or a similar math class taught by a teacher in the same grade who did not enter teaching through *TFA*. Most *TFA* teachers were current corps members within their 2-year teaching commitment, but some were *TFA* alumni who stayed in schools past their 2-year commitment. The mean years of teaching experience was 1.9 for *TFA* teachers and 10.1 for comparison teachers. The authors measured mathematics achievement using state-required end-of-year standardized tests for middle school students and study-administered end-of-course assessments for high school students. The analytic sample included 4,573 students (2,292 *TFA*, 2,281 comparison) in grades 6–12, who participated in either the 2009–10 or 2010–11 school year. Clark et al. (2013) also reported subgroup findings for school levels, years of teaching experience, and comparison group route to certification (traditional or alternative). These supplemental findings are reported in Appendix D and do not factor into the intervention's rating of effectiveness.[11]

Clark et al. (2015) assessed the effectiveness of elementary school *TFA* teachers compared to other teachers in their schools using a randomized controlled trial conducted in 36 schools (including traditional public schools and charter schools) in 10 *TFA* regions in 10 states. In each participating school, students in the same grade were randomly assigned to either a class taught by a *TFA* teacher or a similar class taught by a teacher who was not a *TFA* teacher. All but one of the *TFA* teachers were in their first or second year of teaching. The mean years of teaching experience was 1.7 years for *TFA* teachers and 13.7 years for comparison teachers. The authors measured mathematics and English language arts achievement using end-of-year math and reading scores from study-administered tests for lower elementary grades (pre-K–2) and from state-required assessments for upper elementary grades (3–5). The analytic samples included 2,065 students (855 *TFA*, 1,210 comparison) for the math outcome and 2,123 students (877 *TFA*, 1,246 comparison) for the reading outcome. Clark et al. (2015) also reported findings for several subgroups, including by grades, years of experience, and type of certification. These supplemental findings are reported in Appendix D and do not factor into the intervention's rating of effectiveness.[12]

Glazerman et al. (2006) conducted a randomized controlled trial in 17 elementary schools in six *TFA* regions (Baltimore, Chicago, Houston, Los Angeles, the Mississippi Delta, and New Orleans). In each participating school, students were randomly assigned within each grade to either a class taught by a *TFA* teacher or a class taught by a teacher who was not a *TFA* teacher. Most *TFA* teachers were current corps members, but some were *TFA* alumni. The median years of experience was 2 years for *TFA* teachers and 6 years for comparison teachers. The authors measured mathematics and English language arts achievement using a test they administered to students in grades 1–5 as a pretest in the fall and as a posttest in the spring. The analytic sample included 1,715 students (759 *TFA* and 956 comparison) who participated in either the 2001–02 or 2002–03 school year.

Summary of studies meeting WWC group design standards with reservations

Henry, Purtell, et al. (2014) examined the effectiveness of *TFA* teachers in North Carolina public schools during the 2005–06 through 2009–10 school years using a quasi-experimental design. The authors compared the achievement outcomes for students taught by *TFA* teachers versus students taught by "in-state public undergraduate prepared" teachers—that is, teachers who had completed their initial licensure requirements prior to beginning teaching by receiving a bachelor's degree from a North Carolina public university. All intervention and comparison teachers had less than 5 years of teaching experience. The authors measured achievement using state standardized assessments in mathematics and English language arts achievement for students in grades 3–8. End-of-course tests assessing mathematics and social studies achievement were administered to high school students.[13] Analytic sample sizes by grade level and subject area for comparisons that meet WWC group design standards are provided in Appendix A. A related publication (Henry et al., 2012) also reported findings that compared achievement outcomes for students of *TFA* teachers to outcomes for students of in-state prepared teachers during the same time period as Henry, Purtell, et al. (2014).[14] These supplemental findings are reported in Appendix D and do not factor into the intervention's rating of effectiveness.

Turner et al. (2012) examined the effectiveness of *TFA* teachers in four Texas regions (Dallas–Fort Worth, Houston, the Rio Grande Valley, and San Antonio) using a quasi-experimental design. The study authors identified schools that employed at least one *TFA* teacher. The authors used state-based mathematics and English language arts achievement as outcomes. The authors analyzed two sets of teacher comparisons: (a) *TFA* corps members versus novice comparison teachers, who had less than 3 years of teaching experience; and (b) *TFA* alumni versus experienced comparison teachers, who had 3 or more years of experience. The WWC based its effectiveness ratings on the novice teacher comparison; analytic sample sizes by grade level and subject area are provided in Appendix A.[15] Findings from the experienced teacher comparison are reported in Appendix D and do not factor into the intervention's rating of effectiveness.

Ware et al. (2011) assessed the effectiveness of two cohorts of *TFA* teachers in four Texas school districts using a quasi-experimental design. The study authors identified students who took the English version of the state-required mathematics and English language arts assessments. The authors compared year-to-year passing rate gains for students of two cohorts of *TFA* teachers to year-to-year passing rate gains for students of non-*TFA* teachers who had less than 3 years of teaching experience. Analytic sample sizes by grade level and subject area for comparisons that meet WWC group design standards are provided in Appendix A. Ware et al. (2011) also reported subgroup findings for African-American, Hispanic, and economically disadvantaged students. These supplemental findings are reported in Appendix D and do not factor into the intervention's rating of effectiveness.

Xu et al. (2011) examined the effectiveness of *TFA* high school teachers in North Carolina schools during the 2000–01 through 2006–07 school years using a quasi-experimental design. The study authors restricted the sample to teachers and students in schools that employed at least one *TFA* teacher. The authors compared achievement outcomes on state standardized tests for students of *TFA* teachers to students of non-*TFA* teachers. The authors measured science achievement using high school end-of-course exams.[16] The analytic sample for the comparison that meets WWC group design standards included 36,104 high school students (3,495 *TFA*, 32,609 comparison) for science achievement.[17] Xu et al. (2011) also reported subgroup findings based on whether teachers are licensed in the subject they teach and the specific license held. These supplemental findings are reported in Appendix D and do not factor into the intervention's rating of effectiveness.

Effectiveness Summary

The WWC review of studies of teachers trained through *TFA* for the Teacher Training, Evaluation, and Compensation topic area includes both student and teacher outcomes. The review covers six domains for student outcomes and eleven domains for teacher outcomes.[18] The seven studies of *TFA* teachers that meet WWC group design standards reported findings in four of the six domains for student outcomes: (a) mathematics achievement, (b) science achievement, (c) social studies achievement, and (d) English language arts achievement. The seven studies did not report any findings that meet WWC group design standards in the eleven domains for teacher outcomes.[19] The findings below present the authors' estimates and WWC-calculated estimates of the size and statistical significance of the effects of *TFA* teachers on students in grades pre-K–12. Additional comparisons are presented as supplemental findings in Appendix D. The supplemental findings do not factor into the intervention's rating of effectiveness. For a more detailed description of the rating of effectiveness and extent of evidence criteria, see the WWC Rating Criteria on p. 45.

Summary of effectiveness for the mathematics achievement domain

Table 3. Rating of effectiveness and extent of evidence for the mathematics achievement domain

Rating of effectiveness	Criteria met
Positive effects *Strong evidence of a positive effect with no overriding contrary evidence.*	In four of the six studies, the estimated impact of *TFA* teachers on *mathematics achievement* was positive and statistically significant; two of these studies meet WWC group design standards without reservations. In the other two studies, the estimated impact of *TFA* teachers was neither statistically significant nor large enough to be substantively important.
Extent of evidence	**Criteria met**
Medium to large	Six studies that included 65,324[a] students reported evidence of effectiveness in the *mathematics achievement* domain.[b]

[a] The reported sample sizes may count some individual students more than once because some studies examined data from multiple school years.
[b] The number of schools included in each study were 45 for Clark et al. (2013), 36 for Clark et al. (2015), 17 for Glazerman et al. (2006), and 493 for Turner et al. (2012). Henry, Purtell, et al. (2014) and Ware et al. (2011) did not report the number of schools included in their studies.

Three studies that meet WWC group design standards without reservations and three studies that meet WWC group design standards with reservations reported findings in the mathematics achievement domain.

Clark et al. (2013) examined one outcome in the mathematics achievement domain: the authors created a standardized achievement measure (called a *z*-score) based on two different assessments (state-required assessments for students in grades 6–8 and Northwest Evaluation Association [NWEA] end-of-course assessments for students in grades 9–12). The authors reported, and the WWC confirmed, a positive and statistically significant difference between the *TFA* group and the comparison group. The WWC characterizes this study finding as a statistically significant positive effect.

Clark et al. (2015) examined one outcome in the mathematics achievement domain: the authors created a standardized achievement measure (called a *z*-score) based on two different assessments (Woodcock-Johnson tests for students in grades pre-K–2 and state-required assessments for students in grades 3–5). The authors reported, and the WWC confirmed, that the difference between the *TFA* group and the comparison group was not statistically significant. According to WWC criteria, the effect size was not large enough to be considered substantively important (i.e., an effect size of at least 0.25). The WWC characterizes this study finding as an indeterminate effect.

Glazerman et al. (2006) examined one outcome in the mathematics achievement domain: the Iowa Test of Basic Skills (ITBS) mathematics subtest. The authors reported, and the WWC confirmed, a positive and statistically significant difference between the *TFA* group and the comparison group. The WWC characterizes this study finding as a statistically significant positive effect.

Henry, Purtell, et al. (2014) examined two outcomes in the mathematics achievement domain: (a) the North Carolina end-of-grade mathematics assessments, analyzed separately for elementary grades (3–5) and middle school grades (6–8); and (b) the North Carolina end-of-course high school mathematics assessments. The authors reported positive and statistically significant differences between the *TFA* group and the comparison group for all three grade level samples. The WWC determined that the differences for each grade level sample were not statistically significant after correcting for clustering. However, the WWC found that the average effect size across both grade spans was positive and statistically significant. The WWC characterizes these study findings as a statistically significant positive effect.

Turner et al. (2012) examined one outcome in the mathematics achievement domain: the Texas Assessment of Knowledge and Skills (TAKS) mathematics. The authors analyzed students in elementary grades (4–5) separately from students in middle grades (6–8). The authors reported, and the WWC confirmed, that the difference between the *TFA* group and the comparison group was not statistically significant in the elementary grades sample but was statistically significant in the middle grades sample. However, the WWC-calculated average effect size pooled across the elementary and middle school samples is positive and statistically significant; therefore, the WWC characterizes these study findings as a statistically significant positive effect.

Ware et al. (2011) examined one outcome in the mathematics achievement domain: the TAKS mathematics. The authors analyzed students in grades 3–8 and grades 9–11 in 2009–10. The authors reported that the difference between the *TFA* group and comparison group was not statistically significant for the grades 3–8 sample but was positive and statistically significant for the grades 9–11 sample. The WWC determined that the finding for students in grades 9–11 was not statistically significant after correcting for clustering. The WWC-calculated average effect size pooled across the elementary and middle school samples was not large enough to be considered substantively important. The WWC characterizes these study findings as an indeterminate effect.

Thus, for the mathematics achievement domain, four studies showed statistically significant positive effects, and two studies showed indeterminate effects. This results in a rating of positive effects, with a medium to large extent of evidence.

Summary of effectiveness for the science achievement domain

Table 4. Rating of effectiveness and extent of evidence for the science achievement domain

Rating of effectiveness	Criteria met
Potentially positive effects *Evidence of a positive effect with no overriding contrary evidence.*	In the one study, the estimated impact of *TFA* teachers on *science achievement* was positive and statistically significant.
Extent of evidence	**Criteria met**
Small	One study that included 36,104 students reported evidence of effectiveness in the *science achievement* domain.[a]

[a] The authors (Xu et al., 2011) did not report the number of schools included in the study.

One study that meets WWC group design standards with reservations reported findings in the science achievement domain.

Xu et al. (2011) examined one outcome in the science achievement domain: the authors created a standardized achievement measure (called a *z*-score) based on different North Carolina end-of-course high school science assessments. The authors reported, and the WWC confirmed, a positive and statistically significant difference between the *TFA* group and the comparison group. The WWC characterizes this study finding as a statistically significant positive effect.

Thus, for the science achievement domain, one study showed a statistically significant positive effect. This results in a rating of potentially positive effects, with a small extent of evidence.

Summary of effectiveness for the social studies achievement domain

Table 5. Rating of effectiveness and extent of evidence for the social studies achievement domain

Rating of effectiveness	Criteria met
No discernible effects *No affirmative evidence of effects.*	In the one study, the estimated impact of *TFA* teachers on *social studies achievement* was neither statistically significant nor large enough to be substantively important.
Extent of evidence	**Criteria met**
Small	One study that included 6,051[a] students reported evidence of effectiveness in the *social studies achievement* domain.[b]

[a] The reported sample sizes may count some individual students more than once because some studies examined data from multiple school years.
[b] The authors (Henry, Purtell, et al., 2014) did not report the number of schools included in the study.

One study that meets WWC group design standards with reservations reported findings in the social studies achievement domain.

Henry, Purtell, et al. (2014) did not find a statistically significant effect of *TFA* teachers on social studies achievement using North Carolina end-of-course high school assessments. The WWC-calculated average effect size was not large enough to be considered substantively important. The WWC characterizes this study finding as an indeterminate effect.

Thus, for the social studies achievement domain, one study showed an indeterminate effect. This results in a rating of no discernible effects, with a small extent of evidence.

Summary of effectiveness for the English language arts achievement domain

Table 6. Rating of effectiveness and extent of evidence for the English language arts achievement domain

Rating of effectiveness	Criteria met
No discernible effects *No affirmative evidence of effects.*	In the five studies, the estimated impact of *TFA* teachers on *English language arts achievement* was neither statistically significant nor large enough to be substantively important.
Extent of evidence	**Criteria met**
Medium to large	Five studies that included 53,595[a] students reported evidence of effectiveness in the *English language arts achievement* domain.[b]

[a] The reported sample sizes may count some individual students more than once because some studies examined data from multiple school years.
[b] The number of schools included in each study was 36 for Clark et al. (2015), 17 for Glazerman et al. (2006), and 483 for Turner et al. (2012). Henry, Purtell, et al. (2014) and Ware et al. (2011) did not report the number of schools included in their studies.

Two studies that meet WWC group design standards without reservations and three studies that meet WWC group design standards with reservations reported findings in the English language arts achievement domain.

Clark et al. (2015) did not find a statistically significant effect of *TFA* teachers on English language arts achievement based on the full sample using a *z*-score to standardize the Woodcock-Johnson tests for grades pre-K–2 and the state-required assessments for grades 3–5. The WWC-calculated effect size was not large enough to be considered substantively important. The WWC characterizes this study finding as an indeterminate effect.

Glazerman et al. (2006) did not find a statistically significant effect of *TFA* teachers on English language arts achievement using the ITBS reading subtest. The WWC-calculated effect size was not large enough to be considered substantively important. The WWC characterizes this study finding as an indeterminate effect.

Henry, Purtell, et al. (2014) did not find a statistically significant effect of *TFA* teachers on English language arts achievement using North Carolina end-of-grade reading assessments for either the elementary or middle grades samples. The WWC-calculated average effect size was not large enough to be considered substantively important. The WWC characterizes these study findings as an indeterminate effect.

Turner et al. (2012) did not find a statistically significant effect of *TFA* corps members on English language arts achievement using the TAKS reading assessment for the elementary and middle grades samples. The WWC-calculated average effect size was not large enough to be considered substantively important. The WWC characterizes these study findings as an indeterminate effect. Supplemental findings for experienced teachers do not factor into the intervention's rating of effectiveness, but are presented in Appendix D. As part of these supplemental findings, Turner et al. (2012) found, and the WWC confirmed, a statistically significant positive effect of *TFA* alumni on middle grade students' English language arts achievement; students in the intervention group, who were taught by *TFA* alumni who had completed their 2-year contract assignment but continued teaching, had higher TAKS reading assessment scores than students in the comparison group, who were taught by teachers who did not participate in *TFA* and had 3 or more years of teaching experience.

Ware et al. (2011) did not find a statistically significant effect of *TFA* teachers on English language arts achievement using the TAKS English Language Arts/Reading (ELA/R) assessments for either grades 3–8 students in 2008–09 or grades 9–11 students in 2009–10. The WWC-calculated average effect size was not large enough to be considered substantively important. The WWC characterizes these study findings as an indeterminate effect.

Thus, for the English language arts achievement domain, five studies showed indeterminate effects. This results in a rating of no discernible effects, with a medium to large extent of evidence.

References

Studies that meet WWC group design standards without reservations

Clark, M. A., Chiang, H. S., Silva, T., McConnell, S., Sonnenfeld, K., Erbe, A., & Puma, M. (2013). *The effectiveness of secondary math teachers from Teach For America and the Teaching Fellows programs* (NCEE 2013-4015). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. http://files.eric.ed.gov/fulltext/ED544171.pdf.

Additional source:

Chiang, H. S., Clark, M.A., & McConnell, S. (2014). *Supplying disadvantaged schools with effective teachers: Experimental evidence on secondary math teachers from Teach For America* (Working Paper 31). Princeton, NJ: Mathematica Policy Research.

Clark, M. A., Isenberg, E., Liu, A. Y., Makowsky, L., & Zukiewicz, M. (2015). *Impacts of the Teach for America Investing in Innovation scale-up*. Princeton, NJ: Mathematica Policy Research. Retrieved from http://www.mathematica-mpr.com

Glazerman, S., Mayer, D., & Decker, P. (2006). Alternative routes to teaching: The impacts of Teach for America on student achievement and other outcomes. *Journal of Policy Analysis and Management, 25*(1), 75–96.

Additional sources:

Antecol, H., Eren, O., & Ozbeklik, S. (2013a). The effect of Teach for America on the distribution of student achievement in primary school: Evidence from a randomized experiment. *Economics of Education Review, 37*, 113–125.

Antecol, H., Eren, O., & Ozbeklik, S. (2013b). *The effect of Teach for America on the distribution of student achievement in primary school: Evidence from a randomized experiment* (Discussion Paper 7296). Bonn, Germany: IZA.

Decker, P., Mayer, D., & Glazerman, S. (2004). *Quality in the classroom: How does Teach For America measure up? Issue brief #1*. Princeton, NJ: Mathematica Policy Research. http://files.eric.ed.gov/fulltext/ED496298.pdf.

Decker, P. T., Mayer, D. P., & Glazerman, S. (2004a). *The effects of Teach for America on students: Findings from a national evaluation*. Princeton, NJ: Mathematica Policy Research. Retrieved from http://www.mathematica-mpr.com

Decker, P. T., Mayer, D. P., & Glazerman, S. (2004b). *The effects of Teach for America on students: Findings from a national evaluation* (Discussion Paper 1285-04). Madison: Institute for Research on Poverty, University of Wisconsin–Madison.

Studies that meet WWC group design standards with reservations

Henry, G. T., Purtell, K. M., Bastian, K. C., Fortner, C. K., Thompson, C. L., Campbell, S. L., & Patterson, K. M. (2014). The effects of teacher entry portals on student achievement. *Journal of Teacher Education, 65*(1), 7–23.

Additional sources:

Henry, G. T., Bastian, K. C., Fortner, C. K., Kershaw, D. C., Purtell, K. M., Thompson, C. L., & Zulli, R. A. (2014). Teacher preparation policies and their effects on student achievement. *Education Finance and Policy, 9*(3), 264–303.

Henry, G. T., Bastian, K. C., & Smith, A. A. (2012). Scholarships to recruit the best and brightest into teaching: Who is recruited, where do they teach, how effective are they, and how long do they stay? *Educational Researcher, 41*(3), 83–92.

Henry, G. T., Thompson, C. L., Bastian, K. C., Fortner, C. K., Kershaw, D. C., Purtell, K. M., & Zulli, R. A. (2010). *Portal report: Teacher preparation and student test scores in North Carolina*. Chapel Hill: Carolina Institute for Public Policy, The University of North Carolina at Chapel Hill.

Turner, H. M., Goodman, D., Adachi, E., Brite, J., & Decker, L. E. (2012). *Evaluation of Teach For America in Texas schools*. San Antonio, TX: Edvance Research, Inc.

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

Xu, Z., Hannaway, J., & Taylor, C. (2011). Making a difference? The effects of Teach For America in high school. *Journal of Policy Analysis and Management, 30*(3), 447–469.

Additional source:

Xu, Z., Hannaway, J., & Taylor, C. (2009). *Making a difference? The effects of Teach For America in high school* (Working Paper 17. Revised). Washington, DC: The Urban Institute and the National Center for Analysis of Longitudinal Data in Education Research. http://files.eric.ed.gov/fulltext/ED509654.pdf.

Studies that do not meet WWC group design standards

Boyd, D., Grossman, P., Hammerness, K., Lankford, H., Loeb, S., Ronfeldt, M., & Wyckoff, J. (2012). Recruiting effective math teachers: How do math immersion teachers compare? Evidence from New York City. *American Educational Research Journal, 49*(6), 1008–1047. The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Additional source:

Boyd, D., Grossman, P., Hammerness, K., Lankford, H., Loeb, S., Ronfeldt, M., & Wyckoff, J. (2010). *Recruiting effective math teachers: How do math immersion teachers compare? Evidence from New York City* (NBER Working Paper 16017). Cambridge, MA: National Bureau of Economic Research.

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. The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Additional source:

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2005). *How changes in entry requirements alter the teacher workforce and affect student achievement* (NBER Working Paper 11844). Cambridge, MA: National Bureau of Economic Research.

Carroll, C. A. (2013). *The influence of Teach for America on algebra I student achievement* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3594054) The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Darling-Hammond, L., Holtzman, D. J., Gatlin, S. J., & Heilig, J. V. (2005a). Does teacher preparation matter? Evidence about teacher certification, Teach for America, and teacher effectiveness. *Education Policy Analysis Archives, 13*(42), 1–47. http://files.eric.ed.gov/fulltext/EJ846746.pdf. The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Additional source:

Darling-Hammond, L., Holtzman, D. J., Gatlin, S. J., & Heilig, J. V. (2005b). *Does teacher preparation matter? Evidence about teacher certification, Teach for America, and teacher effectiveness*. Stanford, CA: Stanford University.

Hansen, M., Backes, B., Brady, V., & Xu, Z. (2014). *Examining spillover effects from Teach for America corps members in Miami–Dade County Public Schools* (CALDER Working Paper 113). Washington, DC: National Center for Analysis of Longitudinal Data in Education Research. The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Houston Independent School District Department of Research and Accountability. (2011). *Teach for America (TFA) 2009-2010*. Houston, TX: Author. The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

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. The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Additional source:

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2006). *What does certification tell us about teacher effectiveness? Evidence from New York City* (NBER Working Paper 12155). Cambridge, MA: National Bureau of Economic Research.

Klein, N. (2009). *A comparative study of self-efficacy, outcome expectancy, and retention of beginning urban science teachers* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3355425) The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Laczko-Kerr, I., & Berliner, D. C. (2002). The effectiveness of "Teach for America" and other under-certified teachers on student academic achievement: A case of harmful public policy. *Education Policy Analysis Archives, 10*(37). The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Additional source:

Laczko-Kerr, I. I. (2002). *Teacher certification does matter: The effects of certification status on student achievement* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3045652)

Mac Iver, M. A., & Vaughn, E. S., III (2007). But how long will they stay? Alternative certification and new teacher retention in an urban district. *ERS Spectrum, 25*(2), 33–44. The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Noell, G. H., & Gansle, K. A. (2009). *Teach for America teachers' contribution to student achievement in Louisiana in grades 4-9: 2004-2005 to 2006-2007*. Baton Rouge: Louisiana State University. The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Prescott, S. H. (2010). *The effects of affirmative quality feedback on low socio-economic students' zone of proximal development reading gains (ZPDRL): A causal-comparative study* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3447103) The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Raymond, M., Fletcher, S. H., & Luque, J. (2001). *Teach for America: An evaluation of teacher differences and student outcomes in Houston, Texas*. Stanford, CA: Center for Research on Education Outcomes, The Hoover Institution, Stanford University. The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Additional sources:

Luque, J. A. (2003). *Essays on economics of education* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3102286)

Raymond, M., & Fletcher, S. (2002). The Teach for America evaluation. *Education Next, 2*(1), 62–68.

Schoeneberger, J. A., Dever, K. A., & Tingle, L. (2009). *Teach For America evaluation report*. Charlotte, NC: Center for Research & Evaluation, Charlotte–Mecklenburg Schools. The study does not meet WWC group design standards because the analysis does not provide a credible measure of the effectiveness of the intervention.

Additional source:

Schoeneberger, J. A. (2011). *Teach For America evaluation report*. Charlotte, NC: Center for Research & Evaluation, Charlotte–Mecklenburg Schools.

Strategic Data Project. (2012). *SDP human capital diagnostic: Los Angeles Unified School District*. Cambridge, MA: Author, Center for Education Policy Research, Harvard University. Retrieved from http://www.gse.harvard.edu The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Tennessee State Board of Education. (2014). *2014 report card on the effectiveness of teacher training programs*. Nashville, TN: Author. The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Additional sources:

Tennessee State Board of Education. (2010). *2010 report card on the effectiveness of teacher training programs*. Nashville, TN: Author. http://files.eric.ed.gov/fulltext/ED514363.pdf.

Tennessee State Board of Education. (2011). *2011 report card on the effectiveness of teacher training programs*. Nashville, TN: Author. http://files.eric.ed.gov/fulltext/ED530920.pdf.

Tennessee State Board of Education. (2012). *2012 report card on the effectiveness of teacher training programs*. Nashville, TN: Author.

Tennessee State Board of Education. (2013). *2013 report card on the effectiveness of teacher training programs*. Nashville, TN: Author.

Urdegar, S. M. (2015). *Teach For America: An analysis of placement and impact, 2013-14*. Miami, FL: Office of Assessment, Research, and Data Analysis, Miami–Dade County Public Schools. The study does not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Additional sources:

Urdegar, S. M. (2011). *Teach For America: An analysis of placement and impact*. Miami, FL: Office of Assessment, Research, and Data Analysis, Miami–Dade County Public Schools.

Urdegar, S. M. (2013a). *Teach For America: An analysis of placement and impact, 2011-12*. Miami, FL: Office of Assessment, Research, and Data Analysis, Miami–Dade County Public Schools.

Urdegar, S. M. (2013b). *Teach For America: An analysis of placement and impact*, 2012-13. Miami, FL: Office of Assessment, Research, and Data Analysis, Miami–Dade County Public Schools.

Studies that are ineligible for review using the Teacher Training, Evaluation, and Compensation Evidence Review Protocol

Alicea, M. M. (2013). *To give and to receive: Examining feedback in three coaching dyads from the perspective of a university coach and Teach For America corps members* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3571367) The study is ineligible for review because it does not use an eligible design.

Anderson, A. (2013). Teach For America and symbolic violence: A Bourdieuian analysis of education's next quick-fix. *The Urban Review, 45*(5), 684–700. The study is ineligible for review because it does not use an eligible design.

Brandt, C. (2005). *Recruitment, retention and the effects of participation: The case of Teach For America* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3171679) The study is ineligible for review because it does not use an eligible design.

Cochran-Smith, M. (2005). Taking stock in 2005: Getting beyond the horse race. *Journal of Teacher Education, 56*(1), 3–7. The study is ineligible for review because it does not use an eligible design.

Darling-Hammond, L. (2002). Research and rhetoric on teacher certification: A response to "Teacher Certification Reconsidered." *Education Policy Analysis Archives, 10*(36), 1–55. The study is ineligible for review because it does not use an eligible design.

Additional sources:

Darling-Hammond, L. (2001). *The research and rhetoric on teacher certification: A response to "Teacher Certification Reconsidered."* Arlington, VA: National Commission on Teaching & America's Future. http://files.eric.ed.gov/fulltext/ED477296.pdf.

Walsh, K. (2001a). *Teacher certification reconsidered: Stumbling for quality*. Baltimore, MD: The Abell Foundation. Retrieved from http://files.eric.ed.gov/fulltext/ED460100.pdf

Walsh, K. (2001b). *Teacher certification reconsidered: Stumbling for quality: A rejoinder*. Baltimore, MD: The Abell Foundation. Retrieved from http://files.eric.ed.gov/fulltext/ED481389.pdf

Diaz, V. H. (2012). *Beginning teachers' production of pedagogical content knowledge: A cultural historical perspective* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3505592) The study is ineligible for review because it does not use an eligible design.

Dobbie, W., & Fryer, Jr., R. G. (2015). The impact of voluntary youth service on future outcomes: Evidence from Teach For America. *The B.E. Journal of Economic Analysis & Policy, 15*(3), 1031–1065. The study is ineligible for review because it is out of the scope of the protocol.

Additional source:

Dobbie, W., & Fryer, Jr., R. G. (2011). *The impact of youth service on future outcomes: Evidence from Teach For America* (NBER Working Paper 17402). Cambridge, MA: National Bureau of Economic Research.

Harding, H. (2012). Teach for America: Leading for change. *Educational Leadership, 69*(8), 58–61. The study is ineligible for review because it does not use an eligible design.

Heilig, J. V., & Jez, S. J. (2010). *Teach For America: A review of the evidence*. Boulder, CO and Tempe, AZ: Education and the Public Interest Center and the Education Policy Research Unit. Retrieved from http://files.eric.ed.gov/fulltext/ED510247.pdf The study is ineligible for review because it does not use an eligible design.

Higgins, M., Hess, F. M., Weiner, J., & Robison, W. (2011). Creating a corps of change agents. *Education Next, 11*(3), 18–25. The study is ineligible for review because it is out of the scope of the protocol.

Kopp, W. (2009). Building the movement to end educational inequity. *Education Digest: Essential Readings Condensed for Quick Review, 74*(7), 10–13. The study is ineligible for review because it does not use an eligible design.

Lang, N., Buchanan, T., & Morin, L. (2013). *Perception of preparedness of novice teachers from alternative and traditional licensing programs* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3602833) The study is ineligible for review because it is out of the scope of the protocol.

Lewis, A. S. (2013). *The impact of Teach For America's summer institute on first-year TFA's experience in the Kansas City public schools* (Doctoral dissertation). Retrieved from: https://mospace.umsystem.edu/ The study is ineligible for review because it does not use an eligible design.

Mead, S., Chuong, C., & Goodson, C. (2015). *Exponential growth, unexpected challenges: How Teach For America grew in scale and impact*. Sudbury, MA: Bellwether Education Partners. Retrieved from http://bellwethereducation.org/ The study is ineligible for review because it does not use an eligible design.

Murnane, R. J. (2010). Progress and puzzles in educational policy research. *Harvard Education Letter, 26*(2), 8. The study is ineligible for review because it does not use an eligible design.

Ness, M. K. (2010). Resisting traditional notions of teacher certification. In D. M. Moss & T. A. Osborn (Eds.), *Critical essays on resistance in education* (pp. 17–34). New York: Peter Lang Publishing. The study is ineligible for review because it does not use an eligible design.

Rochkind, J., Ott, A., Immerwahr, J., Doble, J., & Johnson, J. (2007). *Lessons learned: New teachers talk about their jobs, challenges and long-range plans. Issue no. 2. Working without a net: How new teachers from three prominent alternate route programs describe their first year on the job*. New York: National Comprehensive Center for Teacher Quality and Public Agenda. Retrieved from http://files.eric.ed.gov/fulltext/ED499415.pdf The study is ineligible for review because it is out of the scope of the protocol.

Shaw, M. E. (2006). *The impact of alternative teacher certification programs on teacher shortages in Florida, Idaho, New Hampshire, Pennsylvania, and Utah* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3228661) The study is ineligible for review because it is out of the scope of the protocol.

Storm, M. D. (2004). *Beginning and experienced teachers' beliefs about students, teaching, and learning* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3131431) The study is ineligible for review because it is out of the scope of the protocol.

Tatel, E. S. (1997, January). *Teach for America: An effective emergency teaching corps*. Paper presented at the annual meeting of the American Association of Colleges for Teacher Education, Phoenix, AZ. http://files.eric.ed.gov/fulltext/ED405331.pdf. The study is ineligible for review because it does not use an eligible design.

Terry, J. D. (2004). *The effects of short-term teacher preparation on the efficacy of mathematics teachers in Teach for America* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3141228) The study is ineligible for review because it is out of the scope of the protocol.

Appendix A.1: Research details for Clark et al. (2013)

Clark, M. A., Chiang, H. S., Silva, T., McConnell, S., Sonnenfeld, K., Erbe, A., & Puma, M. (2013). *The effectiveness of secondary math teachers from Teach For America and the Teaching Fellows programs* (NCEE 2013-4015). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Additional source:

Chiang, H. S., Clark, M.A., & McConnell, S. (2014). *Supplying disadvantaged schools with effective teachers: Experimental evidence on secondary math teachers from Teach For America* (Working Paper 31). Princeton, NJ: Mathematica Policy Research.

Table A1. Summary of findings

Meets WWC group design standards without reservations

		Study findings	
Outcome domain	Sample size	Average improvement index *(percentile points)*	Statistically significant
Mathematics achievement	136 teachers/4,573 students	+3	Yes

Setting

The study was conducted in 45 secondary schools in 11 school districts in 10 *TFA* regions in eight states.[20]

Study sample

The study included two cohorts of students in grades 6–12, one that participated in the 2009–10 school year and one that participated in the 2010–11 school year. In each participating school, students were randomly assigned within "classroom matches" to either a class taught by a *TFA* teacher or a class taught by a comparison teacher. A classroom match consisted of two or more classes covering the same eligible middle or high school math course that were deemed comparable by the study authors based on factors such as level (for example, honors or regular), length (one or two semesters), and arrangements made for the inclusion of English learners and special education students.[21] After 6,178 students (3,075 *TFA*, 3,103 comparison) were randomly assigned, attrition occurred due to students leaving the school prior to the start of the school year, lack of parental consent, or students not having valid end-of-year math achievement scores. The analytic sample included 4,573 students (2,292 *TFA*, 2,281 comparison) taught by 136 teachers (66 *TFA*, 70 comparison) in 45 schools. The mean age of the students was 13.4 years.[22] Among the sample, 75% of students were in grades 6–8, 49% were female, 90% were eligible for free or reduced-price lunch, 8% were limited English proficient, and 6% had an individualized education plan. The racial/ethnic demographics were as follows: 62% were Black, 28% were Hispanic, 7% were White, 2% were Asian, and 1% were another race/ethnicity.

In addition, the authors present subgroup findings for school levels (middle or high school), years of teaching experience, and comparison group teachers' route to certification (traditional or less selective alternative). The years of teaching experience comparisons include: (a) *TFA* teachers in their first 3 years of teaching versus non-*TFA* teachers in their first 3 years of teaching, (b) *TFA* teachers in their first 3 years of teaching versus non-*TFA* teachers with more than 3 years of experience, (c) *TFA* teachers in their first 2 years of teaching versus non-*TFA* teachers with more than 5 years of experience, (d) *TFA* teachers in their first year of teaching versus non-*TFA* teachers with more than 5 years of experience, and (e) *TFA* teachers in their second year of teaching versus non-*TFA* teachers with more than 5 years of experience. The subgroup findings are reported in Appendix D. The supplemental findings do not factor into the intervention's rating of effectiveness.

Intervention group

Students were taught by *TFA* teachers. Most teachers (83%) were current corps members within their 2-year teaching commitment, but some were *TFA* alumni who had completed the commitment and continued teaching. The mean years of teaching experience at the end of the study year was 1.9. Among *TFA* teachers, 81% had a bachelor's degree from a most, highly, or very competitive college or university; 8% majored in math, none majored in secondary math education, and 27% majored in other math-related subjects.[23] Regarding math content knowledge, the mean score was 162 among teachers who took the Praxis II Mathematics Content Knowledge Test (0.93 standard deviations higher than comparison teachers) and 180 among teachers who took the Praxis II Middle School Mathematics Test (1.19 standard deviations higher than comparison teachers). The mean age of *TFA* teachers at the time of the study was 24.5 years, and 61% of *TFA* teachers were female, 89% were White, 9% were Asian, 8% were Black, and 5% were Hispanic. The authors did not report any deviations from the *TFA* model.

Comparison group

Students in the comparison group were taught by teachers who did not enter teaching through *TFA*, *Teaching Fellows*, or other highly selective alternative routes to certification. The majority (59%) of comparison teachers entered teaching through a traditional route to certification (that is, they became certified teachers after completing a standard postsecondary program for teaching and related certification requirements), with the remainder entering through a less selective alternative route. The mean years of teaching experience at the end of the study year was 10.1. Among comparison teachers, 23% had a bachelor's degree from a most, highly, or very competitive college or university; 26% majored in math, 16% majored in secondary math education, and 12% majored in other math-related subjects. Regarding math content knowledge, the mean score was 140 among teachers who took the Praxis II Mathematics Content Knowledge Test and 158 among teachers who took the Praxis II Middle School Mathematics Test. The mean age of comparison teachers at the time of the study was 37.9 years, and 79% of comparison teachers were female, 57% were Black, 30% were White, 13% were Hispanic, and 11% were Asian.

Outcomes and measurement

An outcome in the mathematics achievement domain was reported. All assessment scores were converted into *z*-scores, thus providing a single outcome for the analysis that expressed math achievement in standard deviation units. For students in grades 6–8, study authors obtained scores from state-required assessments administered in the spring semester of the school year in which the students were randomly assigned. For students in grades 9–12, study authors administered end-of-course math assessments. For a more detailed description of these outcome measures, see Appendix B. The study also examined measures of student absences and teacher job satisfaction; these outcomes are ineligible for review because they are not within a domain specified in the Teacher Training, Evaluation, and Compensation protocol.

Support for implementation

Training provided to *TFA* participants prior to their becoming classroom teachers involves an intensive 5-week summer institute that includes instructor-led coursework, practice teaching, independent work and reflection, and discussions with advisors. During their 2-year commitment, *TFA* staff observe teachers in their classrooms; provide training on topics such as classroom management, goal setting, lesson planning, pedagogy, and student assessment; and offer individualized support as needed.

Appendix A.2: Research details for Clark et al. (2015)

Clark, M. A., Isenberg, E., Liu, A. Y., Makowsky, L., & Zukiewicz, M. (2015). *Impacts of the Teach for America Investing in Innovation scale-up*. Princeton, NJ: Mathematica Policy Research. Retrieved from http://www.mathematica-mpr.com

Table A2. Summary of findings **Meets WWC group design standards without reservations**

Outcome domain	Sample size	Study findings	
		Average improvement index *(percentile points)*	Statistically significant
Mathematics achievement	150 teachers/2,065 students	+2	No
English language arts achievement	154 teachers/2,123 students	+1	No

Setting

The study was conducted in 36 schools in 13 *TFA* placement partners in 10 *TFA* regions in 10 states. The 13 placement partners included 11 traditional public school districts, one charter school district, and one community-based organization that manages an early childhood education program.

Study sample

The study included students in grades pre-K–5 who participated in the 2012–13 school year. In each participating school, students in the same grade level were randomly assigned within "classroom matches" to a class taught by a *TFA* teacher or a class taught by a comparison teacher. A classroom match consisted of two or more classes taught under similar circumstances; for example, all classes in a given match had to be taught in the same language or combination of languages (English versus bilingual English/Spanish or classes for English learners) and all classes in the match were either self-contained or departmentalized. After 3,724 students (1,544 *TFA*, 2,180 comparison) in the 13 placement partners were randomly assigned, attrition occurred due to students not enrolling in the study school, lack of parental consent, or students not having valid end-of-year test score data. The authors included in their analysis a total of 2,153 students (895 *TFA*, 1,258 comparison) taught by 156 teachers (66 *TFA*, 90 comparison). The analytic samples by outcome included 2,065 students (855 *TFA*, 1,210 comparison) for mathematics achievement and 2,123 students (877 *TFA*, 1,246 comparison) for English language arts achievement.[24] Among the students, 47% were female, 7% had an individualized education plan, 34% were limited English proficient, and 84% were eligible for free or reduced-price lunch. The racial/ethnic demographics were as follows: 47% were Black; 42% were Hispanic; 7% were White, 3% were another race, 2% were Asian.

In addition, the authors present subgroup findings for student grade levels (early childhood students [pre-K and K], lower elementary students [pre-K–2], and upper elementary students [grades 3–5]), *TFA* teachers versus comparison teachers in their first or second year of teaching, and *TFA* teachers versus traditionally certified comparison teachers. The subgroup findings are reported in Appendix D. The supplemental findings do not factor into the intervention's rating of effectiveness.

Intervention group

Students were taught by *TFA* teachers. All but one of the *TFA* teachers were in their first or second year of teaching. The mean years of teaching experience was 1.7, and 76% of *TFA* teachers had a bachelor's degree from a most, highly, or very competitive college or university.[25] Among majors, 20% of *TFA* teachers majored in early childhood or elementary education, and 84% majored in a field unrelated to education. The mean age of *TFA* teachers at the time of the study was 24.4 years, and 90% of *TFA* teachers were female, 70% were White, 12% were Black, 12% were Asian, and 7% were Hispanic.

The study assessed the effectiveness of *TFA* teachers during the second year of a *TFA* expansion effort, partially funded by a 5-year Investing in Innovation (i3) scale-up grant from the U.S. Department of Education. *TFA* increased its recruitment among less selective colleges, Historically Black Colleges and Universities, and the Hispanic Association of Colleges and Universities; however, the study authors found no evidence of a change in the program's academic selection standards, as measured by undergraduate grade point average and SAT score.[26] The authors found few substantive changes to *TFA* training and support under the scale-up. However, the authors noted declines in corps members' satisfaction with the program.[27]

Comparison group

Students in the comparison group were taught by teachers who did not enter teaching through *TFA*. The majority (85%) of comparison teachers were traditionally certified teachers (that is, they completed all certification requirements through a traditional university-based program prior to beginning teaching), with the remainder being alternatively certified (that is, they began teaching prior to completing all certification requirements). The mean years of teaching experience was 13.7. Among comparison teachers, 40% had a bachelor's degree from a most, highly, or very competitive college or university; 81% majored in early childhood or elementary education, and 26% majored in a field unrelated to education. The mean age of comparison teachers at the time of the study was 42.8 years, and 99% of comparison teachers were female, 55% were White, 34% were Black, 11% were Hispanic, and 3% were Asian.

Outcomes and measurement

Outcomes in the mathematics and English language arts achievement domains were reported. The study authors converted all assessment scores in each domain into *z*-scores to provide a single mathematics outcome and a single English language arts outcome that expressed achievement in standard deviation units. The authors administered end-of-year math and reading assessments for grades pre-K–2 and obtained end-of-year state-required assessment scores for grades 3–5. For a more detailed description of these outcome measures, see Appendix B. The study also examined measures of teachers' perceptions of issues that hinder student learning in their classrooms, job satisfaction, and career plans; these outcomes are ineligible for review because they are not within a domain specified in the Teacher Training, Evaluation, and Compensation protocol.[28]

Support for implementation

Training provided to *TFA* participants prior to their becoming classroom teachers involves a 5-week summer institute that includes group instruction on curriculum, literacy, and diversity; supervised teaching; observations of other teachers; feedback from advisors; small-group sessions on teaching practice, and lesson-planning clinics. During their 2-year commitment, *TFA* staff provide ongoing training and support that includes one-on-one coaching, grade/subject-specific group meetings, and access to online classroom resources and assessments. As noted above, the authors found few substantive changes to *TFA* training and support under the scale-up.

Appendix A.3: Research details for Glazerman et al. (2006)[29]

Glazerman, S., Mayer, D., & Decker, P. (2006). Alternative routes to teaching: The impacts of Teach For America on student achievement and other outcomes. *Journal of Policy Analysis & Management, 25*(1), 75–96.

Additional source:

Decker, P. T., Mayer, D. P., & Glazerman, S. (2004). *The effects of Teach For America on students: Findings from a national evaluation*. Princeton, NJ: Mathematica Policy Research. Retrieved from http://www.mathematica-mpr.com

Table A3. Summary of findings **Meets WWC group design standards without reservations**

Outcome domain	Sample size	Study findings	
		Average improvement index *(percentile points)*	Statistically significant
Mathematics achievement	100 classes/1,715 students	+6	Yes
English language arts achievement	100 classes/1,715 students	+1	No

Setting

The study took place in 17 schools located in six *TFA* regions: Baltimore, Chicago, Houston, Los Angeles (Compton school district), the Mississippi Delta, and New Orleans.

Study sample

The study included students in grades 1–5. In each participating school, students were randomly assigned within grade to either a class taught by a *TFA* teacher or a class taught by a comparison teacher. The sample included 44 classes taught by *TFA* teachers and 56 classes taught by comparison teachers. Of the 1,969 randomly assigned students who enrolled in study schools (875 *TFA*, 1,094 comparison), 1,715 (759 *TFA*, 956 comparison) were included in the test score analytic sample. Among the students, 49% were female, 20% were overage for their grade, and 95% were eligible for free or reduced-price lunch. The racial/ethnic demographics were as follows: 67% were African American; 26% were Hispanic; 4% were unknown; and 3% were another race/ethnicity, non-Hispanic.

Intervention group

Students were taught by *TFA* teachers. Most teachers were current *TFA* corps members within their 2-year teaching commitment, but some were *TFA* alumni who had completed the commitment and continued teaching. The median years of teaching experience was 2. Among *TFA* teachers, 70% had a bachelor's degree from a most, highly, or very competitive college or university.[30] By the end of the study year, 51% of *TFA* teachers had received a regular or initial teacher certification, and 25% had either a bachelor's or master's degree in education. The median age at the time of the study was 24 years, and 69% of *TFA* teachers were female, 67% were White, 16% were African American, 11% were another race/ethnicity, and 6% were Hispanic. The authors did not report any deviations from the *TFA* model.

Comparison group

Students were taught by individuals who had never been a *TFA* corps member. The median years of teaching experience was 6. Among the comparison teachers, 2% had a bachelor's degree from a most, highly, or very competitive college or university. By the end of the study year, 67% of comparison group teachers had received a regular or initial teacher certification, and 55% had either a bachelor's or a master's degree in education. The median age of comparison group teachers at the time of the study was 35 years, and 87% of comparison group teachers were female, 76% were African American, 11% were White, 11% were Hispanic, and 3% were another race/ethnicity.

Outcomes and measurement

The study examined outcomes in three domains: mathematics achievement, English language arts achievement, and student progression; however, only the findings for outcomes in the mathematics and English language arts achievement domains meet WWC standards. The analysis of student retention in grade, which falls in the student progression domain, does not meet WWC group design standards.[31] The study authors administered math and reading assessments in the fall as a pretest and again in the spring as a posttest. For a more detailed description of these outcome measures, see Appendix B.

The study also examined a number of outcomes that are ineligible for review because they are not within a domain specified in the Teacher Training, Evaluation, and Compensation protocol: attended summer school; number of days absent; being chronically absent; number of days suspended; ever suspended or expelled; teacher reports of a serious problem with student tardiness, student absenteeism/class-cutting, physical conflicts among students, verbal abuse of teachers, and general misbehavior; and teacher reports of the average number of times in the most recent week that students were tardy or absent without excuse, the teacher interrupted class to deal with student disruptions, and the teacher sent a student out of the room.

Support for implementation

TFA teachers received the typical support prescribed by the *TFA* model, which includes attending a 5-week summer institute prior to becoming a classroom teacher and receiving ongoing support during the 2-year teacher commitment from local *TFA* staff who conduct classroom observations and connect corps members with resources to address their specific professional development needs.

Appendix A.4: Research details for Henry, Purtell, et al. (2014)[32]

Henry, G. T., Purtell, K. M., Bastian, K. C., Fortner, C. K., Thompson, C. L., Campbell, S. L., & Patterson, K. M. (2014). The effects of teacher entry portals on student achievement. *Journal of Teacher Education, 65*(1), 7–23.

Additional source:

Henry, G. T., Bastian, K. C., & Smith, A. A. (2012). Scholarships to recruit the best and brightest into teaching: Who is recruited, where do they teach, how effective are they, and how long do they stay? *Educational Researcher, 41*(3), 83–92.

Table A4. Summary of findings

Meets WWC group design standards with reservations

		Study findings	
Outcome domain	**Sample size[a]**	**Average improvement index *(percentile points)***	**Statistically significant**
Mathematics achievement	431 teachers/22,056 students	+5	Yes
Social studies achievement	45 teachers/6,051 students	+3	No
English language arts achievement	433 teachers/18,044 students	+1	No

[a] The reported sample sizes may count some individual students and teachers more than once because the study examined data from multiple school years.

Setting

The study was conducted in North Carolina public schools.

Study sample

The study authors identified teachers of tested grades and subjects during the 2005–06 through 2009–10 school years who had less than 5 years of teaching experience and could be linked to their students using class rosters. The authors then compared outcomes for students of *TFA* teachers to outcomes for students of in-state public undergraduate prepared teachers. For the analyses that meet WWC group design standards, only schools in which both *TFA* and comparison teachers worked were included in the analytic samples. This report presents findings for the six analytic samples for which baseline equivalence between the *TFA* and comparison groups was demonstrated: (a) elementary grades (3–5) math scores for students of *TFA* teachers (2,691 students and 103 teachers) versus students of comparison teachers (4,753 students and 171 teachers); (b) elementary reading scores for students of *TFA* teachers (2,736 students and 107 teachers) versus students of comparison teachers (4,792 students and 175 teachers); (c) middle grades (6–8) math scores for students of *TFA* teachers (4,198 students and 58 teachers) versus students of comparison teachers (3,376 students and 38 teachers); (d) middle grades reading scores for students of *TFA* teachers (6,408 students and 92 teachers) versus students of comparison teachers (4,108 students and 59 teachers); (e) high school math scores for students of *TFA* teachers (3,226 students and 36 teachers) versus students of comparison teachers (3,812 students and 25 teachers); and (f) high school social studies scores for students of *TFA* teachers (2,556 students and 20 teachers) versus students of comparison teachers (3,495 students and 25 teachers). The authors did not report the demographic characteristics of the students and teachers included in these analytic samples.

In addition, the authors of a related publication (Henry et al., 2012) present findings for students of *TFA* teachers versus students of in-state prepared teachers, excluding North Carolina Teaching Fellows Program scholarship recipients.[33] These results are based on students and teachers during the 2005–06 through 2009–10 school years. The Henry et al. (2012) findings that meet WWC group design standards are reported as supplemental findings in Appendix D. The supplemental findings do not factor into the intervention's rating of effectiveness.

Intervention group

Students were taught by *TFA* teachers with less than 5 years of teaching experience. The authors did not report any deviations from the *TFA* model.

Comparison group

Students were taught by individuals with less than 5 years of teaching experience who were in-state undergraduate prepared teachers. These teachers completed their initial licensure requirements prior to beginning teaching by receiving a bachelor's degree from a North Carolina public university.

For the Henry et al. (2012) findings presented in Appendix D, comparison students were taught by in-state prepared teachers, excluding North Carolina Teaching Fellows Program scholarship recipients.[34]

Outcomes and measurement

Henry, Purtell, et al. (2014) examined outcomes in four domains: mathematics achievement, science achievement, social studies achievement, and English language arts achievement; however, only the analyses of outcomes in the mathematics, social studies, and English language arts achievement domains meet WWC group design standards. End-of-grade and end-of-course science test scores do not meet WWC group design standards.[35] For elementary and middle school students, end-of-grade math and reading test scores were reported, with prior year or, for grade 3, grade 3 pretest scores being used as a pretest measure in the analysis; test scores were standardized within subject, grade, and year. For high school students,

end-of-course test scores were reported, with grade 8 math and reading tests used as pre-tests; test scores were standardized within subject and year. For a more detailed description of these outcome measures, see Appendix B.

Henry et al. (2012) examined outcomes in four domains: mathematics achievement, English language arts achievement, general achievement, and teacher retention in the state; however, only the analyses of outcomes in the mathematics and English language arts achievement domains meet WWC standards. The high school test score outcome, which falls in the general achievement domain because it standardizes end-of-course test scores across subjects, and two teacher retention in the state outcomes (the percentages of teachers who return to North Carolina public schools for a third year or fifth year of teaching) do not meet WWC group design standards.[36] The mathematics and English language arts outcomes in Henry et al. (2012) are the same as the outcomes in Henry, Purtell, et al. (2014).

Support for implementation

TFA teachers received training through a 5-week summer program they attend prior to beginning teaching. They also received mentoring and professional development from *TFA* throughout their 2-year teaching commitment.

Appendix A.5: Research details for Turner et al. (2012)

Turner, H. M., Goodman, D., Adachi, E., Brite, J., & Decker, L. E. (2012). *Evaluation of Teach For America in Texas schools*. San Antonio, TX: Edvance Research, Inc.

Table A5. Summary of findings **Meets WWC group design standards with reservations**

		Study findings	
Outcome domain	Sample size	Average improvement index *(percentile points)*	Statistically significant
Mathematics achievement	nr/9,146 students	+6	Yes
English language arts achievement	nr/11,202 students	+2	No

nr = not reported. The study authors did not report the number of teachers or classes in the analytic sample.

Setting

The study took place in more than 400 elementary and middle schools ("campuses") in four regions of Texas: Dallas–Fort Worth, Houston, the Rio Grande Valley, and San Antonio.[37]

Study sample

Beginning with an initial sample of 316 Texas campuses that employed at least one *TFA* teacher in the 2010–11 school year, the authors matched at the campus and student levels to obtain a comparison sample of students at campuses that did not employ a *TFA* teacher in the 2010–11 school year. The final sample included students in grades 4–8 who were enrolled in the Texas public education system for more than 150 days in the 2010–11 school year and took the regular version of the 2010–11 TAKS in math or reading at the first administration. The study analyzed two sets of teacher comparisons: (a) *TFA* corps members, who were within their 2-year contract assignment during the 2010–11 school year, versus novice comparison teachers, who had less than 3 years of teaching experience; and (b) *TFA* alumni, who had completed their 2-year contract assignment before the 2010–11 school year but continued teaching in Texas schools, versus experienced comparison teachers, who had 3 or more years

of experience. This report focuses on findings for students taught by *TFA* corps members and students taught by novice comparison teachers.[38] Findings from the comparison between students taught by *TFA* alumni and students taught by experienced comparison teachers are presented as supplemental findings in Appendix D and do not factor into the intervention's rating of effectiveness.

The analysis of *TFA* corps members includes four samples defined by grade level and subject: (a) the elementary grades (4–5) math sample included 545 students of *TFA* corps members from 25 campuses and 545 comparison group students from 90 campuses; (b) the elementary grades reading sample included 830 students of *TFA* corps members from 37 campuses and 830 comparison group students from 103 campuses; (c) the middle grades (grades 6–8) math sample included 4,028 students of *TFA* corps members from 51 campuses and 4,028 comparison group students from 205 campuses; and (d) the middle grades reading sample included 4,771 students of *TFA* corps members from 55 campuses and 4,771 comparison group students from 157 campuses. Average baseline characteristics varied across samples and groups. Among the sample, 49%–54% were female, 71%–88% were Hispanic, 12%–29% were African American, 94%–97% were economically disadvantaged, and 10%–22% were limited English proficient.

Intervention group

Students were taught by *TFA* corps members who were in either their first or second year of *TFA* assignment. Corps members chosen through a highly selective process undergo a 5-week summer training before beginning a 2-year teaching assignment in a low-income urban or rural public school. The authors did not report any deviations from the *TFA* model.

Comparison group

The study authors created a matched comparison group from within Texas public schools using students taught by teachers who did not participate in *TFA* and had less than 3 years of teaching experience. The authors first used propensity score matching to identify 924 comparison campuses matched based on campus-level demographic variables and prior-year achievement variables; they then added 717 campuses that had not been matched but were located in the same districts as the campuses employing *TFA* participants.[39] In a second stage of propensity score matching, students were matched within grade level and teacher experience category (novice versus experienced) based on student-level demographic and prior-year achievement variables.[40]

Outcomes and measurement

Outcomes in the mathematics and English language arts achievement domains were reported. A state-based assessment was administered each spring, with the prior year's scores being used as a pretest measure. For a more detailed description of these outcome measures, see Appendix B.

Support for implementation

TFA corps members received the typical support prescribed by the *TFA* model. The *TFA* support was grounded in classroom leadership training. In addition to observations during the 5-week summer training, mentors observed corps members at least four times a year during the 2-year assignment and provided support through coaching, instructional demonstrations, and discussions.

Appendix A.6: Research details for Ware et al. (2011)

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

Table A6. Summary of findings

Meets WWC group design standards with reservations

Outcome domain	Sample size[a]	Study findings	
		Average improvement index *(percentile points)*	Statistically significant
Mathematics achievement	527 teachers/25,769 students	+2	No
English language arts achievement	569 teachers/20,511 students	+1	No

[a] The reported sample sizes may count some individual students and teachers more than once because the study examined data from multiple school years.

Setting

The study analyzed student achievement data from four Texas school districts: Donna ISD, Houston ISD, McAllen ISD, and IDEA Public Schools.

Study sample

The study authors identified four Texas school districts with data sufficient for conducting their analyses, including data that link students with their teachers. The necessary achievement data were available for two cohorts of teachers and students: the 2008–09 cohort and the 2009–10 cohort. The authors identified students who took the English version of the mathematics and TAKS ELA/R tests and contributed to their schools' accountability ratings. Specifically, this "accountability subset" included only students who (a) took the TAKS or the TAKS (Accommodated) (TAKS-Alternative and TAKS-Modified scores were excluded), (b) were enrolled at the school on a specific fall semester day, and (c) took the TAKS at that same school in the spring semester. After linking these students to their mathematics and ELA/R teachers and using national *TFA* office data to identify *TFA* teachers, the authors compared TAKS year-to-year passing rate gains for students of *TFA* teachers to TAKS year-to-year passing rate gains for students of teachers who did not participate in the *TFA* program and had less than 3 years of experience ("novice non-*TFA* teachers"). This report presents findings for four analytic samples for which baseline equivalence between the *TFA* and comparison groups was demonstrated: (a) TAKS math gains for grades 3–8 students of *TFA* teachers (3,059 students and 70 teachers) versus students of comparison teachers (10,032 students and 291 teachers) in the 2009–10 cohort; (b) TAKS math gains for grades 9–11 students of *TFA* teachers (2,314 students and 32 teachers) versus students of comparison teachers (10,364 students and 134 teachers) in the 2009–10 cohort; (c) TAKS ELA/R gains for grades 3–8 students of *TFA* teachers (2,044 students and 67 teachers) versus students of comparison teachers (10,169 students and 374 teachers) in the 2008–09 cohort; and (d) TAKS ELA/R gains for grades 9–11 students of *TFA* teachers (1,923 students and 28 teachers) versus students of comparison teachers (6,375 students and 100 teachers) in the 2009–10 cohort. The authors did not report the demographic characteristics of the students and teachers included in these analytic samples.

In addition, the authors present subgroup findings for African-American students, Hispanic students, and economically disadvantaged students (i.e., students eligible for free or reduced-price lunch). The subgroup findings for which baseline equivalence between the *TFA* and comparison groups was demonstrated are reported as supplemental findings in Appendix D. The supplemental findings do not factor into the intervention's rating of effectiveness.

Intervention group

Students were taught by *TFA* teachers. The authors did not report any deviations from the *TFA* model.

Comparison group

Students were taught by novice non-*TFA* teachers—that is, teachers with less than 3 years of experience who did not participate in *TFA*.

Outcomes and measurement

The study examined outcomes in five domains: mathematics achievement, English language arts achievement, teacher retention at the school, teacher retention in the school district, and teacher retention in the state; however, only analyses of the outcomes in the mathematics and English language arts achievement domains meet WWC group design standards.[41] The study authors measured achievement using the passing rate gain, defined as the difference between the current year passing rate for a teacher's students on a given state-required test (mathematics or English language arts/reading) and those students' prior year passing rate on the same test. For a more detailed description of these outcome measures, see Appendix B.

Support for implementation

Prior to beginning teaching, *TFA* teachers receive training in a 5-week summer institute that aims to develop participants' pedagogical knowledge and skills and includes supervised practice teaching under the direction of an experienced teacher. During their 2-year teaching commitment, *TFA* teachers receive one-on-coaching from *TFA* program directors, participate in learning team meetings with other *TFA* teachers, and have access to online *TFA* resources that include lesson plans and videos.

Appendix A.7: Research details for Xu et al. (2011)[42]

Xu, Z., Hannaway, J., & Taylor, C. (2011). Making a difference? The effects of Teach For America in high school. *Journal of Policy Analysis and Management, 30*(3), 447–469.

Table A7. Summary of findings **Meets WWC group design standards with reservations**

		Study findings	
Outcome domain	Sample size	Average improvement index *(percentile points)*	Statistically significant
Science achievement	574 teachers/ 36,104 students	+ 7	Yes

Setting

The study took place in North Carolina high schools that employed at least one *TFA* teacher.

Study sample

The study included high school students during the 2000–01 through 2006–07 school years. The student data available to the study authors identified each student's exam proctor, who was not necessarily the teacher who taught the student in the course assessed. The authors linked students to their teachers based on exam proctor and classroom demographics. The analytic sample for the comparison that meets WWC group design standards included 3,495 students of 70 *TFA* teachers versus 32,609 students of 504 comparison teachers for science achievement.[43] The authors did not report the characteristics of the students and teachers included in this analytic sample.

In addition, the authors presented findings that meet WWC group design standards for the following subgroups: (a) science achievement of students of in-field *TFA* teachers versus

students of in-field non-*TFA* teachers, (b) science achievement of students of *TFA* teachers versus students of traditional track teachers, and (c) mathematics achievement of students of *TFA* teachers versus students of non-*TFA* teachers who held a Standard Professional I license. In-field teachers held a license in the subject they taught. Traditional track teachers earned their teaching licenses by completing a teacher education program at an accredited North Carolina institution of higher education. The Standard Professional I license is the regular teaching license typically held by teachers with less than 3 years of experience who completed all state requirements. These subgroup findings are reported as supplemental findings in Appendix D. The supplemental findings do not factor into the intervention's rating of effectiveness.

Intervention group

Students were taught by *TFA* teachers. The authors did not report any deviations from the *TFA* model.

Comparison group

Students were taught by individuals who did not enter teaching through *TFA*.

Outcomes and measurement

The study authors examined outcomes in three domains: mathematics achievement, science achievement, and general achievement; however, only analyses of the outcomes in the mathematics and science achievement domains meet WWC group design standards. The analysis of the all-subjects high school test score outcome, which falls in the general achievement domain because it standardizes end-of-course test scores across subjects, does not meet WWC group design standards.[44] Furthermore, the only mathematics outcome analysis that meets standards is conducted for a subgroup and presented as a supplemental finding in Appendix D. The supplemental finding does not factor into the intervention's rating of effectiveness.

End-of-course math and science tests were administered to high school students; all test scores were standardized within subject and year. For a more detailed description of these outcome measures, see Appendix B.

Support for implementation

TFA corps members received training through a 5-week summer institute and a 2-week local orientation and induction program they attend prior to beginning teaching. They also received ongoing professional development from *TFA*.

Appendix B: Outcome measures for each domain

Mathematics achievement	
End-of-Year Math Assessments	Clark et al. (2015) used study-administered Woodcock-Johnson III (W-J III) Normative Update Tests of Achievement for students in grades pre-K–2 and state-required assessments for students in grades 3–5. In particular, the study authors measured math achievement for grades pre-K–2 students using W-J III Broad Math *W* scores from the Applied Problems subtest administered to students in grades pre-K–2 and the Calculation subtests administered to students in grades 1 and 2. For students in grades 3–5, the study authors obtained state math test scores from school districts. The authors converted each score to a *z*-score, using the Woodcock-Johnson national norming populations for grades pre-K–2 and the full population of students in the same state and grade who took the same assessment for grades 3–5 as reference populations (as cited in Clark et al., 2015).
Iowa Test of Basic Skills (ITBS) Mathematics Subtest	Glazerman et al. (2006) administered an abbreviated form of the mathematics subtest of the ITBS to students in grades 1–5. The authors do not describe how the abbreviated form differs from the full form of the standardized test. Test scores were reported as normal curve equivalent (NCE) scores, which have a mean nationally of 50 and a standard deviation of 21.06 (as cited in Glazerman et al., 2006).
Mathematics Assessments	Clark et al. (2013) used state-required math assessments for students in grades 6–8 and study-administered NWEA end-of-course math assessments for students in grades 9–12. The state-required assessments were criterion-referenced tests. The tests differed across states, with each test being part of the state's accountability system. Each score was converted to a *z*-score using as a reference population the full population of students in the same state, year, and grade who took the same assessment. The NWEA assessments were computer-adaptive tests administered in general high school math, Algebra I, geometry, or Algebra II, depending on the content of the student's math course. The administration and scoring of the tests for the study differed from standard NWEA procedures in that the study authors imposed a 35-minute time limit and obtained scores for incomplete tests. The study authors reported marginal reliability coefficients of .927 or greater for the analytic sample. Each score was converted to a *z*-score using the NWEA's nationwide norming sample for the assessment as the reference population (as cited in Clark et al., 2013 and Chiang, Clark, & McConnell, 2014).
North Carolina End-of-Course Mathematics Assessment	These state-required standardized tests were administered to high school students at the end of courses in Algebra I, Algebra II, and geometry. Test scores were standardized within subject and year (as cited in Henry, Purtell, et al., 2014 and Xu et al., 2011). For Xu et al. (2011), this outcome is presented as a supplemental finding.
North Carolina End-of-Grade Mathematics Assessment	This standardized test was administered to students in grades 3–8, at the end of the school year. Henry, Purtell, et al. (2014) and Henry et al. (2012) standardized test scores within subject, grade, and year.
Texas Assessment of Knowledge and Skills (TAKS) Mathematics	TAKS Mathematics is a statewide, criterion-referenced assessment administered to students in grades 3–11 in the spring of each school year. The study authors reported that internal consistency estimates of reliability exceeded 0.80 (as cited in Turner et al., 2012).
TAKS Mathematics Passing Rate Gain	TAKS Mathematics is a state test administered to students in grades 3–11 and used for accountability purposes. For each teacher in the analytic sample, Ware et al. (2011) calculated the percentage of the teacher's students who met the minimum standard on the TAKS Mathematics test in the current school year (i.e., the current year passing rate) and the percentage of those same students who met the minimum standard on the test in the prior school year (i.e., the prior year passing rate). The difference between the current year passing rate and the prior year passing rate indicates the passing rate gain (as cited in Ware et al., 2011).
Science achievement	
North Carolina End-of-Course Science Assessments	The state-required tests were administered to high school students at the end of courses in biology, chemistry, physics, and physical sciences. Tests were standardized within subject and year (as cited in Xu et al., 2011).
Social studies achievement	
North Carolina End-of-Course Social Studies Assessments	These state-required standardized tests were administered to high school students at the end of courses in U.S. history and civics/economics. Test scores were standardized within subject and year (as cited in Henry, Purtell, et al., 2014).

English language arts achievement	
End-of-Year Reading Assessments	Clark et al. (2015) used study-administered W-J III Normative Update Tests of Achievement for students in grades pre-K–2 and state-required assessments for students in grades 3–5. In particular, the study authors measured English language achievement for grades pre-K–2 students using W-J III Broad Reading *W* scores derived from the Letter-Word Identification subtest administered to grades pre-K–2 students and the Passage Comprehension subtest administered to grades K–2 students. For students in grades 3–5, the study authors obtained state reading test scores from school districts. The authors converted each score to a *z*-score, using the Woodcock-Johnson national norming populations for grades pre-K–2 and the full population of students in the same state and grade who took the same assessment for grades 3–5 as reference populations (as cited in Clark et al., 2015).
ITBS Reading Subtest	Glazerman et al. (2006) administered an abbreviated form of the reading subtest of the ITBS to students in grades 1–5. The authors do not describe how the abbreviated form differed from the full form of the standardized test. Test scores were reported as NCE scores, which have a mean nationally of 50 and a standard deviation of 21.06 (as cited in Glazerman et al., 2006).
North Carolina End-of-Grade Reading Assessment	This standardized test was administered to students in grades 3–8. Henry, Purtell, et al. (2014) and Henry et al. (2012) standardized test scores within subject, grade, and year.
TAKS English Language Arts/Reading (ELA/R) Passing Rate Gain	TAKS ELA/R is a state test administered to students in grades 3–11 and used for accountability purposes. For each teacher in the analytic sample, Ware et al. (2011) calculated the percentage of the teacher's students who met the minimum standard on the TAKS ELA/R test in the current school year (i.e., the current year passing rate) and the percentage of those same students who met the minimum standard on the test in the prior school year (i.e., the prior year passing rate). The difference between the current year passing rate and the prior year passing rate indicates the passing rate gain (as cited in Ware et al., 2011).
TAKS Reading	TAKS Reading is a statewide, criterion-referenced assessment administered to students in grades 3–11 in the spring of each school year. The study authors reported that internal consistency estimates of reliability exceed 0.80 (as cited in Turner et al., 2012).

Appendix C.1: Findings included in the rating for the mathematics achievement domain

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Clark et al. (2013)[a]								
Mathematics Assessments	Grades 6–12	136 teachers/ 4,573 students	–0.52 (0.95)	–0.60 (0.91)	0.07	0.08	+3	< .01
Domain average for mathematics achievement (Clark et al., 2013)						**0.08**	**+3**	**Statistically significant**
Clark et al. (2015)[b]								
End-of-Year Math Assessments	Grades pre-K–5	150 teachers/ 2,065 students	nr (0.92)	nr (1.14)	0.05	0.05	+2	.28
Domain average for mathematics achievement (Clark et al., 2015)						**0.05**	**+2**	**Not statistically significant**
Glazerman et al. (2006)[c]								
Iowa Test of Basic Skills (ITBS) Mathematics Subtest	Grades 1–5	100 classes/ 1,715 students	30.44 (nr)	28.01 (nr)	2.43	0.15	+6	< .01
Domain average for mathematics achievement (Glazerman et al., 2006)						**0.15**	**+6**	**Statistically significant**
Henry, Purtell, et al. (2014)[d]								
North Carolina End-of-Grade Mathematics Assessment	Grades 3–5	274 teachers/ 7,444 students	–0.36 (0.91)	–0.43 (0.94)	0.07	0.07	+3	< .01
North Carolina End-of-Grade Mathematics Assessment	Grades 6–8	96 teachers/ 7,574 students	–0.21 (0.91)	–0.34 (0.98)	0.13	0.14	+5	< .01
North Carolina End-of-Course Mathematics Assessment	Grades 9–12	61 teachers/ 7,038 students	–0.31 (0.85)	–0.49 (0.90)	0.18	0.20	+8	< .01
Domain average for mathematics achievement (Henry, Purtell, et al., 2014)						**0.14**	**+5**	**Statistically significant**
Turner et al. (2012)[e]								
Texas Assessment of Knowledge and Skills (TAKS) Mathematics	Grades 4–5, novice teachers	nr/ 1,090 students	688.65 (97.00)	678.66 (91.82)	9.99	0.11	+4	.28
TAKS Mathematics	Grades 6–8, novice teachers	nr/ 8,056 students	742.93 (91.65)	725.99 (87.80)	16.94	0.19	+7	< .01
Domain average for mathematics achievement (Turner et al., 2012)						**0.15**	**+6**	**Statistically significant**
Ware et al. (2011)[f]								
TAKS Mathematics Passing Rate Gain	Grades 3–8, 2009–10 cohort	361 teachers/ 13,091 students	3.0 (38)	3.5 (38)	–0.5	–0.01	–1	≥ .10

TAKS Mathematics Passing Rate Gain	Grades 9–11, 2009–10 cohort	166 teachers/ 12,678 students	10.3 (45)	5.6 (47)	4.7	0.10	+4	< .05
Domain average for mathematics achievement (Ware et al., 2011)						**0.04**	**+2**	**Not statistically significant**
Domain average for mathematics achievement across all studies						**0.11**	**+4**	**na**

Table Notes: 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 intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. 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 each study's domain average was determined by the WWC. Some statistics may not sum as expected due to rounding. na = not applicable. nr = not reported.

[a] For Clark et al. (2013), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-value presented here was reported in the original study. The study authors calculated the intervention group mean by adding the impact of the intervention (the regression-adjusted difference between the intervention and comparison groups) to the unadjusted comparison group mean. The unadjusted standard deviations were provided by the study authors at the WWC's request. This study is characterized as having a statistically significant positive effect because the estimated effect for the one measure in this domain is positive and statistically significant. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

[b] For Clark et al. (2015), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-value presented here was reported in the original study. The mean difference is the impact of the intervention (the estimated coefficient on the intervention group indicator from a regression model) as reported in the original study. The unadjusted standard deviations were provided by the study authors at the WWC's request. This study is characterized as having an indeterminate effect because the estimated effect for the one measure in this domain is neither statistically significant nor substantively important. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

[c] For Glazerman et al. (2006), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-value presented here was reported in the original study. The sample size and intervention and comparison group means were obtained from Decker et al. (2004b); the intervention group mean is unadjusted, and the comparison group mean is regression-adjusted. The effect size presented here was reported in Glazerman et al. (2006); the authors calculated the effect size using the standard deviation from the comparison group's math pretest (15.9). This study is characterized as having a statistically significant positive effect because the estimated effect for the one measure in this domain is positive and statistically significant. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

[d] For Henry, Purtell, et al. (2014), the *p*-values presented here were reported in the original study. A correction for clustering was needed and resulted in WWC-computed *p*-values of .21 for grades 3–5 end-of-grade math, .15 for grades 6–8 end-of-grade math, and .08 for grades 9–12 end-of-course math; therefore, the WWC does not find the results to be statistically significant. The WWC calculated the intervention group mean by adding the impact of the intervention (the estimated coefficient on the intervention group indicator from a regression model) to the unadjusted comparison group posttest mean. The analytic sample sizes, unadjusted means, and unadjusted standard deviations were provided by the study authors at the WWC's request. This study is characterized as having a statistically significant positive effect because the WWC determined that the omnibus effect for all outcome measures together is positive and statistically significant. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

[e] For Turner et al. (2012), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-values presented here were reported in the original study. The reported group means were estimated using a hierarchical linear model. This study is characterized as having a statistically significant positive effect because the mean effect is positive and statistically significant, accounting for multiple comparisons. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

[f] For Ware et al. (2011), the *p*-values presented here were reported in the original study. A correction for clustering was needed and resulted in a WWC-computed *p*-value of .82 for the TAKS mathematics passing rate gain for grades 3–8 in the 2009–10 cohort and .27 for grades 9–11 in the 2009–10 cohort; therefore, the WWC does not find the results for either outcome to be statistically significant. The intervention and comparison group means are the mean passing rate gains reported in the original study. The standard deviations are student-level standard deviations calculated by the WWC based on the study-reported current year passing rates. Specifically, the WWC first calculated the standard deviation of the dichotomous student-level passing variable using the formula $\sqrt{(p*(1-p))}$, where p is the unadjusted current year passing rate for the teacher's students; the WWC then multiplied the result by 100 because the authors reported the passing rates as percentages rather than decimals. This study is characterized as having an indeterminate effect because the mean effect is neither statistically significant nor substantively important. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

Appendix C.2: Findings included in the rating for the science achievement domain

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Xu et al. (2011)[a]								
North Carolina End-of-Course Science Assessments	High school	574 teachers/ 36,104 students	nr	nr	0.19	0.19	+7	< .05
Domain average for science achievement (Xu et al., 2011)						**0.19**	**+7**	**Statistically significant**
Domain average for science achievement across all studies						**0.19**	**+7**	**na**

Table Notes: 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 intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. The statistical significance of the study's domain average was determined by the WWC. Some statistics may not sum as expected due to rounding. na = not applicable. nr = not reported.

[a] For Xu et al. (2011), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-value presented here was reported in the original study. The mean difference is the impact of the intervention (the estimated coefficient on the intervention group indicator from a regression model) as reported in the original study. The effect size is the impact estimate from the study because the outcome was scaled to be in standard deviation units. The study authors estimated this impact using a model that included teacher experience variables. The analytic sample sizes were provided by the study authors at the WWC's request. This study is characterized as having a statistically significant positive effect because the effect for the one measure in this domain is positive and statistically significant. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

Appendix C.3: Findings included in the rating for the social studies achievement domain

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Henry, Purtell, et al. (2014)[a]								
North Carolina End-of-Course Social Studies Assessments	Grades 9–12	45 teachers/ 6,051 students	–0.11 (0.91)	–0.17 (0.98)	0.06	.07	+3	≥ .05
Domain average for social studies achievement (Henry, Purtell, et al., 2014)						**.07**	**+3**	**Not statistically significant**
Domain average for social studies achievement across all studies						**.07**	**+3**	**na**

Table Notes: 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 intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. The statistical significance of the study's domain average was determined by the WWC. Some statistics may not sum as expected due to rounding. na = not applicable.

[a] For Henry, Purtell, et al. (2014), a correction for clustering was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-value presented here was reported in the original study. The WWC calculated the intervention group mean by adding the impact of the intervention (the estimated coefficient on the intervention group indicator from a regression model) to the unadjusted comparison group posttest mean. The analytic sample sizes, unadjusted means, and unadjusted standard deviations were provided by the study authors at the WWC's request. This study is characterized as having an indeterminate effect because the estimated effect for the one measure in this domain is neither statistically significant nor substantively important. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

Appendix C.4: Findings included in the rating for the English language arts achievement domain

Outcome measure	Study sample	Sample size	Mean (standard deviation) Intervention group	Mean (standard deviation) Comparison group	WWC calculations Mean difference	WWC calculations Effect size	WWC calculations Improvement index	*p*-value
Clark et al. (2015)[a]								
End-of-Year Reading Assessments	Grades pre-K–5	154 teachers/ 2,123 students	nr (1.07)	nr (1.40)	0.03	0.02	+1	.57
Domain average for English language arts achievement (Clark et al., 2015)						**0.02**	**+1**	**Not statistically significant**
Glazerman et al. (2006)[b]								
Iowa Test of Basic Skills (ITBS) Reading Subtest	Grades 1–5	100 classes/ 1,715 students	28.17 (nr)	27.61 (nr)	0.56	0.03	+1	.37
Domain average for English language arts achievement (Glazerman et al., 2006)						**0.03**	**+1**	**Not statistically significant**
Henry, Purtell, et al. (2014)[c]								
North Carolina End-of-Grade Reading Assessment	Grades 3–5	282 teachers/ 7,528 students	–0.40 (0.94)	–0.43 (0.93)	0.03	0.03	+1	≥ .05
North Carolina End-of-Course Reading Assessment	Grades 6–8	151 teachers/ 10,516 students	–0.22 (0.97)	–0.24 (1.02)	0.02	0.02	+1	≥ .05
Domain average for English language arts achievement (Henry, Purtell, et al., 2014)						**0.02**	**+1**	**Not statistically significant**
Turner et al. (2012)[d]								
Texas Assessment of Knowledge and Skills (TAKS) Reading	Grades 4–5, novice teachers	nr/ 1,660 students	678.69 (87.61)	674.57 (95.06)	4.11	0.04	+2	.49
TAKS Reading	Grades 6–8, novice teachers	nr/ 9,542 students	754.89 (93.49)	751.10 (90.13)	3.79	0.04	+2	.33
Domain average for English language arts achievement (Turner et al., 2012)						**0.04**	**+2**	**Not statistically significant**
Ware et al. (2011)[e]								
TAKS English Language Arts/Reading (ELA/R) Passing Rate Gain	Grades 3–8, 2008–09 cohort	441 teachers/ 12,213 students	3.4 (34)	2.9 (35)	0.5	0.01	+1	≥ .10
TAKS ELA/R Passing Rate Gain	Grades 9–11, 2009–10 cohort	128 teachers/ 8,298 students	3.2 (32)	4.5 (31)	–1.3	–0.04	–2	≥ .10
Domain average for English language arts achievement (Ware et al., 2011)						**–0.01**	**–1**	**Not statistically significant**
Domain average for English language arts achievement across all studies						**0.02**	**+1**	**na**

Table Notes: 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 intervention on outcomes, representing the average change expected for all who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. 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 each study's domain average was determined by the WWC. Some statistics may not sum as expected due to rounding. na = not applicable. nr = not reported.

[a] For Clark et al. (2015), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-value presented here was reported in the original study. The mean difference is the impact of the intervention (the estimated coefficient on the intervention group indicator from a regression model) as reported in the original study. The unadjusted standard deviations were provided by the study authors at the WWC's request. This study is characterized as having an indeterminate effect because the estimated effect for the one measure in this domain is neither statistically significant nor substantively important. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

[b] For Glazerman et al. (2006), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-value presented here was reported in the original study. The sample size, intervention and comparison group means, and *p*-value were obtained from Decker et al. (2004b); the intervention group mean is unadjusted, and the comparison group mean is regression-adjusted. The effect size presented here was reported in Glazerman et al. (2006); the authors calculated the effect size using the standard deviation from the comparison group's reading pretest (17.1). This study is characterized as having an indeterminate effect because the estimated effect for the one measure in this domain is neither statistically significant nor substantively important. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

[c] For Henry, Purtell, et al. (2014), a correction for clustering was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-values presented here were reported in the original study. The WWC calculated the intervention group mean by adding the impact of the intervention (the estimated coefficient on the intervention group indicator from a regression model) to the unadjusted comparison group posttest mean. The analytic sample sizes, unadjusted means, and unadjusted standard deviations were provided by the study authors at the WWC's request. This study is characterized as having an indeterminate effect because the mean effect is neither statistically significant nor substantively important. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

[d] For Turner et al. (2012), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-values presented here were reported in the original study. The reported group means were estimated using a hierarchical linear model. This study is characterized as having an indeterminate effect because the mean effect is neither statistically significant nor substantively important. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

[e] For Ware et al. (2011), a correction for clustering was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-values presented here were reported in the original study. The intervention and comparison group means are the mean passing rate gains reported in the original study. The standard deviations are student-level standard deviations calculated by the WWC based on the study-reported current year passing rates. Specifically, the WWC first calculated the standard deviation of the dichotomous student-level passing variable using the formula $\sqrt{(p*(1-p))}$, where p is the unadjusted current year passing rate for the teacher's students; the WWC then multiplied the result by 100 because the authors reported the passing rates as percentages rather than decimals. This study is characterized as having an indeterminate effect because the mean effect is neither statistically significant nor substantively important. For more information, please refer to the WWC Procedures and Standards Handbook (version 3.0), p. 26.

Appendix D.1: Supplemental grade level and student subgroup findings for the mathematics achievement domain

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Clark et al. (2013)[a]								
Mathematics Assessments	Grades 6–8	103 teachers/ 3,373 students	–0.47 (0.87)	–0.52 (0.84)	0.06	0.07	+3	.01
Mathematics Assessments	Grades 9–12	33 teachers/ 1,200 students	–0.69 (1.12)	–0.82 (1.07)	0.13	0.12	+5	< .01
Clark et al. (2015)[b]								
End-of-Year Math Assessments	Grades pre-K and K	67 teachers/ 878 students	nr (1.04)	nr (1.18)	0.08	0.07	+3	.49
End-of-Year Math Assessments	Grades pre-K–2	123 teachers/ 1,653 students	nr (0.87)	nr (0.95)	0.09	0.10	+4	.14
End-of-Year Math Assessments	Grades 3–5	27 teachers/ 412 students	nr (0.96)	nr (1.31)	0.01	0.01	0	.92
Ware et al. (2011)[c]								
Texas Assessment of Knowledge and Skills (TAKS) Mathematics Passing Rate Gain	Grades 3–8, 2008–09 cohort, African-American students	282 teachers/ 2,811 students	8.5 (44)	5.1 (45)	3.4	0.08	+3	≥ .10
TAKS Mathematics Passing Rate Gain	Grades 3–8, 2009–10 cohort, African-American students	231 teachers/ 2,543 students	1.0 (42)	–0.6 (43)	1.6	0.04	+1	≥ .10
TAKS Mathematics Passing Rate Gain	Grades 9–11, 2009–10 cohort, African-American students	134 teachers/ 2,858 students	14.8 (46)	6.4 (49)	8.4	0.17	+7	< .05
TAKS Mathematics Passing Rate Gain	Grades 3–8, 2009–10 cohort, Hispanic students	336 teachers/ 9,502 students	3.3 (38)	5.0 (37)	–1.7	–0.05	–2	< .10
TAKS Mathematics Passing Rate Gain	Grades 9–11, 2009–10 cohort, Hispanic students	162 teachers/ 8,963 students	9.4 (45)	5.3 (46)	4.1	0.09	+4	< .05

Table Notes: The supplemental findings presented in this table are additional findings from studies in this report that meet WWC design standards with or without reservations, but do not factor into the determination of the intervention rating. 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 intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. Some statistics may not sum as expected due to rounding. nr = not reported.

[a] For Clark et al. (2013), a correction for multiple comparisons was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-values presented here were reported in the original study. The study authors calculated the intervention group mean by adding the impact of the intervention (the regression-adjusted difference between the intervention and comparison groups) to the unadjusted comparison group mean. The unadjusted standard deviations were provided by the study authors at the WWC's request.

[b] For Clark et al. (2015), a correction for multiple comparisons was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-values presented here were reported in the original study. The mean difference is the impact of the intervention (the estimated coefficient on the intervention group indicator from a regression model) as reported in the original study. The unadjusted standard deviations were provided by the study authors at the WWC's request.

[c] For Ware et al. (2011), the *p*-values presented here were reported in the original study. A correction for clustering was needed and resulted in WWC-computed *p*-values for the TAKS mathematics passing rate gain of .16 for African-American students in grades 9–11 in the 2009–10 cohort and .32 for Hispanic students in grades 9–11 in the 2009–10 cohort; therefore, the WWC does not find the results for either outcome to be statistically significant. The intervention and comparison group means are the mean passing rate gains reported in the original study. The standard deviations are student-level standard deviations calculated by the WWC based on the study-reported current year passing rates. Specifically, the WWC first calculated the standard deviation of the dichotomous student-level passing variable using the formula $\sqrt{(p*(1-p))}$, where p is the unadjusted current year passing rate for the teacher's students; the WWC then multiplied the result by 100 because the authors reported the passing rates as percentages rather than decimals. The subgroup finding for economically disadvantaged students in grades 9–11 in cohort 2009–10 also meets WWC group design standards with reservations; however, the finding is not presented here because the authors reported an implausibly high value (332) for the number of *TFA* teachers in the analytic sample. A sample size of 332 is out of line with the *TFA* teacher sample size for other subgroups and exceeds the number of teachers reported for the full sample of all students in grades 9–11 in the 2009–10 cohort. The WWC cannot apply the clustering correction and determine the statistical significance of the finding without the correct teacher sample sizes. The authors define "economically disadvantaged students" as students who are eligible for free or reduced-price lunch.

Appendix D.2: Supplemental teacher subgroup findings for the mathematics achievement domain

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Chiang et al. (2014)[a]								
Mathematics Assessments	Grades 6–12, *TFA* in their first 2 years of teaching vs. non-*TFA* with more than 5 years of experience	~90 teachers/ 2,815 students	–0.59 (0.98)	–0.66 (0.91)	0.07	0.07	+3	.01
Mathematics Assessments	Grades 6–12, *TFA* in their first year of teaching vs. non-*TFA* with more than 5 years of experience	~50 teachers/ 1,434 students	–0.68 (0.99)	–0.69 (0.88)	0.01	0.01	0	.87
Mathematics Assessments	Grades 6–12, *TFA* in their second year of teaching vs. non-*TFA* with more than 5 years of experience	~40 teachers/ 1,381 students	–0.50 (0.97)	–0.63 (0.94)	0.13	0.14	+5	< .01
Clark et al. (2013)[a]								
Mathematics Assessments	Grades 6–12, *TFA* vs. traditional route to certification	82 teachers/ 2,477 students	–0.52 (1.03)	–0.58 (0.95)	0.06	0.06	+2	.03
Mathematics Assessments	Grades 6–12, *TFA* vs. less selective alternative route to certification	58 teachers/ 2,096 students	–0.52 (0.84)	–0.62 (0.86)	0.09	0.11	+4	< .01
Mathematics Assessments	Grades 6–12, teachers in their first 3 years of teaching	23 teachers/ 710 students	–0.24 (0.81)	–0.32 (0.78)	0.08	0.10	+4	.01
Mathematics Assessments	Grades 6–12, *TFA* in their first 3 years of teaching vs. non-*TFA* with more than 3 years of experience	107 teachers/ 3,642 students	–0.59 (0.97)	–0.66 (0.92)	0.07	0.07	+3	< .01

Clark et al. (2015)[b]								
End-of-Year Math Assessments	Grades pre-K–5, teachers in their first 2 years of teaching	23 teachers/ 313 students	nr (0.93)	nr (0.84)	0.04	0.04	+2	.77
End-of-Year Math Assessments	Grades pre-K–5, *TFA* vs. traditionally certified	130 teachers/ 1,836 students	nr (0.92)	nr (1.14)	0.06	0.06	+2	.18
Turner et al. (2012)[c]								
Texas Assessment of Knowledge and Skills (TAKS) Mathematics	Grades 4–5, experienced teachers	nr/ 846 students	697.74 (nr)	694.24 (nr)	3.50	0.04	+2	.67
TAKS Mathematics	Grades 6–8, experienced teachers	nr/ 1,796 students	764.09 (85.24)	740.85 (84.51)	23.25	0.27	+11	< .01
Xu et al. (2011)[d]								
North Carolina End-of-Course Math Assessment	*TFA* vs. non-*TFA* with Standard Professional I license	213 teachers/ 8,662 students	nr	nr	0.11	0.11	+4	≥ .05

Table Notes: The supplemental findings presented in this table are additional findings from studies in this report that meet WWC design standards with or without reservations, but do not factor into the determination of the intervention rating. 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 intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. Some statistics may not sum as expected due to rounding. nr = not reported.

[a] For Clark et al. (2013) and the additional source (Chiang et al., 2014), a correction for multiple comparisons was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-values presented here were reported in the original study. For Clark et al. (2013), the study authors calculated the intervention group mean by adding the impact of the intervention (the regression-adjusted difference between the intervention and comparison groups) to the unadjusted comparison group mean. For Chiang et al. (2014), the WWC calculated the intervention group mean by adding the impact of the intervention (the estimated coefficient on the intervention group indicator from a regression model) to the unadjusted comparison group posttest mean. The unadjusted standard deviations for both publications and the unadjusted comparison group means, student sample sizes, and *p*-values for Chiang et al. (2014) were provided by the study authors at the WWC's request. Chiang et al. (2014) reported teacher sample sizes rounded to the nearest 10.

[b] For Clark et al. (2015), a correction for multiple comparisons was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-values presented here were reported in the original study. The mean difference is the impact of the intervention (the estimated coefficient on the intervention group indicator from a regression model) as reported in the original study. The unadjusted standard deviations were provided by the study authors at the WWC's request. The subgroup contrast involving non-*TFA* teachers in their first 2 years of teaching also excludes the one *TFA* teacher in the study sample who taught for 2 years prior to entering *TFA* and therefore had a total of 3 years of teaching experience.

[c] For Turner et al. (2012), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-values presented here were reported in the original study. The reported group means were estimated using a hierarchical linear model. The effect size presented here for the grades 4–5, experienced teachers subgroup was reported in Turner et al. (2012); the authors calculated the effect size as Hedges' *g*. The intervention group teachers were *TFA* alumni who had completed their 2-year contract assignment but continued teaching in Texas schools. The comparison group teachers were individuals who did not participate in *TFA* and had 3 or more years of teaching experience.

[d] For Xu et al. (2011), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-value presented here was reported in the original study. The mean difference is the impact of the intervention (the estimated coefficient on the intervention group indicator from a regression model) as reported in the original study. The effect size is the impact estimate from the study because the outcome was scaled to be in standard deviation units. The comparison group teachers were non-*TFA* teachers who held the Standard Professional I license, which is the regular teaching license typically held by teachers with less than 3 years of experience who had completed all state requirements. The analytic sample sizes were provided by the study authors at the WWC's request.

Appendix D.3: Supplemental findings for teacher subgroups in the science achievement domain

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Xu et al. (2011)[a]								
North Carolina End-of-Course Science Assessment	In-field *TFA* vs. in-field non-*TFA* teachers	501 teachers/ 33,980 students	nr	nr	0.17	0.17	+7	< .05
North Carolina End-of-Course Science Assessment	*TFA* vs. traditional track teachers	300 teachers/ 20,706 students	nr	nr	0.15	0.15	+6	< .05

Table Notes: The supplemental findings presented in this table are additional findings from studies in this report that meet WWC design standards with or without reservations, but do not factor into the determination of the intervention rating. 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 intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. Some statistics may not sum as expected due to rounding. nr = not reported.

[a] For Xu et al. (2011), a correction for multiple comparisons was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-values presented here were reported in the original study. The mean difference is the impact of the intervention (the estimated coefficient on the intervention group indicator from a regression model) as reported in the original study. The effect size is the impact estimate from the study because the outcome was scaled to be in standard deviation units. In-field teachers held a license in the subject they taught. Traditional track teachers earned their teaching licenses by completing a teacher education program at an accredited North Carolina institution of higher education. The analytic sample sizes were provided by the study authors at the WWC's request.

Appendix D.4: Supplemental grade level and student subgroup findings in the English language arts achievement domain

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Clark et al. (2015)[a]								
End-of-Year Reading Assessments	Grades pre-K and K	67 teachers/ 878 students	nr (1.02)	nr (0.99)	0.15	0.15	+6	.21
End-of-Year Reading Assessments	Grades pre-K–2	123 teachers/ 1,653 students	nr (0.92)	nr (0.91)	0.12	0.13	+5	.04
End-of-Year Reading Assessments	Grades 3–5	31 teachers/ 470 students	nr (1.19)	nr (1.76)	–0.07	–0.05	–2	.40
Ware et al. (2011)[b]								
Texas Assessment of Knowledge and Skills (TAKS) English Language Arts/ Reading (ELA/R) Passing Rate Gain	Grades 9–11, 2008–09 cohort, Hispanic students	115 teachers/ 6,370 students	1.9 (38)	1.0 (38)	0.9	0.02	+1	≥ .10
TAKS ELA/R Passing Rate Gain	Grades 9–11 2008–09 cohort, economically disadvantaged students	118 teachers/ 7,549 students	1.8 (38)	0.7 (39)	1.1	0.03	+1	≥ .10

Table Notes: The supplemental findings presented in this table are additional findings from studies in this report that do not factor into the determination of the intervention rating. 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 intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. Some statistics may not sum as expected due to rounding. nr = not reported.

[a] For Clark et al. (2015), the *p*-values presented here were reported in the original study. A correction for multiple comparisons was needed and resulted in a WWC-computed critical *p*-value of .02 for End-of-Year Reading Assessments for lower elementary students (pre-K–2); therefore, the WWC does not find the result to be statistically significant. The mean difference is the impact of the intervention (the estimated coefficient on the intervention group indicator from a regression model) as reported in the original study. The unadjusted standard deviations were provided by the study authors at the WWC's request.

[b] For Ware et al. (2011), a correction for clustering was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-values presented here were reported in the original study. The intervention and comparison group means are the mean passing rate gains reported in the original study. The standard deviations are student-level standard deviations calculated by the WWC based on the study-reported current year passing rates. Specifically, the WWC first calculated the standard deviation of the dichotomous student-level passing variable using the formula √(p*(1-p)), where p is the unadjusted current year passing rate for the teacher's students; the WWC then multiplied the result by 100 because the authors reported the passing rates as percentages rather than decimals. The authors define "economically disadvantaged students" as students who are eligible for free or reduced-price lunch.

Appendix D.5: Supplemental teacher subgroup findings in the English language arts achievement domain

Outcome measure	Study sample	Sample size	Mean (standard deviation)		WWC calculations			*p*-value
			Intervention group	Comparison group	Mean difference	Effect size	Improvement index	
Clark et al. (2015)[a]								
End-of-Year Reading Assessments	Grades pre-K–5, teachers in their first 2 years of teaching	23 teachers/ 313 students	nr (1.00)	nr (0.92)	0.13	0.13	+5	.26
End-of-Year Reading Assessments	Grades pre-K–5, *TFA* vs. traditionally certified	132 teachers/ 1,884 students	nr (1.04)	nr (1.41)	0.03	0.02	+1	.64
Henry et al. (2012)[b]								
North Carolina End-of-Grade Reading Assessment	Elementary, *TFA* vs. in-state prepared	263 teachers/ 6,895 students	–0.43 (0.93)	–0.44 (0.94)	0.02	0.02	+1	≥ .05
North Carolina End-of-Grade Reading Assessment	Middle, *TFA* vs. in-state prepared	152 teachers/ 10,346 students	–0.31 (0.97)	–0.35 (0.99)	0.04	0.04	+2	≥ .05
Turner et al. (2012)[c]								
Texas Assessment of Knowledge and Skills (TAKS) Reading	Grades 4–5, experienced teachers	nr/ 596 students	683.11 (nr)	687.20 (nr)	–4.09	–0.05	–2	.64
TAKS Reading	Grades 6–8, experienced teachers	nr/ 2,556 students	774.55 (95.04)	764.19 (91.71)	10.36	0.11	+4	.04

Table Notes: The supplemental findings presented in this table are additional findings from studies in this report that do not factor into the determination of the intervention rating. 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 intervention on outcomes, representing the average change expected for all individuals who are given the intervention (measured in standard deviations of the outcome measure). The improvement index is an alternate presentation of the effect size, reflecting the change in an average individual's percentile rank that can be expected if the individual is given the intervention. Some statistics may not sum as expected due to rounding. nr = not reported.

[a] For Clark et al. (2015), a correction for multiple comparisons was needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-values presented here were reported in the original study. The mean difference is the impact of the intervention (the estimated coefficient on the intervention group indicator from a regression model) as reported in the original study. The unadjusted standard deviations were provided by the study authors at the WWC's request. The subgroup contrast involving non-*TFA* teachers in their first 2 years of teaching also excludes the one *TFA* teacher in the study sample who taught for 2 years prior to entering *TFA* and therefore had a total of 3 years of teaching experience.

[b] For Henry et al. (2012), corrections for clustering and multiple comparisons were needed but did not affect whether any of the contrasts were found to be statistically significant. The *p*-values presented here were reported in the original study. The WWC calculated the program group mean using a difference-in-differences approach by adding the impact of the program (i.e., difference in mean gains between the intervention and comparison groups) to the unadjusted comparison group posttest mean. Please see the WWC Procedures and Standards Handbook (version 3.0) for more information. The intervention group teachers were *TFA* teachers who had less than 5 years of teaching experience. The comparison group teachers were in-state prepared teachers (excluding North Carolina Teaching Fellows Program scholarship recipients) who had less than 5 years of teaching experience. The analytic sample sizes, unadjusted means, and unadjusted standard deviations were provided by the study authors at the WWC's request.

[c] For Turner et al. (2012), the WWC did not need to make corrections for clustering, multiple comparisons, or to adjust for baseline differences. The *p*-values presented here were reported in the original study. The reported group means were estimated using a hierarchical linear model. The effect size presented here for the grades 4–5, experienced teachers subgroup was reported in Turner et al. (2012); the authors calculated the effect size as Hedges' *g*. The intervention group teachers were *TFA* alumni who had completed their 2-year contract assignment but continued teaching in Texas schools. The comparison group teachers were individuals who did not participate in *TFA* and had 3 or more years of teaching experience.

Endnotes

[1] The descriptive information for this program was obtained from publicly available sources: the program's website (www.teachforamerica.org, downloaded April 2016) and the research literature (Clark et al., 2013 and Clark et al., 2015). The WWC requests developers review the program description sections for accuracy from their perspective. The program description was provided to the developer in April 2016, and the WWC incorporated feedback from the developer. Further verification of the accuracy of the descriptive information for this program is beyond the scope of this review.

[2] According to Clark et al. (2015), *TFA* defines selective colleges and universities as those ranked as "selective," "more selective," or "most selective" by *U.S. News & World Report*.

[3] The exact nature of the support *TFA* provides its teachers has varied over time.

[4] The literature search reflects documents publicly available by August 2015. A single study review of Clark et al. (2013) was released in May 2014 and modified in September 2015. Some of the effect sizes reported in the single study review differ from the effect sizes reported in this intervention report because the WWC calculated the effect sizes in this intervention report using unadjusted standard deviations provided by the study authors at the WWC's request. The single study review and intervention report effect sizes differ by no more than 0.02 standard deviations. Both the single study review and this intervention report characterize the study as having a statistically significant positive effect in the mathematics achievement domain.

[5] The studies in this report were reviewed using the Standards from the WWC Procedures and Standards Handbook (version 3.0) and the Teacher Training, Evaluation, and Compensation review protocol (version 3.2). The evidence presented in this report is based on available research. Findings and conclusions may change as new research becomes available.

[6] Absence of conflict of interest: This intervention report includes studies conducted by staff from American Institutes for Research or Mathematica Policy Research. Because American Institutes for Research and Mathematica Policy Research are two of the contractors that administer the WWC, the studies were reviewed by staff members from a different organization. This report was reviewed by the lead methodologist, a WWC Quality Assurance reviewer, and an external peer reviewer.

[7] *TFA* regions across studies included: Baltimore, Chicago, Houston, Los Angeles, the Mississippi Delta, and New Orleans in Glazerman et al. (2006); and Dallas–Fort Worth, Houston, the Rio Grande Valley, and San Antonio in Turner et al. (2012). Clark et al. (2013) and Clark et al. (2015) did not name the 10 *TFA* regions included in each of their studies. Henry, Purtell, et al. (2014), Ware et al. (2011), and Xu et al. (2011) did not report the number or names of the *TFA* regions included in each their studies.

[8] Please see the Teacher Training, Evaluation, and Compensation review protocol (version 3.2) for a list of all the outcome domains.

[9] For criteria used in the determination of the rating of effectiveness and extent of evidence, see the WWC Rating Criteria on p. 45. These improvement index numbers show the average and range of individual-level improvement indices for all findings across the studies.

[10] In the absence of *TFA*, teaching vacancies may have been filled by newly-hired teachers (novice or veteran) or covered by veteran teachers already employed by the district. Because the relevant counterfactual may have included a mix of novice and veteran teachers, comparison groups that included a mix of novice and veteran teachers are acceptable. Differences between intervention and comparison group teachers in background characteristics (e.g., demographics and educational background) may reflect the type of teacher that *TFA* attracts and selects. In other words, teachers' background characteristics may be considered part of the intervention.

[11] In a sensitivity analysis, Clark et al. (2013) also presented complier average causal effect estimates of *TFA* teachers' effectiveness. The authors reported that the findings from this sensitivity analysis were consistent with the estimates that the WWC includes in the intervention's effectiveness rating.

[12] In a sensitivity analysis, Clark et al. (2015) also presented complier average causal effect estimates of *TFA* teachers' effectiveness. The authors reported that the findings from this sensitivity analysis were consistent with the estimates that the WWC includes in the intervention's effectiveness rating.

[13] Henry, Purtell, et al. (2014) also analyzed scores from end-of-grade science assessments and end-of-course science and English language arts assessments. However, the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated for these achievement outcomes; therefore, the results do not meet WWC group design standards.

[14] Henry et al. (2012) describes the in-state prepared teachers as "traditionally prepared" (p. 86), but do not otherwise define the comparison group. Based on the set of teacher preparation categories examined and the definitions given in Henry, Purtell, et al. (2014), the WWC assumes that the in-state prepared teacher category includes North Carolina public school teachers who earned an undergraduate or graduate degree from a North Carolina institution (public or private) and qualified for an initial license before beginning teaching, excluding North Carolina Teaching Fellows Program scholarship recipients. The North Carolina Teaching Fellows Program provides competitive, merit-based scholarships for high school graduates to attend a North Carolina public or private institution

and earn a teaching credential; it also offers program participants fieldwork, cultural opportunities, and other experiences beyond the teacher education curriculum they receive at their university.

[15] The comparison in Turner et al. (2012) between students taught by *TFA* corps members and students taught by novice comparison teachers is presented as the main analysis and included in the effectiveness rating in this report because the novice comparison teachers' experience was more similar to the experience of comparison teachers from other studies included in the effectiveness rating.

[16] Xu et al. (2011) also analyzed end-of-course test scores in the mathematics and general achievement domains. However, the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated for these achievement outcomes; therefore, the results do not meet WWC group design standards.

[17] The Xu et al. (2011) science achievement findings presented in this intervention report are from a model that included controls for teacher experience. The authors also examined the effectiveness of *TFA* teachers on science achievement using a model that excluded controls for experience. Both models use similar analytic samples and produce similar results. Both analyses meet WWC group design standards with reservations. The analytic sample sizes by group were provided by the study authors at the WWC's request.

[18] The six student outcome domains are: English language arts achievement, mathematics achievement, science achievement, social studies achievement, general achievement, and student progression. The eleven teacher outcome domains are: teacher instruction, teacher attendance, teacher retention at the school, teacher retention in the school district, teacher retention in the state, teacher retention in the profession, measures of teacher or school effectiveness in English language arts achievement, measures of teacher or school effectiveness in mathematics achievement, measures of teacher or school effectiveness in science achievement, measures of teacher or school effectiveness in social studies achievement, and measures of teacher or school effectiveness in general achievement.

[19] The following domains were not examined by studies that meet standards: teacher instruction, teacher attendance, teacher retention in the profession, and measures of teacher or school effectiveness in general achievement. Two of the seven studies (Henry et al., 2012 and Henry et al., 2010, which are additional sources for Henry, Purtell, et al., 2014; and Xu et al., 2011) reported findings for outcomes in the general achievement domain. However, equivalence of the analytic intervention and comparison groups is necessary and not demonstrated for these general achievement outcomes; therefore, the results do not meet WWC group design standards. One of the seven studies (Glazerman et al., 2006) reported a finding for an outcome in the student progression domain; however, the authors did not provide the sample sizes needed to determine attrition or the information needed to assess equivalence of the intervention and comparison groups for this outcome, so the results are rated *does not meet WWC group design standards*. One of the seven studies (Ware et al., 2011) reported findings for outcomes in the teacher retention in the school and teacher retention in the school district domains, and two studies (Henry et al., 2012, which is an additional source for Henry, Purtell, et al., 2014; and Ware et al., 2011) reported findings for outcomes in the teacher retention in the state domain. However, equivalence of the analytic intervention and comparison groups is necessary and not demonstrated for these teacher retention outcomes; therefore, the results do not meet WWC group design standards. One of the seven studies (Henry, Bastian, et al., 2014, which is an additional source for Henry, Purtell, et al. 2014) reported findings for outcomes in the measures of teacher or school effectiveness in English language arts achievement, mathematics achievement, science achievement, and social studies achievement domains. However, equivalence of the analytic intervention and comparison groups is necessary and not demonstrated for these measures of teacher or school effectiveness outcomes; therefore, the results do not meet WWC group design standards.

[20] Clark et al. (2013) contained two studies examining the effectiveness of teachers from two different interventions, *TFA* and *Teaching Fellows*. This report only reviews findings for the *TFA* study.

[21] Eligible math courses included any middle school math course and the following high school courses: general math (for example, pre-algebra or remedial math), Algebra I, Algebra II, and geometry.

[22] These sample characteristics are the simple average of the characteristics that the authors reported separately for the intervention and comparison groups. The difference in age—13.44 years for the *TFA* group versus 13.39 years for the comparison group—was statistically significant ($p = .002$). The *TFA* and comparison groups differed by less than two percentage points for each of the remaining demographic characteristics; none of the differences was statistically significant.

[23] Other math-related subjects included statistics, engineering, computer science, finance, economics, physics, and astrophysics. College competitiveness was defined based on *Barron's Profiles of American Colleges 2003*.

[24] Students included in the mathematics achievement analytic sample were taught by 150 teachers, and students included in the English language arts achievement analytic sample were taught by 154 teachers. The authors did not report these teacher counts separately by research condition.

[25] College competitiveness was defined based on *Barron's Profiles of American Colleges 2013*.

[26] *TFA* defines selective colleges and universities as those ranked as "selective," "more selective," or "most selective" by *U.S. News & World Report*.

[27] Observed training and support changes included a decrease in the number of hours of curriculum and literacy sessions assigned during the summer institute from 60 hours 2 years prior to the scale-up to 52 hours in the second year of scale-up. Declines in corps members' satisfaction included decreases in the percentage of corps members who reported feeling that the summer institute was critical for being an effective teacher (from 85% 2 years before scale-up to 75% in the second year of scale-up) and in the percentage who reported positive or very positive overall satisfaction with the program (from 64% to 57% over the same period).

[28] The career plans outcomes do not fall in the teacher retention domains because the review protocol specifies that only teacher retention outcomes reflecting *actual* movement from a teaching position (not expected movement) are eligible for review.

[29] The WWC identified four other additional sources related to Glazerman et al. (2006). These studies do not contribute unique information to Appendix A.3 and are not listed here.

[30] College competitiveness was defined based on *Barron's Profiles of American Colleges 2003*.

[31] For student retention in grade, Glazerman et al. (2006) did not provide the information needed to determine attrition, and the analytic intervention and comparison groups were not shown to be equivalent.

[32] The WWC identified two other additional sources related to Henry, Purtell, et al. (2014). These studies do not contribute unique information to Appendix A.4 and are not listed here.

[33] The Henry, Purtell, et al. (2014) findings are presented as primary findings that factor into the intervention's rating of effectiveness because they reflect the statistical analyses conducted by the authors. The statistical analyses conducted by the authors in Henry et al. (2012) control for the number of days the student was absent. Because days absent is a potentially endogenous covariate (that is, days absent could have been influenced by the intervention), the WWC does not deem the results of these statistical analyses to be a credible source of information about the intervention's effectiveness. The WWC requested, and the authors provided, unadjusted posttest means and standard deviations for the outcomes reported in Henry et al. (2012). In the analytic samples for the school fixed effects analysis of two outcomes (elementary reading and middle reading) examined in Henry et al. (2012), the baseline differences between the *TFA* and comparison groups were less than 0.05 standard deviations. The WWC applied a post-hoc difference-in-differences adjustment to the unadjusted posttest means and standard deviations for these two outcomes and presents the results as supplemental findings. All others results reported in Henry et al. (2012) do not meet WWC group design standards, because either (a) equivalence of the analytic intervention and comparison groups is necessary and not demonstrated, or (b) the analysis does not provide a credible measure of the effectiveness of the intervention.

[34] Henry et al. (2012) describes the in-state prepared teachers as "traditionally prepared" (p. 86), but do not otherwise define the comparison group. Based on the set of teacher preparation categories examined and the definitions given in Henry, Purtell, et al. (2014), the WWC assumes that the in-state prepared teacher category includes North Carolina public school teachers who earned an undergraduate or graduate degree from a North Carolina institution (public or private) and qualified for an initial license before beginning teaching, excluding North Carolina Teaching Fellows Program scholarship recipients. The North Carolina Teaching Fellows Program provides competitive, merit-based scholarships for high school graduates to attend a North Carolina public or private institution and earn a teaching credential; it also offers program participants fieldwork, cultural opportunities, and other experiences beyond the teacher education curriculum they receive at their university.

[35] Henry, Purtell, et al. (2014) also analyzed end-of-grade science test scores (grades 5 and 8 only) and end-of-course (high school) science test scores. The results for these science achievement outcomes do not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated. In addition to the Henry et al. (2012) publication referenced in this appendix, the study includes two other related publications: Henry, Bastian, et al. (2014) and Henry et al. (2010). Henry, Bastian, et al. (2014) analyzed outcomes in the mathematics achievement; science achievement; social studies achievement; English language arts achievement; and measures of teacher or school effectiveness in mathematics achievement, science achievement, social studies achievement, and English language arts domains. Henry et al. (2010) analyzed outcomes in the mathematics achievement, science achievement, social studies achievement, English language arts achievement, and general achievement domains. The results for the outcomes in Henry, Bastian, et al. (2014) and Henry et al. (2010) do not meet WWC group design standards, either because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated, or because the analysis does not provide a credible measure of the effectiveness of the intervention.

[36] The high school test score outcome standardized scores from end-of-course tests in Algebra 1, Algebra 2, biology, chemistry, economics and civics, geometry, physical science, physics, and U.S. history. The high school test score outcome and the retention outcomes do not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

[37] The study included 493 campuses in the analysis of mathematics scores and 483 campuses in the analysis of reading scores. These campus counts are the totals across both the analysis of *TFA* corps members versus novice comparison teachers and the analysis of *TFA* alumni versus experienced comparison teachers. Sample sizes for the analysis of *TFA* corps members are presented in the study sample section of Appendix A.5. Sample sizes for the analysis of *TFA* alumni are as follows: (a) the elementary grade math sample included 423 students of *TFA* alumni from 14 campuses and 423 comparison group students from 98 campuses; (b) the elementary grade reading sample included 298 students of *TFA* alumni from 14 campuses and 298 comparison group students from 80 campuses; (c) the middle grade math sample included 898 students of *TFA* alumni from 12 campuses and 898 comparison group students from 200 campuses; and (d) the middle grade reading sample included 1,278 students of *TFA* alumni from 18 campuses and 1,278 comparison group students from 185 campuses. The sum of campus counts across analyses does not equal the total number of campuses due to overlap across samples.

[38] The comparison between students taught by *TFA* corps members and students taught by novice comparison teachers is presented as the main analysis and included in the effectiveness rating in this report because the novice comparison teachers' experience is more similar to the experience of comparison teachers from other studies included in the effectiveness rating.

[39] Campus-level demographic variables included campus size; number of full-time equivalents; percentage of teachers who were in their first year of teaching; and percentages of students by ethnicity, economically disadvantaged status, special education status, limited English proficiency, and mobility. Campus-level achievement variables included the rates at which students met state standards on the 2009–10 TAKS mathematics and reading assessments.

[40] Student-level demographic variables included gender, ethnicity, economically disadvantaged status, special education status, limited English proficiency, and mobility. Student-level achievement variables included 2009–10 TAKS mathematics and reading achievement scores and all other available TAKS achievement scores for content areas tested at the student's grade level.

[41] Ware et al. (2011) examined the following teacher retention outcomes: cumulative same school retention rate of first-year Texas teachers, cumulative same district retention rate of first-year Texas teachers, and cumulative in-state retention rate of first-year Texas teachers. These retention outcomes were rated *does not meet WWC group design standards* because the equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

[42] The WWC identified an additional source related to Xu et al. (2011). The study does not contribute unique information to Appendix A.7 and is not listed here.

[43] The science achievement findings presented in this report are from a model that included controls for teacher experience. The authors also examined the effectiveness of *TFA* teachers on science achievement using a model that excluded controls for experience. Both models use similar analytic samples and produce similar results. Both analyses meet WWC group design standards with reservations. The analytic sample sizes by group were provided by the study authors at the WWC's request.

[44] The all-subjects high school test score outcome standardized scores from end-of-course tests in Algebra I, Algebra II, biology, chemistry, geometry, physics, physical science, and English I; for some of the subgroup analyses, the outcome did not include English I test scores. Results for the all-subjects outcome do not meet WWC group design standards because equivalence of the analytic intervention and comparison groups is necessary and not demonstrated.

Recommended Citation

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2016, August). *Teacher Training, Evaluation, and Compensation intervention report: Teach For America.* Retrieved from http://whatworks.ed.gov

WWC Rating Criteria

Criteria used to determine the rating of a study

Study rating	Criteria
Meets WWC group design standards without reservations	A study that provides strong evidence for an intervention's effectiveness, such as a well-implemented RCT.
Meets WWC group design standards with reservations	A study that provides weaker evidence for an intervention's effectiveness, such as a QED or an RCT with high attrition that has established equivalence of the analytic samples.

Criteria used to determine the rating of effectiveness for an intervention

Rating of effectiveness	Criteria
Positive effects	Two or more studies show statistically significant positive effects, at least one of which met WWC group design standards for a strong design, AND No studies show statistically significant or substantively important negative effects.
Potentially positive effects	At least one study shows a statistically significant or substantively important positive effect, AND No studies show a statistically significant or substantively important negative effect AND fewer or the same number of studies show indeterminate effects than show statistically significant or substantively important positive effects.
Mixed effects	At least one study shows a statistically significant or substantively important positive effect AND at least one study shows a statistically significant or substantively important negative effect, but no more such studies than the number showing a statistically significant or substantively important positive effect, OR At least one study shows a statistically significant or substantively important effect AND more studies show an indeterminate effect than show a statistically significant or substantively important effect.
Potentially negative effects	One study shows a statistically significant or substantively important negative effect and no studies show a statistically significant or substantively important positive effect, OR Two or more studies show statistically significant or substantively important negative effects, at least one study shows a statistically significant or substantively important positive effect, and more studies show statistically significant or substantively important negative effects than show statistically significant or substantively important positive effects.
Negative effects	Two or more studies show statistically significant negative effects, at least one of which met WWC group design standards for a strong design, AND No studies show statistically significant or substantively important positive effects.
No discernible effects	None of the studies shows a statistically significant or substantively important effect, either positive or negative.

Criteria used to determine the extent of evidence for an intervention

Extent of evidence	Criteria
Medium to large	The domain includes more than one study, AND The domain includes more than one school, AND The domain findings are based on a total sample size of at least 350 students, OR, assuming 25 students in a class, a total of at least 14 classrooms across studies.
Small	The domain includes only one study, OR The domain includes only one school, OR The domain findings are based on a total sample size of fewer than 350 students, AND, assuming 25 students in a class, a total of fewer than 14 classrooms across studies.

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 and inclusion in this report 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.

Extent of evidence An indication of how much evidence supports the findings. The criteria for the extent of evidence levels are given in the WWC Rating Criteria on p. 45.

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

Intervention An educational program, product, practice, or policy aimed at improving student outcomes.

Intervention report A summary of the findings of the highest-quality research on a given program, product, practice, or policy in education. The WWC searches for all research studies on an intervention, reviews each against design standards, and summarizes the findings of those that meet WWC design standards.

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 study participants 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 eligible study participants are randomly assigned to intervention and comparison groups.

Rating of effectiveness The WWC rates the effects of an intervention in each domain based on the quality of the research design and the magnitude, statistical significance, and consistency in findings. The criteria for the ratings of effectiveness are given in the WWC Rating Criteria on p. 45.

Single-case design 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.

Glossary of Terms

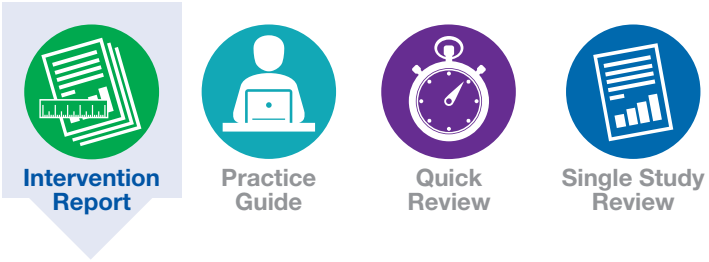
Standard deviation The standard deviation of a measure shows how much variation exists across observations in the sample. A low standard deviation indicates that the observations in the sample tend to be very close to the mean; a high standard deviation indicates that the observations in the sample tend to be 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% ($p < .05$).

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

Systematic review A review of existing literature on a topic that is identified and reviewed using explicit methods. A WWC systematic review has five steps: 1) developing a review protocol; 2) searching the literature; 3) reviewing studies, including screening studies for eligibility, reviewing the methodological quality of each study, and reporting on high quality studies and their findings; 4) combining findings within and across studies; and, 5) summarizing the review.

Please see the WWC Procedures and Standards Handbook (version 3.0) for additional details.



An **intervention report** summarizes the findings of high-quality research on a given program, practice, or policy in education. The WWC searches for all research studies on an intervention, reviews each against evidence standards, and summarizes the findings of those that meet standards.

This intervention report was prepared for the WWC by Mathematica Policy Research under contract ED-IES-13-C-0010.